Observatoire ARGA

INTERNATIONAL ANALYTICAL REPORT

Study No. 19

Investigating Serious AI Incidents: Evidentiary Standards, Reporting and Accountability

Preserving Digital Evidence, Causal Investigation and Protecting Affected Persons in International Practice

Organisation: Observatoire ARGA

English

21 rue de l'Aviation, 64600 Anglet, France

info@argaobservatory.org | www.argaobservatory.org

22 October 2026

Contents

Annotation — 3
1. Scope of the study, limits of conclusions and research method — 4
Scientific grounds and research design — 5
2. Concepts: events, hazards, incidents and serious incidents — 7
3. Recording initial information without prejudging causes or liability — 9
4. Roles: developer, provider, deployer, integrator and human actors — 10
5. Detection: signals, monitoring and the limits of automated detectors — 13
6. The evidence preservation rule: preserve the object before changing it — 14
7. Chronology and unsynchronised clocks — 16
8. Model logs, prompts, retrieval context and external tools — 18
9. Data sources and the provenance of disputed attributes — 19
10. Human oversight: effective supervision or merely formal approval — 20
11. Causation: from a sequence of events to a testable explanation — 22
Verification of causal hypotheses in case of false sanctions match — 23
12. Statistical testing, rare events and sampling sensitivity — 24
13. Signs of a deliberate attack: prompt injection, privilege bypass and memory manipulation — 26
14. Cyber incidents, AI incidents and their overlap — 28
15. Model failure and process failure: separating five sources of defects — 29
16. The EU AI Act: the legal definition and scope of a serious incident — 30
17. Article 55 of the AI Act: serious incidents involving general-purpose models with systemic risk — 32
18. Article 73 of the AI Act: reporting serious incidents involving high-risk systems — 34
19. The GDPR: AI incidents and personal data breaches — 36
20. NIS2 and DORA: when an AI incident is also an ICT incident — 37
21. The content and stages of regulatory notification — 39
22. Territorial scope, conflicting rules and cross-border transfers — 40
23. Administrative investigations and independent expert assessment — 41
24. Criminal and civil proceedings: admissibility and limits of technical conclusions — 42
25. Human rights, remedies and restoring the affected person's position — 43
26. Sanctions screening and the risk of mistaken identification — 45
27. International police cooperation, wanted-person alerts and extradition — 46
28. AML/KYC: automated suspicion and its consequences for customers — 48
29. Algorithmic credit scoring, insurance and private-sector decisions — 49
30. Medical AI and the safety of critical infrastructure — 50
31. Synthetic documents, fabricated media and evidentiary errors — 51
32. Independence of expert organisations and conflict management — 52
33. Protection of sources, trade secrets and personal data — 53
34. Root-cause reports: structure, quality and limitations — 54
35. Corrective measures, retesting and resumption of operation — 55
36. Recording losses, compensation and correction of third-party data — 56
37. Public incident registers and the limits of disclosure — 57
38. International notification exchange and confidential coordination — 58
39. The independent Observatoire ARGA methodology: five quality standards — 59
40. Contractual architecture for expert investigations — 60
41. Quality control of the investigation and safeguards against AI-generated errors — 61
42. Development priorities for 2027-2028 — 62
43. Conclusions and institutional proposals — 63
Practical annexes: forms, protocols and monitoring procedures — 64
Appendix 1. Initial AI incident report record — 64
Appendix 2. Priority and immediate protection scheme — 65
Appendix 3. Model order to preserve electronic footprints — 67
Appendix 4. Digital artifact register and integrity check — 68
Appendix 5. Chain of society magazine — 69
Appendix 6. A.I. passport at the time of the event — 70
Appendix 7. Protocol on Time Normalization — 70
Appendix 8. Mandatory and voluntary communications matrix — 71
Appendix 9. Form of early communication to the competent authority — 72

Appendix 10. Request to the supplier of the basic model and the integrator — 74
Appendix 11. Report of the interview of the operator and the decision maker — 75
Appendix 12. A matrix of causality — 75
Appendix 13. Test environment control playback and restriction plan — 77
Appendix 14. Statistical form for testing rare harm — 78
Appendix 15. Checking discriminatory effects — 78
Appendix 16. Expert independence checklist — 79
Appendix 17. Instruction to a judicial or technical expert — 80
Appendix 18. Investigation calendar and establishment of control points — 81
Appendix 19. Notification of the victim and explanation of the restriction — 82
Appendix 20. CAPA Register of Corrective Actions — 82
Appendix 21. Protocol of the decision on the renewal of the — 83
Appendix 22. Table of levels of confidentiality and legal transfer — 84
Appendix 23. Structure of the expert opinion for external verification — 85
Appendix 24. Criteria for closure and follow-up — 85
Scenario Annexes: Methodology of application in model cases — 86
Appendix 25. Scenario: Identity error in the sanctions check — 86
Appendix 26. Scenario: algorithmic match at border crossing — 88
Appendix 27. Scenario: Automatic classification in the asylum procedure — 90
Appendix 28. Scenario: Mass AML operation after model update — 91
Appendix 29. Scenario: The agent changed the monitoring settings during maintenance — 92
Appendix 30. Scenario: mistaken medical sorting of patient — 93
Appendix 31. Scenario: The generic summary changed the content of the legal instrument — 94
Appendix 32. Scenario: Deepfake as an alleged proof of action — 95
Appendix 33. Scenario: A major language agent disclosed personal data — 96
Appendix 34. Scenario: the rating model was subjected to data poisoning — 97
Appendix 35. Scenario: When the model was updated, the possibility of replication disappeared. — 97
Appendix 36. Scenario: critical infrastructure and incompatibility of investigative objectives — 98
Appendix 37. Scenario: AI classification of mass complaints covered a serious problem — 99
Appendix 38. Scenario: AI expert was not provided with data due to commercial secrecy — 100
Appendix 39. Scenario: correct AI-suggestion illegally used — 101
Appendix 40. Scenario: Regulator dates started to count as a wrong date — 103
Appendix 41. Comparison of ISO and NIST standards in the investigation programme — 104
Appendix 42. Harmonization of methodology with human rights — 104
Appendix 43. Institutional system for the prevention of repeat AI incidents — 105
Appendix 44. Practical protocol for the private accounter — 106
Appendix 45. Report of whistleblower protection and internal error report — 107
Appendix 46. Qualitative evaluation of the unreliable AI-output — 107
Appendix 47. Report completeness test without disclosure of prohibited information — 108
Appendix 48. Legal review of categories of automated solutions — 109
Appendix 49. Comparison of AI-incident public bases — 109
Appendix 50. Coordination with AI Office and national authorities — 110
Index of Practical Recommendations — 111
Sources and normative materials — 113



Annotation

The report develops a methodology for investigating AI incidents, which divides event reconstruction, causal mechanism testing, and legal characterisation. The scientific basis includes digital forensics, causal analysis, system security and statistics. The central task is to determine which observations confirm a particular component's involvement in an adverse outcome, which

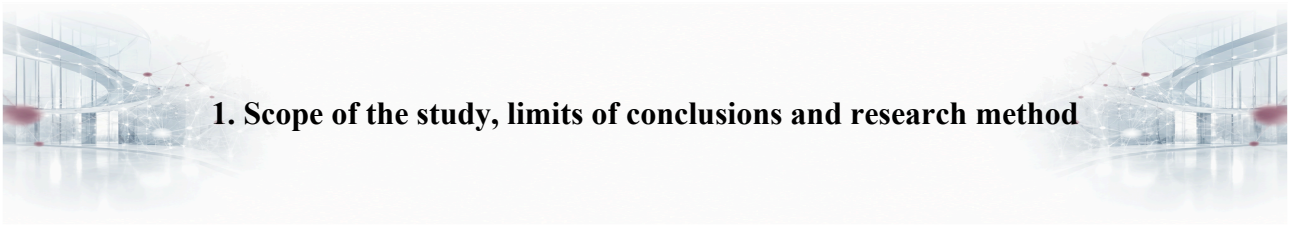
alternatives remain open, and how to verify the outcome of the correction. Every substantial conclusion is linked to the preserved material and the research operation described. [M1–M3]

The study specifically distinguishes the scientific and methodological category AI, legally defined "serious incident" AI Act, security violation of GDPR personal data, significant incident NIS2 and major ICT-related internal DORA. The legal threshold, the address, the time of commencement of the period and the applicability of the regime are set separately. The consolidated AI Act of 27 July 2026 and the EU acts applicable on 22 October 2026 are used. The 50 Applications are expert forms and training scenarios; they are not regulatory-approved templates and do not describe the applicants' actual cases.

The proposed program includes a map of available and lost data, temporary reconstruction, comparison of explanations, limited reproductive checks, and a separate analysis of the implementation of protection measures. The expanded hypothetical example of a false sanction match shows how to distinguish between a defect in the original record, a comparison algorithm, and an organizational solution. The practical outcome for the ARGA is a causal investigation protocol and verifiable completion criteria adapted to the powers and applicable law.

Brief map of notification regimes

Mode	Condition and address	Time frame / starting point
AI Act, Article 55	Supplier of GPAI-models with systemic risk; AI Office and, if necessary, national competent authority	Without undue delay; conditions of art. 55(1)(c)
AI Act, Article 73	High-risk AI system supplier when Chapter III is applicable; market surveillance authority	Usually up to 15 days; special cases 2 or 10 days. Start date of the separate mode
GDPR, Article 33	Monitor, if personal data breach creates a risk of rights and freedoms; DPA	If possible, up to 72 hours from the moment of knowledge if notice is required
NIS2, Article 23	Organization subject to the directive; CSIRT or competent authority	Early warning 24h, notification 72h; with reservations of the national regime
DORA + EU 2025/301	Financial entity at major ICT-related entity; competent authority	The first message before 4 hours after classification and not later than 24 hours after being informed; further on the established stages



1. Scope of the study, limits of conclusions and research method

The subject of the study is the evidentiary sufficiency of AI-incident explanation. It is necessary to establish not only the sequence of actions, but also the basis for each transition from observation to a supposed cause. The question is formulated in relation to a specific outcome: incorrect identification, inaccurate access restriction, distribution of inaccurate records or other confirmed consequences. The involvement of AI, technical defect and legally significant causal contribution are checked separately. [1; 3; 5; M1–M3]

The study was completed as of 22 October 2026. It is divided into four layers: compulsory law in force; voluntary technical standards; actual practices of organizations; and observational proposals

from ARGA. Neither the algorithmic evaluation itself nor the certification of the control system is a legal conclusion about guilt. In each case, it is necessary to determine the applicable jurisdiction, the type of system, the stage of its life cycle, the time when the obligation arises and the position of the parties.

The research method combines comparative analysis of primary normative materials, principles for working with digital objects, structuring of causal hypotheses and constructing learning scenarios. The report offers a way of organizing the investigation and the criteria for verifying its results. It doesn't involve an experiment on a real client case and it doesn't establish the effectiveness of the methodology on a set of incidents. The pilot of such a test is described separately; the authorship procedure is not presented by a state or international standard.

1.1. Qualitative issues and criteria tested

For each incident, research questions and material can be recorded. What acts have been committed, and what records have been confirmed? Where does the first unfounded information come from? How could she influence the decision? What barriers should have revealed the error? What alternative explanations are compatible with the data? What changed after the correction? The issue of notification is parallel: the need for regulatory action is not automatically deferred until the cause is finally established.

Criterion	Minimum sufficient confirmation	Unacceptable replacement
Event	Primary artifact and register	Retold without origin
Time	UTC and local reference area	Undated memory
Object	Model and configuration ID	Name of commercial product
Cause	Verification of the chain of events and alternatives	Time overlap
Damage	Impact data and source of verification	Non-critical evaluation in PR release
Law	Norma, addressee, deadline, start point	"Any incident = 72 hours notice"

1.2. Research restrictions and criteria for external verification

The report suggests a distinction between a universal analytical procedure and a specific judicial or administrative proceeding. The same technical facility may be sufficient for internal security checks, but not enough to establish guilt in a criminal case. Therefore, the methodology does not promise a single international standard of proof. It offers a minimal infrastructure of evidence control that can be adapted to the parties' legal powers, procedural rules and access levels to the source material.

The force of approval depends on the type of conclusion. The integrity of the file confirms the preservation of the object received, but not the authenticity of its contents. Reproduction shows the behavior of the restored configuration, but does not in itself prove the historical course of the event. The causal model explains the supposed connection, and the legal characterisation evaluates its significance in a particular procedure. In the report, these grounds are listed directly next to the conclusion so that the next reader can check each transition.

Scientific grounds and research design

Research design must be consistent with the historical nature of the incident. The existing records are incomplete, the various components change over time, and the adverse outcome is already

known to the researcher. Therefore, the methodology controls separately the preservation of the material, the validity of the reconstruction and the impact of the knowledge of the outcome on interpretation. Digital forensics, causal analysis, and system security are used for a variety of interconnected operations. [M1–M3]

Scientific basis	Operation	Limit of result
Digital forensics	Collection, preservation and research of digital objects	Integrity does not prove true content
Causal analysis	Formulation of intervention and comparison of explanations	The impact of reconstruction is not the same as the historical cause
System security	Verification of restrictions and feedback	Lack of control does not automatically determine guilt
Statistics	Assessment of frequency and variability with known sample	One incident doesn't make it widespread.
Procedural law	Evaluation of material receipt and use	The legal value depends on the procedure

Subject matter and accessibility map

The object is described as the version of the system at the time of the event: data, model, settings, tools, interface and organizational rules. For each component, one status is recorded: directly preserved; reliably restored; reconstructed with permissions; inaccessible. This allows you to determine in advance which experimental conclusions are achievable and where the documentary or probable explanation will remain.

When working with digital objects, the original, working copy and derivative materials are shared. The checksum allows you to verify the consistency of a recorded copy, but does not establish the author or truth of the information. Transformations, timing corrections and journaling are recorded so that another researcher can restore their sequence. The right of access and the admissibility of use are determined separately. [M1]

Research hypotheses and observed investigations

The cause of the hypothesis indicates the mechanism and the observation that would help to distinguish it. If an error occurs in the original register, it must be present before processing by the model. If it is entered by transformation, the correct input and incorrect output are required in a comparable configuration. If the damage is caused by the execution procedure, it is necessary to investigate the transition from signal to action. Several such hypotheses may be jointly true.

To compare hypotheses, supportive and contradictory data, source dependence and lack of verification are indicated. The absence of an error record is not the same as the confirmed absence of an error if the fullness of the registration is unknown. The participant's statement is evaluated together with the terms of his knowledge. The severity of the conclusion should not increase just because alternative sources are unavailable.

Reproduction and counterfactual verification

Causal analysis requires precise determination of the conditions to be compared [M2]. The authoring procedure first checks the reproducibility of the original script, then changes one significant factor while preserving the rest. For a probability model, the distribution of pre-determined outcomes is compared rather than one convenient answer. All departures from the historical configuration are described before the result is interpreted.

The experiment doesn't restore an unobserved human decision automatically. The document may show that the operator was required to verify the identifier; this does not yet prove that he would have done so in a different interface. The CAST system approach helps to explore the constraints, information and feedback available to participants [M3]. The conclusion on legal causality and liability remains a separate legal characterisation.

Quality assurance of the methodology itself

The pilot is offered training files with pre-prepared refusal mechanisms and controlled data gaps. Different researchers build an explanation independently. The main outcomes are unfounded assertion of the cause, failure of the confirmed mechanism and ability to restore the chain of evidence. The reference breakdown is hidden until the work is completed; several explanations are allowed if the original materials do not really distinguish them.

The methodology is considered suitable for the next stage if researchers maintain substantial uncertainties and formulate reproducible grounds for their conclusions. Differences are analyzed for reasons such as incomplete sources, ambiguous categories, miscalculation or legal interpretation. The pilot is not imitating the court ruling and confirming the real effectiveness of the ARGA investigations until the relevant data is obtained.

Practical case analysis 19.01

Recommendation: Segregate event and explanation of cause

Hypothetical example: Once the assistant was upgraded, the client received an erroneous notice.

Analysis: Time sequence doesn't prove why.

Practical steps: The manager records the set notice, version and several explanations; assigns different checks.

Outcome criterion: The conclusion separates the facts, the hypotheses, and the confirmed mechanism.

2. Concepts: events, hazards, incidents and serious incidents

The OECD proposes to distinguish AI incident as an event that led to actual harm and AI hazard as a circumstance capable of causing it. These categories are designed for international monitoring comparability, not automatically determining the existence of a breach. AI Act uses its own legal definition of a serious incident in Article 3 ((49)), so the same facts may be included in the OECD scientific statistics, but do not meet the requirements of a specific EU standard. [1], [3], [4]

In developing an internal classification, two independent fields should be maintained: "observed harm or hazard" and "legal classification according to applicable regime". The third field is the degree of confirmation: message, prior confirmation, independent verification, disputed fact and rebuttal. It is wrong to place these three axles in one label "critical". The event may be catastrophic in terms of potential severity and not have any proven damage or be limited in scope and a defined violation of the rights of a particular person.

The European definition of Article 3*49 includes death or serious damage to health, serious and irreversible breaches of critical infrastructure, violations of EU law's obligations to protect fundamental rights, as well as serious damage to property or the environment. In an investigation, it's helpful to record the event separately, the legally significant consequences and the causal contribution of AI system; any serious event that happened next to AI has not yet proven such a contribution. [1]

2.1. Practical taxonomy for the registry

It is proposed to conduct classes I0 - unconfirmed signal, I1 - danger without established harm, I2 - confirmed harm without assessment of severity, I3 - probable serious incident, I4 - serious incident, corresponding to a specific legal definition. Class I4 is assigned with the norm and date of assessment indicated; a change in class must never rewrite history. This is an internal ARGA classification that does not replace legal criteria and does not involve automatic communication to the regulator.

Example of fact	Analytical expertise	What More to Find
The model gave a false alarm, it didn't come to a decision.	Risk or technical event	Was there an intervention, risk of repetition?
False match caused the account to be blocked.	Actual injury incident	Grounds of blocking, role of man
The agent deleted the audit log.	Control incident and possible IB	Whether the data were lost, the consequences
The wrong decision led to the unlawful deprivation of liberty	A Potential Serious Event	Chain of decisions, applicable law

2.2. Why do you need an independent scale of gravity and confirmation?

The external statistics of AI incidents are not automatically comparable to the legal data of market surveillance. The OECD offers a convenient international language for actual and potential harm, but jurisdictions can set their own thresholds. The internal register must keep an unprocessed description of the event, the degree of confirmation, the possible type of harm, the extent of the impact, the status of the investigation and individual legal characterisations. This allows data to be aggregated in the future without erasing the original context.

The risk of escalation with incomplete information deserves particular attention. The lack of photographs or medical records should not make life-threatening messages a low priority. However, timely preventive protection does not require an organization to declare the victim or statistically confirmed harm frequency prematurely. The classification system should allow protective actions of high severity of probable consequence with limited confidence in the causal mechanism.

Practical case analysis 19.02

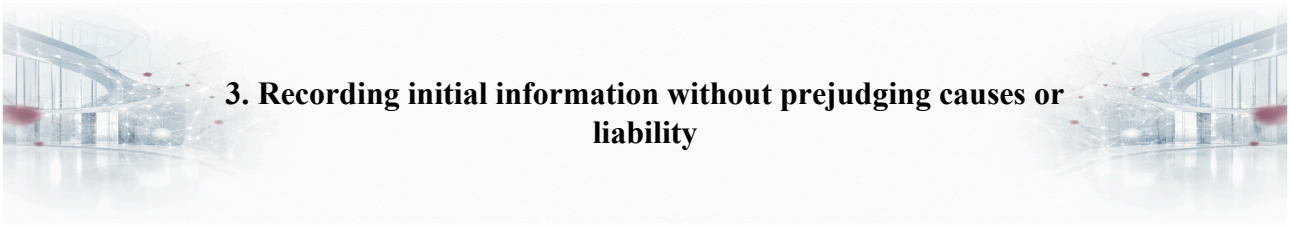
Recommendation: Classify an event for proven consequences

Hypothetical example: The system formed an erroneous draft that the staff member corrected before shipping.

Analysis: The potential harm is different from the actual one.

Practical steps: The analyst records the hazardous scenario and prevention; the legal category checks separately.

Outcome criterion: No actual damage was reported in the registry without confirmation.



3. Recording initial information without prejudging causes or liability

Any investigation begins with a record of the report, not a finding of the perpetrator. The receipt includes the source, method of obtaining, date in UTC, original formulation, system identifier, available artifacts, affected persons and known damage control actions. The complainant's words should be distinguished from those of the examiner: for example, the phrase "the system has placed me under international surveillance" does not establish that the system had such powers or was used at all in a particular procedure.

The original appeal shall be retained without editorial changes. When the screenshots are received, they record the original file, metadata and transmission method, not just the embedded image in the presentation. If contact is made via a messenger, the correspondence archive should be kept in a manner consistent with local law and privacy rules; it is not necessary to copy all personal correspondence without reason. If there is an urgent risk to life or liberty, the defense procedure should not wait for forensics to be completed.

Three regimes should exist in parallel: the protection of the victim, the preservation of evidence and the assessment of the obligation to notify the external authorities. Separation prevents a common mistake when the legal department first requires an internal investigation to be completed, although the regulatory deadline may have already begun. The log not only records the moment of the event, but also every moment of the legal entity's knowledge - it often determines the timing. [1], [8], [10]

3.1. Conditions for immediate escalation

An internal challenge of an independent specialist is required in the event of possible death or serious health damage, a threat to fundamental rights, interference with critical infrastructure, massive data breaches, unexplained logging or evidence of concealment. Escalation is an organizational security measure, not a legal characterisation statement. If it is not possible to assess the risk promptly, a temporary suspension of the hazardous function shall be applied, documenting what has been changed.

First hour	Objective	Document
Recording of referral	Save Primary Story	Form of income
Notification of the responsible party	Do not lose urgent duties	Excess log
Technical preservation	Record the reference state	Act snapshot
Protection measure	Avoiding Further Harm	Act of limitations

3.2. Three parallel magazines and prohibition of retrospective substitution

The first log records all messages received in the original form: from whom, through what channel, which original documents were attached and who was the first to get acquainted with them. The second registers protective actions and technical changes, including failed attempts to limit the function. The third describes the process of legal characterisations and external notifications. This division allows you to trace a situation where the technical team already knew about the dangerous behavior of the model, but the lawyers got the information much later, or when the local operator learned about the incident before the central supplier. This may be relevant for the calculation of time in different modes, but it does not mean that the date of the internal legal agreement automatically replaces the date on which the person concerned became aware.

Every magazine should be as important-only as technically possible. The correction of an error in the record is entered with a new line indicating the previous value and the reason for the correction. UTC markings are not allowed to be brought up to the expected regulatory period. If an organization has received a notification from an external source and could not open the account, it needs to record separately the fact of receiving the message, the fact of technical unavailability and the moment of receiving the relevant information. In real life, a legal assessment of such an episode may require an examination of the responsibilities of the specific addressee, the nature of the message and the reasonableness of the actions taken.

Type of log	Legally relevant issue	Quality control
Intake	Who and when would have known about the event?	Save Original Message, Headers, delivery IDs
Containment	What kind of interference changed the object?	Record change of version, configuration and authority
Legal reporting	Why was the communication sent or not sent?	Legal rule, responsible, time of decision, version of form

Practical case analysis 19.03

Recommendation: Record first information without indictment

Hypothetical example: The client sent a copy of the rejection and suggested the use of AI.

Analysis: The photo confirms the rejection report, but not the solution architecture.

Practical steps: The receiver retains the material, time and words of the applicant; asks for a document on the grounds for refusal.

Outcome criterion: The primary record distinguishes between the message and the verified information.

4. Roles: developer, provider, deployer, integrator and human actors

In large AI chains, functional roles are shared between the base model developer, API provider, interface creator, operator organization, and final decision-making employee. For proper attribution, the architecture of the operation should be drawn: the data came from source A, processed by service B, transformed by rules C, shown to analytics D, formulated into the decision of organ E. Until the pattern is checked, the phrase "guilty of neurosmart" conceals the causal links.

AI Act's responsibilities depend on the legal category "provider" and "deployer" and may change with substantial product modification, re-labelling or integration. Internal service agreements do not automatically transfer public legal obligations to the regulator. Similarly, the lack of access to model weights does not exempt an operator from keeping its own inputs, usage policies and employee decisions. [1], [1]

Investigation requires a register of data possession across five layers: interface and user query, orchestration and tools, model and its versions, knowledge base and source arrays, human resolution and subsequent effect. The registry lists the access manager, the person responsible for keeping the logs, the owner of the logs, the place of storage and the time of removal. This is particularly critical for external APIs, where the provider's logs can be stored for less than administrative correspondence takes.

4.1. Minimum RACI Model

The investigator must not simultaneously confirm his final conclusions and change the original data. It's practically useful to appoint a technical keeper of the artifacts, legal notices coordinator, an independent methodologist for the assessment of causes and a representative of the affected persons. In small organizations, roles can be combined, but the conflict of functions is recorded in writing and compensated by external verification.

Role	Main decision	Required trail
Incident lead	Status and priority	Signed timeline
Evidence custodian	Collection, access, hash	Storage log
Regulatory Lawyer	Addresser and date of communication	Legal memorandum
Independent expert	Methodology and alternatives	Expert advice
Representative of the victims	Correction and protection	Register of applications

4.2. Contractual responsibility for logging and access

Even when the supplier and the operator have agreed in the contract that one of them is "accounting for AI", the public obligation of each entity is determined by the law and its actual role. For an effective investigation, it is not an abstract guarantee of conformity that is required, but specific provisions on keeping logs, sufficient depth of audit, the timing of material release after the incident and the procedure for confidential access by an independent expert. The chain may include model suppliers, cloud providers, integrators, solution operators and primary registry holders, each with only a portion of the evidence.

A map of the refusal points is useful. For example, if the integrator does not have the right to request the history of the model versions from a cloud supplier, its obligation "to provide complete logs" is potentially not fulfilled. Such restrictions must be detected before a serious incident occurs, and the requests, deadlines and responses documented during an already occurring event. The supplier's unilateral certification application does not replace access to the information necessary for causal verification.

Practical case analysis 19.04

Recommendation: Record actual actions of participants

Hypothetical example: The model supplier did not manage the bank's decision, but was named the sole perpetrator.

Analysis: The role of the developer is not equal to that of the operator and the decision maker.

Practical steps: The process investigator is building a matrix of actions, access and authority over documents.

Outcome criterion: Each causal statement relates to the action of a particular participant.

5. Detection: signals, monitoring and the limits of automated detectors

An AI incident may first appear in a customer complaint, a discrepancy between a manual and an automatic solution, an employee complaint, an external vendor notification, a news message, or an anomaly in production metrics. Monitoring only the technical availability of the API is fundamentally inadequate: the system is able to respond quickly and without failures, while systematically generating illegal or unfair results. So the metrics should include the impact on people: the proportion of decisions that are canceled, the time it takes to correct an error, the distribution by group, and the cases where a person refuses to review the machine inference. [5], [6]

It is not possible to declare a public database of incidents representative of frequency statistics. The OECD AIM uses automatically-classified news sources to directly warn of incomplete surveillance. The lack of reports in monitoring does not mean that there is no harm, and the increase in their numbers does not prove that models are deteriorating without analyzing media coverage and changing methodology. [4]

In the recommended alarm mode, it creates a verifiable request for investigation, not automatically determines legal significance. For example, a sharp jump in the number of suspicious transactions after a model update may be due to real environmental variability, faulty feature engineering, threshold crossing, or upstream source disruption. The signal classifier should also keep the wrong triggers, as systematically ignoring false alarms creates a risk of missing a serious event.

5.1. Detector design and testing

Checking the quality of monitoring requires a fixed test period, identifying the true signal and taking into account false negative detections. An alarm threshold cannot be chosen only for the low cost of checking: in systems that affect freedom or access to basic services, the cost of passing and the cost of an incorrect escalation are highly asymmetrical. All decisions to change the threshold are recorded as a new version of the monitoring rule.

Signal source	What Shows	Which does not prove
False positive	Change in verification behaviour	Model error itself
The human complaint	Specific adverse effects	Systematicity of the problem
Lack of logs	Observation gap	Intentional concealment
Non-consent of experts	Uncertainty of classification	Arbitraryity of any decision

5.2. Sensitivity test of detection system

Monitoring checks should cover two mistakes: misplacing and missing a truly significant event. Accuracy based on a rare and unbalanced sample can seem high even with a virtually useless detector. So the protocol reflects precision and recall for pre-defined types of events, sample size, method of building a reference and conditions for delivering alerts to employees. If a message is sent but does not arrive in the queue, it is a separate routing failure that cannot be eliminated by improving the model itself.

In a repeat incident group, it is important not to combine events involving different individuals without verification into one "technical ticket". The same root defect may require corrective action at the product level and individual remedies for each affected person. The reverse error is counting the same event multiple times as independent incidents due to duplicate alerts. Monitoring is required to maintain both levels: the individual family and specific activities.

Observed metric	What Uses	Which does not prove
Alert precision	Share of relevant investigated signals	No missing episodes
Independently marked Recall	Level of detection of known events	Coverage of unknown harm classes
Time to triage	Speed of initial analysis	Legitimacy of the outcome
Percentage of successful deliverys	Operation of the notification channel	Actual elimination of effects

Practical case analysis 19.05

Recommendation: Check the detector signal before escalation

Hypothetical example: The leak signal is caused by a conditional training discharge.

Analysis: The operation doesn't confirm a real incident.

Practical steps: The duty officer checks the launch destination, the data composition and the actual recipient.

Outcome criterion: The signal's status is changed for reasons of keeping the original recording.

6. The evidence preservation rule: preserve the object before changing it

At the first sign of a serious incident, an organization must establish a legal hold or functionally similar order to preserve potentially significant information. It includes production logs, configuration images, deployment records, test sets, source documents, incident correspondence and political/managerial decisions prior to the event. The list should be proportionate to the circumstances and privacy requirements: a public indefinite copy of personal data can itself become a new breach.

Evidence work is based on the principles of origin, integrity, verifiability, reproducibility and access control. ISO/IEC 27037 describes recommendations for identifying, collecting, obtaining and preserving digital evidence; the application of the standard does not automatically turn an artifact into permissible evidence before any court. It is necessary to comply with the procedural rules of a particular body and to be able to verify who, when and how worked with the carrier. [13]

For a cloud model, the most fragile evidence is often not files in the usual sense. It's a version of the remote model, the state of the external API, changes to the query routing, dynamic system prompting, policy filters, retrieval-body identifier and call log of the tool. They need to be kept together; a separately extracted LLM answer is sometimes impossible to reproduce and correctly interpret.

6.1. The condition of cold and hot preservation

If disconnecting the system itself is a threat to humans, it is admissible to first perform safe isolation or manual transition while simultaneously recording the condition. The act specifies why complete freezing was technically impossible, what components were changed, whether control copies were created and what information was irreversibly lost. Change of configuration under the euphemism of "normal maintenance" cannot be hidden if it affects future cause testing.

Object	Priority	Means of verification of integrity
Runtime logs	Immediately	Signed export, hash
Model/version manifest	Immediately	ID version, image digit
Staff member's decision	Immediately	Unchangeable entry/act
Policy and instruction	High	The current text
Interview with participants	After stabilization	Protocol and time

6.2. Plan to secure the evidence without deteriorating security

Keeping evidence is not the same as keeping a dangerous function available indefinitely. If an agent is able to send payments or publish personal data on his own, the first step is to limit his authority. However, the command must describe the minimum necessary intervention: disconnecting the external instrument, rotating the key, switching to read-only mode or a complete stop. The protocol must contain the time, initiator, purpose, observed state before intervention and expected impact on the possibility of subsequent reconstruction. If the hazardous function is stopped before the logs are unloaded, this must be specified directly, not filled in with the presumed information.

The digital set of evidence is better divided into unalterable originals and working copies. The original is kept with limited access, a check-up fee is calculated and a log of appeals is kept. A working copy can be filtered, personal data masked, and formatted, but the history of conversions remains. The hash proves the matching of the two bytes at the corresponding moment, rather than the truth of each of the events contained in them. In particular, a cryptographically identical log may reflect incorrect system time, incomplete event coverage, or a record made by a compromised service.

The objects that can disappear the fastest should be identified in advance: short-lived containers, message queues, retrieval service responses, cloud telemetry records, tool caches, model call tokens, human verifier logs and routing settings. For each object, the available storage window and the technically permissible storage strategy are indicated. The transfer of materials from the EU to a third country requires a separate verification of the legal basis, even if the audience is convinced that the foreign expert is "only helping to understand".

Object	Possible cause of loss	Priority action
Temporary transfer agent	Short retension policy	Immediately signed exports
External API-response	Supplier only stores aggregates	Request retension gold and correlation ID
Knowledge base	Documents are rewritten	Snapshot version of index and sources
Human review	Hand-made solution outside workflow	A confirmed protocol of action with a separation of facts and memories

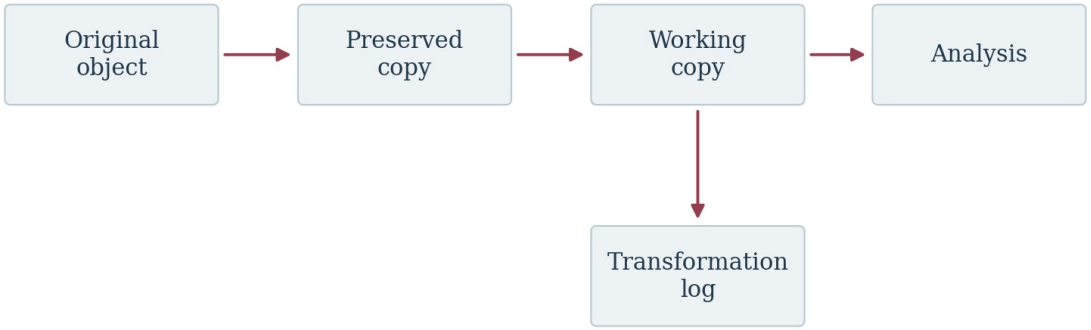


Figure 19.1. Preserving a digital object before intervention. Procedure in Section 6. Preservation and provenance are assessed separately from the truth of the content.

Practical case analysis 19.06

Recommendation: First save the object then correct

Hypothetical example: The administrator is ready to delete the wrong recording and restart the service.

Analysis: The removal could destroy the conditions of reproduction.

Practical steps: Responsible keeps authorized copies, logs, versions and integrity controls; then allows correction.

Outcome criterion: The historical object is available for independent analysis.

7. Chronology and unsynchronised clocks

The chronology is built with the time source, time zone, synchronization and possible delay in recording. First, the original values are stored; normalized time is stored in a separate field with the rule of conversion. If the exact moment is unknown, an interval is used, and the order of events is considered to be established only if the data actually allows it to be distinguished. Artificial accuracy can create a false explanation of causality.

There are at least four distinct points: there was an AI drawdown; an operational decision was made; damage occurred; the relevant subject learned about the event. For regulatory notification it is also important to establish a reasonable link between the system and the consequence, if provided for in the norm. It is necessary to keep a chronology of the decisions of different legal entities: the supplier could get the ticket in the morning, the operator saw the damage during the day, and the body - the next day.

The time scale includes periods of unknownness. For example, two log events are known at 10:02 and 10:08, but the person's notification is removed. Correct conclusion: "The transfer could have taken place between the events recorded"; incorrect: "the transfer was exactly 10:05." For statistical recovery methods, it is necessary to mark the assumptions separately.

7.1. Basic recording scheme

Each event should have an Event ID, a source timestamp, a normalized time, a time zone, a source, a extraction operator, an unchanged object, a degree of confidence, and a hash link. Legal endlines are considered in a separate table indicating the time of launch. This reduces the risk of misinterpretation when changing time zones or switching to summer time.

7.2. Relative time with damaged sync

If the events cannot be reliably placed on a general UTC scale, reconstruction can sometimes be done through relative sequences. For example, a queue log can record the order of transactions, and a server trace can record stable request IDs and causality IDs. These features allow us to determine that the output was formed before the automatic operation, even if the wall clocks of the respective systems differed. But the relative order does not resolve without further evidence the question of an exact absolute moment essential to the notice period.

The report should share the proven "before/after" relationship and the reconstructed time interval. In the proposed correction, the drift shall retain the reference values and the correction function. A change in time in the original log that hides the original inconsistency is unacceptable. At the technical level, you should check the source of the NTP, the configuration of the container, the cloud zones and the possible change in system time after the restart.

Practical case analysis 19.07

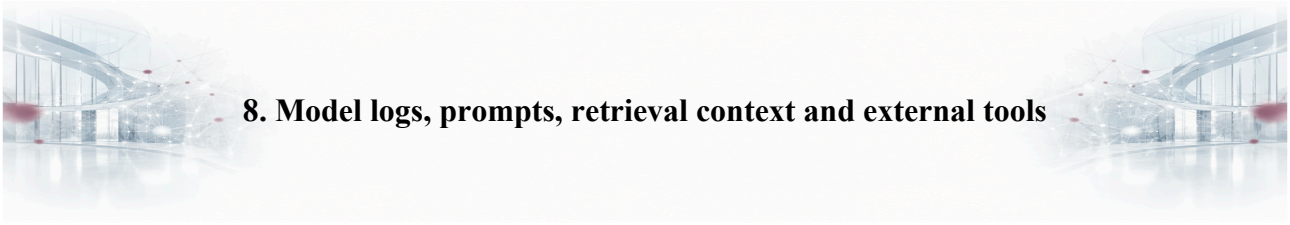
Recommendation: Normalize time by source

Hypothetical example: In the bank's logbook, the transaction is an hour ahead of the service request.

Analysis: Different time zones may create a false causal sequence.

Practical steps: The analyst retains the reference marks and builds a normalized scale with tolerances.

Outcome criterion: The order of events is justified by the uncertainty of the clock.



8. Model logs, prompts, retrieval context and external tools

It's not enough to keep the final answer to AI operation. The system and user instructions, attached documents, retrieval context, decoding parameters, tools, intermediate calls, API responses, security filters, restrictors and user rights status need to be specified. If each element is not fully disclosed for technical or legal reasons, access boundaries and alternative methods of independent verification are recorded.

Agent systems are particularly important. A single user request can trigger a chain of dozens of automated operations; the real damage occurs when the tool is triggered, not when the wrong text is formulated. When analyzing, the content of the model response, the choice of action by the planner, the powers of the tool, the work of the external service and the absence of human verification should be separated. This allows you to assign corrections at the correct system level, rather than prohibit any text generation.

The risk of a hidden change in the clue is also significant. The program code orchestration is often updated regardless of the version of the base model; after the incident, the same model ID can yield different results. "Reproduction" should therefore be understood as an operational configuration reconstruction rather than a single launch of a similar request in today's web interface.

Layer	Protected object	Model verification error
Prompt	System/developer/user message	Only the client's question is saved.
Retrieval	Corps, selected fragments, versions	Context replaced by updated base
Model	ID and snapshot supplier	The general name of the model retained
Tool	Challenges, arguments, responses and credentials	Only the success of the operation is recorded
Human	Reading, correction, confirmation	The interface log equals consent.

8.1. Re-establishment of the agent solution and retrieval source chain

It is not enough for the language system to store only the final answer. Essential objects can be system instructions, user requests, external tool calls, obtained document fragments, routing policy, model version identifier, generation parameters, and a person's decision to accept the result. In the presence of an external supplier, a separate question of evidence is whether the customer has the right to access this information. The refusal of the supplier does not allow the missing trace to be dorsed; the conclusion describes the technical and contractual reason for the lack of access.

The reconstruction should capture the boundaries of the "actual" and "experimental" path. If a researcher repeats a request for the same commercial model within a month, the result may change due to an undocumented update. Such an experiment shows the behavior of a new version available and does not prove that these tokens were created on the day of the incident. Repeatability increases with snapshots of weights, version of tools, frozen retrieval index and controlled generation seeds, but even the same seed does not eliminate all sources of nondeterminism in distributed computation.

It is particularly important to identify the points where the model made the proposition, but the decision was made by a separate system. In sanctions screening, the final blockage is often a consequence of business rules, rather than a direct "wire declaration" of AI. For each link, it is necessary to determine what actual information was received, what threshold worked, what role the algorithmic conclusion played and who had the authority to change the result. This reduces the risk of incorrectly assigning the technical component of the powers of a financial organisation or public authority.

Practical case analysis 19.08

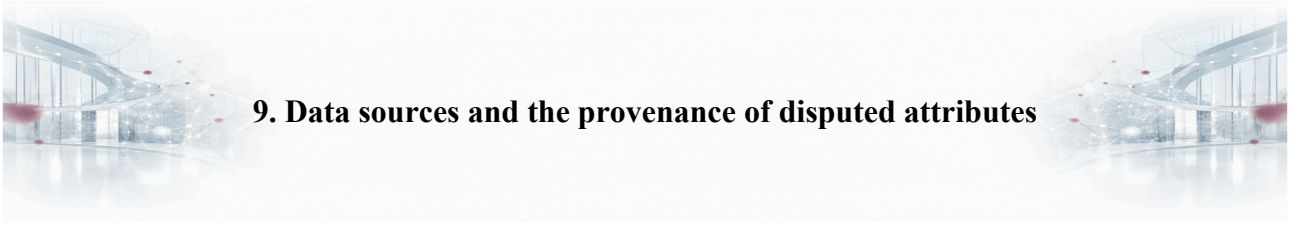
Recommendation: Maintain the full context of the response

Hypothetical example: The model's response is saved, but there is no recovered piece of the document.

Analysis: You can't distinguish a search error from an interpretation error.

Practical steps: The engineer records the input, extraction, instructions, tools, and response with available versions.

Outcome criterion: The entry and transformation chain has been re-established for the disputed claim.



9. Data sources and the provenance of disputed attributes

The system can process incorrect signals even when the algorithm is functioning according to its specification. For example, a name match with a sanction list can be reliably detected as a linear match, but misinterpreted as an identification. In this case, you need to look at the data sources, the rules of normalization, the quality of the additional identifier, the connection graph and the moment when a person makes a risk decision.

Data verification should establish the authenticity of the source, the date of the last synchronization, the permissible context of use, the history of corrections and the presence of signs of displacement. For journalistic and open source, the difference between an independent confirmation and a republication of one original error is needed. One charge reprinted by ten sites does not become ten pieces of evidence. Subsequent adjustments should keep the history of the previous entry.

Information accuracy should be separated from the legality of processing. Even factually accurate data can be used outside the authorized purpose, without adequate measures of necessity, or without an available correction mechanism. The human rights implications may be more severe than the technical error of the classification, especially when the intersection of databases automatically gives the assumption of "verified fact".

9.1. Minimum design package

The package includes a license or legal basis for access, a snapshot of the dataset ID, a fingerprint of the source line, transformations, all filters, exit signs, and a documented connection to the solution. The absence of a single link should be reported as a limitation of evidence rather than compensated by a reference to the overall reputation of the data provider.

9.2. Verification of the original register and independent truth

In a number of systems, the algorithm performs technically correct matching with an incorrectly formed source. If the list contains an outdated address or a person's erroneous connection to the company, high model accuracy does not mean actual signal correctness. Therefore, a separate line of inquiry is devoted to the provenance of the source, the conditions under which it is updated, how it is corrected and whether the end user could legally verify the alternative information. Not only is it helpful to keep a recording record, but also a known history of changes and the date of download.

Often the error is spread through secondary aggregators that include the original entry in their own risk categories. When correcting, it is necessary to establish whether the correction message was sent to these recipients, whether the old cached copies were preserved, and what actual restrictions still apply. The possibility of requesting a correction depends on the contract and the law. The independent expert records the discrepancy and the direction of the corrective chain without having the authority to update the closed state register.

Practical case analysis 19.09

Recommendation: Check the origin of the disputed topic

Hypothetical example: The algorithm found the match correctly in the registry, but the record itself contains an outdated date of birth.

Analysis: A technically accurate comparison with the wrong source doesn't establish the correct identification. A correction of the algorithm will not eliminate an error in the original record.

Practical steps: The expert keeps the version of the register, compares available primary documents and builds a chain of correction to known recipients.

Outcome criterion: The report identifies the source of the discrepancy and lists verified recipients of the adjustment.

10. Human oversight: effective supervision or merely formal approval

The availability of the "confirm" button in the interface does not prove effective human oversight. It is necessary to verify that the operator could recognize an error, obtain the original data, change the conclusion, stop execution, request independent review and reject the automatic recommendation without sanction from the organization. If the system gives out hundreds of worries per shift, the formal possibility of manual checking may not be in line with real working conditions.

Witnesses compare explanations with the interface log, the change schedule, the timing of the record on the screen and the operator's training. This is not a search for a "guilty officer" but a test of which control mechanism was actually available. It was important to examine whether the system had offered false certainty: the red flag, the exact percentage or the "high risk" category often gave the impression that the facts were established when the baseline data were not reliable.

Each manual correction and reason for the cancellation of the result should be recorded. If an organization deletes such records when learning a new model, the trace of a potentially systematic error disappears. It is practical to analyze the relationship between the number of cancellations, their causes, the speed of response and the validity of subsequent decisions.

Sign	Substantive oversight	Oversight simulation
Information	Primary facts and uncertainty	Color assessment only
Credentials	Stop or change action	Only confirm
Time	Enough to check	Mass automatic consent
Reporting	Audit of reasons for cancellation	Error written off on user

10.1. Operational signs of the real human oversight

Human in the loop only makes sense when a person gets enough data, has time to verify, and has the right to disagree with an automatic result without disproportionate organizational costs. The formal availability of the "Approve" button does not guarantee oversight. The expert asks for an interface, instructions, access records and statistics on deviations from the recommendation. With almost 100 percent confirmation of automatic solutions, you have to check not only the quality of the model, but also the process incentives, load and actual reviewing ability.

Monitoring should cover negative cases. An employee could have pressed the rule button without noticing that the original document had been cut or that the denial had disappeared from the translation pipeline. The report distinguishes between a user error and a procedure design error: if the interface hides a critical context, one employee's personal responsibility does not explain the whole reason. Improvement measures include access to originals, mandatory determination of decision motives for serious cases and selective independent audit of deviations.

Practical case analysis 19.10

Recommendation: Checking Real Human Control

Hypothetical example: The staff member confirmed the flag without the right to cancel it.

Analysis: A formal signature does not mean that it is possible to prevent a consequence.

Practical steps: The researcher studies the interface, laws and instructions; reproduces the learning dispute.

Outcome criterion: The report established whether or not a person was able to change the outcome and at what stage.



11. Causation: from a sequence of events to a testable explanation

The underlying question is about specific intervention and outcome. For example: would the signal change when the original identifier is corrected; would the blockage occur when the manual check is correct; would the recording spread during the current recall procedure? These issues are at different stages of the process. The answer to one of them does not allow the others and does not automatically allocate legal responsibility. [M2]

The hypotheses are built on a mechanism, not on a convenient number of versions. The original data, model transformation, threshold adjustment, result presentation to the person and organizational action are considered. Several mechanisms can operate simultaneously. For each person, the expected observations are recorded when they are present, information that contradicts them, and checks that can distinguish them from alternatives. A hypothesis compatible with everything possible does not give a verifiable explanation.

The configuration restoration shall be evaluated before the experiment is interpreted. If the original version, input, rules and dependencies are retained, a restart can explore a specific mechanism. If a modern model or a manually restored context is used instead, the result is a reconstruction. The degree of conformity with historical conditions is described by component. The absence of the old version does not exclude documentary analysis, but limits the experimental causal inference.

11.1. Earl of Evidence

The graph of the study contains separately observed variables and intended relationships. The "model signal" unit is confirmed by the logbook, the "opinion restriction" unit is supported by an execution record, and the link between them is the rules of routement and the material of a particular decision. Time sequence serves as a necessary part of the reconstruction, but it does not replace communication verification. The unknown transition is kept as a gap, not filled with the most plausible story.

11.2. Counterfeit testing and discipline of alternative hypotheses

For counterfactual verification, a minimum of intervention is required: correct the identifier, change one parameter or turn on a certain protective barrier. The remaining conditions shall be maintained as far as the bench permits. If the data, the model and the human procedure are replaced at the same time, the experiment shows the effect of a complex change. It does not single out the contribution of one component. Such decomposition is necessary to select a fix for the installed mechanism. [M2]

Official duties documents allow to establish the prescribed procedure, but do not prove the actual behavior of a person in unobserved circumstances. It requires separate data: logs, interface materials, consistency-checking interviews or a safe learning experiment. Retrospective explanation takes into account the information available to a person at the time of action; knowledge of the outcome is not automatically transferred to his previous situation. [M3]

Audited point	Thesis test	Refusal or restriction
The wrong entry mark caused the lockdown.	Corrected entry changes decision with other equals	The lock-off rule works independently of the label.
The model created a non-existent fact	Absence of fact in the source + saved output	The external source was not available to the researcher
The manual check was valid.	Record of access to originals and reasoned decision	Automatic confirmation only without reading
The correction stopped the damage.	No new action affected on thedownstream	Old copies of the record continue to be used

The result of the causal analysis should indicate the degree of validity of each mechanism and the remaining alternatives. A useful investigation may end in a partial result: an erroneous entry has been established and the possibility of its spread has been proven, but no decisive human step has been restored. This allows you to correct a confirmed defect without drawing a categorical conclusion about the unknown part of the event. The legal assessment shall be prepared separately, taking into account the applicable standard of evidence.

Verification of causal hypotheses in case of false sanctions match

In a conditional case, the client's operation is limited after the automatic signal. The surname and part of the name matched, but the date of birth in the original file is different from the date in the sanctions record. The researcher has access to the input data, the saved signal, and the log of restrictions; the previous version of the operator interface is not available yet. These circumstances set the test program, but they do not yet establish the reason for the final decision.

Hypotheses	Differential check	Limitation of conclusion
Wrong information came in.	Compare entry, converted entry and date source	New unloading does not replace the image at the time of the event
The algorithm didn't take into account the distinguishing sign	Check the stored configuration at the reference and corrected inlet	A modern version explores another configuration
The signal automatically caused a restriction.	Compare router rule and execution log	Convergence of time without connection of records is not enough
The manual check did not correct the error	Re-establish the displayed information and accessible action	The regulations do not prove a person's actual decision

If the distinguishing characteristic was not already in the input, the data transmission defect is confirmed. If it was available to the model, but the signal remained, it is necessary to investigate the rule of comparison and its purpose: a preliminary wide search may allow such a signal when required for subsequent verification. To explain the harm, it is necessary to determine separately why the signal led to the restriction. Various links can make a joint contribution.

In an isolated reconstruction, you can only correct the date and repeat the calculation with previous settings. The change of the signal will support the hypothesis of the relevance of this sign to the configuration. It won't prove that the operator would have historically lifted the restriction. Until the interface and materials of the manual verification are restored, this part of the explanation remains open; the conclusion does not attribute to the person knowledge of the information that is inaccessible to him.

The correction plan follows the established mechanisms: restore correct input, verify routing, ensure a substantive manual check and correct the transfers. The conclusion is confirmed separately for each action. Restarting the model with a favorable outcome does not prove that the customer's restrictions have been lifted or that there is no further dissemination of erroneous information.

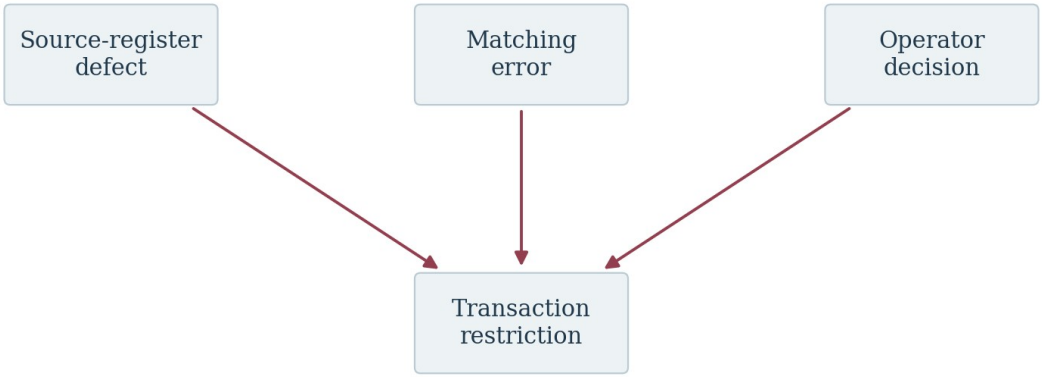


Figure 19.2. Alternative explanations for a false match. Diagram of the causal hypotheses in Section 11. Each path requires evidence that distinguishes it from the others; several mechanisms may operate together.

Practical case analysis 19.11

Recommendation: Compare Differing Causal Hypotheses

Hypothetical example: False matches may have involved an incorrect register, algorithm, or employee decision.

Analysis: The three mechanisms require different evidence.

Practical steps: The expert compares the source record, the input and output of the algorithm, the solution log; each gap is marked separately.

Outcome criterion: The underlying conclusion is supported by an observation that distinguishes it from alternatives.

12. Statistical testing, rare events and sampling sensitivity

The assessment of a serious incident requires the determination of a denominator. The allegation of "five illegal blockades" is hardly indicative of the frequency without the number of persons audited, the exposure, the period and the selection procedure. It should also be noted that customer complaints are selective: a person may not know the reasons for refusal, have no access to the complaint or refuse to submit it.

For an independent repeat test, the general set, the stratification criteria, the observation unit, the true mark, the blind mark scheme and the dispute resolution protocol are recorded. The results are given at intervals of uncertainty. Example: 0 errors detected on 100 independent tests, the upper one-sided 95 per cent limit on the "rule three" is about 3 per cent; this is not proof of zero risk. For

1 error per 1000 checks, the dot frequency is 0.1%, but small numbers require an accurate interval and clear description of selection.

The comparison of groups cannot be limited to general accuracy. In sanctions screening systems, false positives per 10,000 tests are more likely, and in medical screening, the percentage of actually dangerous cases is more likely. The choice of metrics is tied to the mechanism of harm. Multiple checks increase the risk of random conclusions, so it is desirable to fix the hypotheses and primary indicators before the analysis.

Indicator	Formula or definition	Limitation
Error frequency	FP / (FP+TN) for false alarms	We need a confirmed ground truth.
Sensibility	TP / (TP+FN)	Depends on FN detection
Incident frequency	Confirmed incidents/exposures	Incomplete registration
Time for correction	From verification to restoration	Stopping the function of assistance to a person
Zero events	0 observations at n verifications	Not proof of security

12.1. Denominator, incomplete registration and statistical distortion

The number of complaints and the number of persons affected is not equivalent to the number of independent exhibitions. One wrong penalty entry can result in dozens of blocks to various intermediaries, while thousands of people who encountered the error have no access to the complaint mechanism at all. When calculating the frequency, the denominator should be determined in advance: the number of checks, unique individuals, financial transactions, model launches or actual decisions. A change in the denominator after the result can radically change the estimate. Re-observation of one person and correlation of intra-group malfunctions are described separately.

"Zero violations in one hundred checks" does not establish zero risk. For independent trials with the same probability of a rare outcome, a rough guide to the one-sided upper 95% bound with zero observations is approximately 3/n. However, this rule of thumb cannot be applied without qualification to dependent observations, unobserved harm, or a sample assembled only from successful tests. A formal opinion states the interval used, the sampling model and the justification for that model. Where fundamental rights are harmed, even a statistically rare outcome may require immediate intervention.

The selective appearance of the event and the errors of the markings are also present. For example, the classification may appear to be accurate because the "right answer" is formed from an old registry that itself contains false records. In this case, the machine accuracy of a register is not equal to the accuracy of a legally significant fact. The verification must use an independent ground truth and, in its absence, directly qualify the unprovenness of sensitivity and specificity assessments.

Practical case analysis 19.12

Recommendation: Show statistical output base

Hypothetical example: Of the five complaints, four relate to the same language.

Analysis: The ratio itself does not determine the frequency of errors among all users.

Practical steps: The statistician checks the selection and the available denominator; only the observed shell is described.

Outcome criterion: There's no population frequency in the report without a base.

13. Signs of a deliberate attack: prompt injection, privilege bypass and memory manipulation

Not every undesired act of an autonomous agent is his spontaneous behavior. The investigation examines the external attempts to influence separately: malicious instructions in a document, page, letter, search result or response of a third-party service. Prompt injection can get into the model like regular data, but when context is incorrectly isolated, it acquires the practical power of a command. The harm in this case depends simultaneously on the behavior of the model, the architecture of the powers and the quality of control of actions.

The collection of evidence should take into account the existence of hidden instructions and intermediate calls to tools. We need a log of the source of the injection, the level of trust in the channel, what commands the agent saw and what actual actions he was able to perform. The preservation of the "most harmful text" cannot be limited: investigation requires a complete environment, including IAM permits, protection filters and alternative routes of action. The investigator must not reproduce a dangerous attack in a manufacturing environment without an authorized plan and protective restrictions.

Unlike conventional human fraud, an AI agent can execute a chain of actions faster than an organization can evaluate the first signals. Therefore, the proposed procedure involves checking time constraints, amounts, access, number of consecutive actions and the possibility of safe cancellation. Replacement of the prompt is detected at the origin level of the instructions, but damage is prevented at the actual rights level of the instrument.

13.1. Distinguishing features

Hypotheses	Relevant evidence	Alternative explanation
Injection in an external document	Full copy of source and exact prompt	Regular quote from the document
Extension of authority	IAM Logs and tool missions	Previously confirmed role
Concealment	Flats in logs and rules of rotation	Logal failure
Self-replicating	Connected sessions	Manual re-launchs

13.2. Segregation of attack and routine handling of untrustworthy material

Prompt injection is often detected only after the researcher divides the source of information and the source of instructions. If an external document contains text suggesting that the agent change the order of verification, it does not in itself prove a successful attack. It is necessary to show that the system interpreted this text as a command, despite the confidentiality limit, and that subsequent actions depended on such a transition. To do this, they record the received document, its unchanged copy, the place of inclusion in the context, the effective instruction hierarchy, the chosen tool, the query created and the actual response of the infrastructure.

An alternative to a malicious injection could be an error in the design of the interface or an incorrect policy for a reliable source. Qualifying requires caution and separation of what looked like an attack from a confirmed right-to-hand circumvention. The check shall be carried out in an isolated environment with authorized data. Publishing a detailed replicable payload capable of exploiting other systems is not always justified; the researcher may disclose sufficient signs of an attack in a confidential application to an authorized entity, and leave a safe description of the vulnerability class in an open version.

Verification area	Evidence criterion	Conclusion limit
Injection Source	Original file and way of receiving it	The source could have been changed before receiving
The Transition of Trust	Logic of the text being accepted as an instruction	The resemblance of text to a team is not enough.
Implementation	Document call and authority	The success of a call is determined by an external response
Consequences	Objective damage or transition	Attempted and actual harm differs

Practical case analysis 19.13

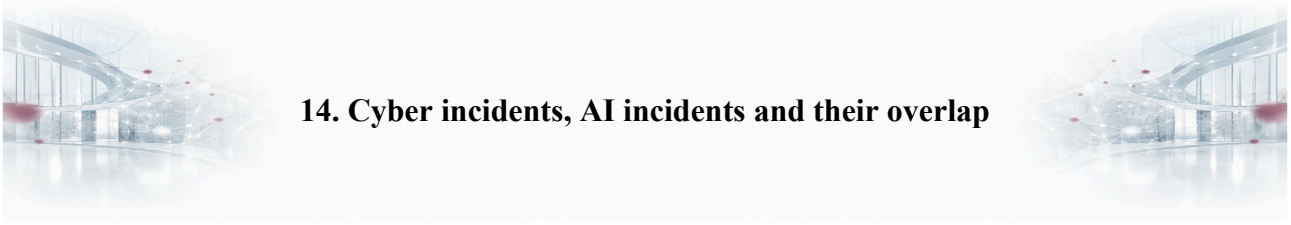
Recommendation: Check the external exposure hypothesis safe

Hypothetical example: The answer changed after processing the inserted text with a foreign instruction.

Analysis: The match doesn't establish the attacker or the intent.

Practical steps: The engineer compares harmless control options in isolation; keeps the materials and access rights.

Outcome criterion: Acquiesced or disproved mechanism of influence without unjustified attribution.



14. Cyber incidents, AI incidents and their overlap

A cyber incident can affect an AI service without the model itself causing the event. For example, API-key compromise allows a third party to initiate operations that the model honestly executes under current authority. The reverse: the discriminatory rule of scoring causes harm without compromising network security. In a mixed scenario, the retrieval-corpus attack changes the outcome, and the control system makes a decision without verification.

NIST SP 800-61 Rev. 3 offers organizational guidance on how to respond to cyber incidents. They're useful for situation management, preservation of digital artifacts, and recovery, but they're no substitute for research on fairness, model validity, and legal impact. The NIST AI RMF offers a voluntary framework for managing specific AI risks; the two documents are useful to use together, emphasizing different subjects. [5], [7]

When double-classified, two investigative flows are formed with a single register of evidence: DFIR - access, network, change events; AI assurance - model versions, data characteristics, conclusions and consequences for affected individuals. If these flows are run by independent teams, we need a single timeline and control that the isolation of the information system doesn't destroy the objects to evaluate the model.

14.1. Example of conflict of purpose

The IB team intends to immediately restore the service and remove the altered container; a panel of experts asks to freeze the environment for RCA. The correct solution depends on the threat of further harm and may include a copy to the forensic carrier, replacement of the work environment, and a ban on access to the original blind. Any choice shall be reflected in the minutes with a reasoning; refusal to retain must not be a hidden administrative decision.

14.2. Joint investigation of the technical and administrative trail

After the cyberattack, the technical team can focus on closing the vulnerability, and the legal team can focus on regulatory communications. But an AI incident sometimes occurs without a breakdown: fairly incorrect assessment and automated action based on pre-allocated powers. A reverse scenario is also possible: technical compromise affects the data, but substantial damage to humans is caused by an independent subsequent administrative decision. So incident review should include both security chain and decision chain without mixing them into one network security report.

The trusted components are analyzed separately: task order, service accounts, approval rules, integration with the payment gateway and data sources. Sometimes the main correction is not a model update, but a change in the resolution architecture or mandatory human control of the irreversible action.

Practical case analysis 19.14

Recommendation: Segregate access violation and response defect

Hypothetical example: The staff member had legal access, but received incorrect generalizations.

Analysis: The wrong answer doesn't prove that access is compromised.

Practical steps: The team will check separately the authorisation, the integrity of the entry and the quality of the transformation.

Outcome criterion: The type of incident is justified by the actual mechanism.

15. Model failure and process failure: separating five sources of defects

For objective investigation purposes, it is useful to distinguish between data defect, model defect, integration defect, institutional control defect and legal purpose defect. Example: the basic classifier correctly recognized the similarity of names; the integrator set too low a threshold; the bank automatically blocked the client; the review procedure is unavailable. There is a legal error, but it's not a statistical classifier.

Each class has different ways of testing. The problem of dataset shift requires time comparison of distributions, a model defect - control testing, an integration defect - analysis of service events, a control defect - evaluation of people's instructions and behavior, and an illegal goal - legal analysis of proportionality and authority. The hypothesis may affect several classes simultaneously; the final conclusion must specify the weight of each essential factor.

For prevention, it is important not to impose a single remedial measure "to retrain the model". If the damage was caused by the impossibility of challenging a decision, improved accuracy will not restore human rights. If the cause is in a modified retrieval source, re-training the base model won't eliminate the technical vulnerability. The corrective actions are linked to a confirmed mechanism.

Failure class	Verification test	Responsible for correction
Data	Control of origin and markings	Data owner
Model	Reproducible evaluation	Model provider
Integration	Replay Service Chains	System integrator
Oversight	Audit of authority and workflow	Deployer
Legitimacy of purpose	Test of necessity, grounds, rights	Controller/body

15.1. Matrix of the place where the error occurred and the place of the injury

Often, the defect arose in one component and the harm manifested itself in another. The vendor of the original list could have given the wrong identifier, the model correctly classified the match, the integrator mistakenly attributed the result to "final" status, and the operator automatically stopped the service. All four links should be described regardless of the subsequent distribution of responsibility. The conclusion of the "guilty model" is too broad and useless if the main measure is to correct the regulations and to access independent identification.

Level	The difference to be checked	Possible correction
Background	Are the identifiers relevant at the date of event	Rectification of recording anddownstream copy
Model	Whether output is in line with task rules	Test, restriction of application
Integration	How output is turned into action	Gate, thresholds, reversibility

Level	The difference to be checked	Possible correction
The human process	Whether source and objection are available	Change of interface and regulation

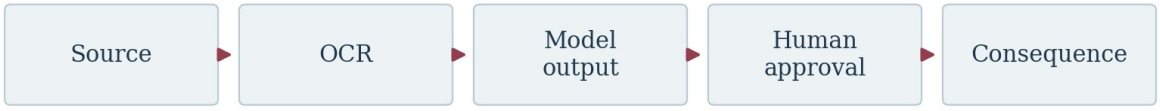


Figure 19.3. Locating a defect within the processing stages. Hypothetical scenario in Section 15: an incorrect date may arise during extraction and persist through later stages. Remediation must address the verified mechanism.

Practical case analysis 19.15

Recommendation: Localize the source of the defect by step

Hypothetical example: OCR read the wrong date, the model accurately repeated it, the employee approved it.

Analysis: The correction of the model alone does not eliminate the primary defect and failure of control.

Practical steps: The expert shall note the extraction, generation and approval; shall designate measures to be taken by each owner.

Outcome criterion: The re-checking covers the entire chain before the outcome document.

16. The EU AI Act: the legal definition and scope of a serious incident

Article 3(49) of Regulation (EU) 2024/1689 provides a distinct definition of a serious incident. It does not automatically cover every chatbot failure, inaccuracy or financial inconvenience: the specific elements of the legal definition must be examined. A breach of obligations intended to protect fundamental rights may nevertheless be legally significant without physical damage to equipment. [1], [1]

For 22 October 2026, work with the consolidated version of 27 July 2026, taking into account Regulation (EU) 2026/1744, is necessary. The second article on the original AI Act of 2024 is not a sufficient basis for a 2026-2028 legal timeline. The internal analysis must contain the date of the standard, the category of the system, the date of its placement on the market, the transitional provisions and the applicability of the exceptions. [1], [2], [2]

The legal classification may change once additional evidence has been obtained. At first, there may be a reasonable suspicion of an AI connection to harm; later it turns out that the cause is completely external. Such changes cannot be reflected by editing the original notification backwards. The original message should be preserved and explanations given for the change of conclusions should be sent.

16.1. Proposed two-pronged analysis

Contour A is the regulatory classification under article 3 (49) and special sector law. Contour B is a broader internal classification of harm for an independent Observatoire. Under B, an organization has the right to investigate harm that is below the legal notification threshold under AI Act and recommend corrections, but must clearly state that it is a voluntary analytical practice.

16.2. Applicability of a legal definition in a mixed technical architecture

A single service may include a general purpose model, a high-risk application system, business analytics, an employee interface, and an external public register. The legal characterisation begins with a description of the object and role of the subject rather than with the general word "AI". Having a language model in the chain doesn't automatically turn the entire commercial service into high-risk AI under AI Act. Similarly, classifying a particular product as high-risk does not mean that every incorrect prediction of it satisfies the definition of a serious incident in Article 3 (pp. 49).

It is proposed to draw up a separate classification record for each event: the legal entity, the basis of coverage of AI Act, the role of the participant, the place where the system is provided, the time of the event and the stage of application of the relevant provisions. At the same time, the 2026 amendments postponed the application of sections 1-3 of Chapter III for certain high-risk categories of AI (Annex III - until 2 December 2027; Annex I - until 2 August 2028). Article 73 is in Chapter IX and is not itself included in the list of deferred sections of Article 113 ((c). So it's a mistake to automatically assume that all serious incident reporting is always postponed until those dates. For a particular system, it is necessary to separate its classification, transition regime, industry legislation, participant status and legal event threshold. [1], [2]

If the same situation implies several industry qualifications, each is checked separately. A model error in medical service may require an analysis of the regimen of medical devices regardless of AI Act; a leak of personal data may result in data breaches; interference with the network may lead to DORA or NIS2 obligations. The researcher should not "take the longest or the shortest time" as a single rule. The applicability of each regime is confirmed by its norm and actual premise.

Practical case analysis 19.16

Recommendation: Check the legal threshold for facts

Hypothetical example: The organization calls any wrong response a serious incident.

Analysis: The technical error and the special legal category are not identical.

Practical steps: The lawyer establishes the subject, object, consequences and applicable criteria according to the current version of the source.

Outcome criterion: Qualifications are motivated and separately the remaining unknown facts.

17. Article 55 of the AI Act: serious incidents involving general-purpose models with systemic risk

Article 55 (1) (c) requires suppliers of general purpose models classified as systemic risk models to monitor, document and report without undue delay to AI Office and, if necessary, to national competent authorities relevant information on serious incidents and possible corrective actions. In addition, the article provides for a model assessment, systemic risk management and appropriate cybersecurity. These responsibilities apply to a specific category of suppliers, not to any enterprise that connects an external AI API. [1]

The fact that there is a large amount of damage is not sufficient evidence that the model involved is a systemic risk. The legal status of the model, the available AI Office notifications, the relationship between the supplier and the downstream deployer, and the link to the event should be checked. If the impact was caused by an integration that does not change the characteristics of the base model, the operator's own responsibility for operation is assessed independently.

For this regime, article 55(1)(c) uses the formula "without undue delay" rather than the universal numerical period of 15 days. The European Commission explains that in the context of GPAI models with systemic risk, serious cyber incidents related to the model or its infrastructure, including parameters exfiltration and attacks affecting relevant commitments, can be significant. This is an interpretation of the Commission, not a replacement for article 55. It was not permissible to transfer the time limits of article 73 to article 55. The organisation has the right to provide for a shorter internal escalation, but such an internal period should be clearly marked as a recommendation rather than a requirement of EU law. [1], [23]

17.1. Mandatory elements of an analytical package

It is recommended to record the description of the model, known hazardous capabilities, the period of the affected versions, the circumstances of the violation of the restrictions, the nature of the potential systemic effect, the checks carried out and the status of the measures. Documents containing sensitive information about vulnerabilities or other personal data are transmitted to the authorized recipient through a protected channel rather than being published in an open registry.

17.2. GPAI incident report: What exactly is relevant to the model

For a systematically risky general purpose model supplier, the researcher must consider the statutory obligations separately from those of the final application developer. The deviation may be the model itself, its distribution, or a combination with an unsafe overlay. Prior to notification of a serious incident, it is necessary to establish why the event is actually related to the Article 55 model, what is known about the systemic risk, what mitigation measures have already been taken and what facts remain unclear. A particular downstream application's imperfection cannot automatically be presented as proof of a general purpose model defect.

It is recommended that the proof pack includes model and version identifiers, detection information, technical reproduction conditions, type and extent of potential harm, affected distribution channels,

security barriers and command action. Not all of this information is publicly disclosed; some may be sent to the competent authority in a protected form. The confidentiality of commercial information, model security, rights of third parties and the regime of personal data should be taken into account. In the absence of a complete trace, the developer has the right to report the limitations of the inspection and the clarification plan without giving a working hypothesis for the confirmed primary cause.

Internal evaluation should additionally consider how a change in model weight, generation policy or routing affects the ability to compare events in different versions. Fixing one publicly known flaw can hide the original fact if the history of the deployments is not preserved. Therefore, the engineering change log and legal notification register is recommended to be synchronized via a stable incident ID.

Practical case analysis 19.17

Recommendation: Link the incident to the right model

Hypothetical example: The operator used a system risk model, but the malfunction occurred in a side converter.

Analysis: The existence of a model does not determine its causal contribution.

Practical steps: The expert checks the fault component and roles documents; the lawyer evaluates the applicable obligations separately.

Outcome criterion: The notification does not attribute the model to an unconfirmed cause.

18. Article 73 of the AI Act: reporting serious incidents involving high-risk systems

Article 73 regulates the notification of high-risk AI system providers of serious incidents in the countries where they occurred, taking into account the allocation of competences between market surveillance authorities and, where appropriate, AI Office. The basic rule requires communication immediately after establishing a causal link or reasonable probability of its existence and, in any case, no later than 15 days after the subject is aware of the norm. In the event of widespread infringement or incident covered by Article 3 ((49) ((b), immediate communication is provided for within a maximum of two days. When a person dies, there is a special rule for immediate notification within 10 days under the terms of article 73(4). Initial incomplete report with further addition is permitted. The deployer's obligations to notify the supplier and the authorities in relevant cases are separately provided for in Article 26 ((5), but their applicability and exceptions are also subject to verification. [1]

A key procedural detail: once notified, the supplier must immediately conduct the necessary research and risk assessment, take corrective action and cooperate with the authorities. The competent authority should be informed before changing the system in a way that may affect the subsequent assessment of causes. This directly supports the logic of documented object preservation proposed in this report. [1]

The terms of application after the amendments of July 2026 should be divided by specific provisions. The general date of application of AI Act is August 2, 2026; GPAI obligations came into effect earlier, August 2, 2025. The transfer to 2 December 2027 for Annex III systems and to 2 August 2028 for Annex I systems is directly established by Article 113 ((c) with respect to sections 1-3 of Chapter III (with the exception mentioned in the norm). Article 73 is in chapter IX and not in these sections. Consequently, a postponement of high-risk requirements cannot be referred to as automatic proof of the absence of an obligation under Article 73 on any date. However, the question of its actual application to a given system, including transitional provisions, classification features and the competence of oversight bodies, needs to be addressed separately; the legal matrix should record the applicable wording and the reasoned conclusion, rather than the generic 2027/2028. [1], [2]

18.1. Prohibition of the Universal Timetable

Context	Basic rule is normal.	Reservations
AI Act art. 55(1)(c)	Without undue delay	GPAI with systemic risk
AI Act art. 73(2)	Immediately after the establishment of a connection or its reasonable likelihood; maximum 15 days from becoming aware	Communication and applicability
AI Act art. 73(3)	Immediately; maximum 2 days from awareness	These special cases
AI Act art. 73(4)	Immediately following the establishment or suspicion of communication; maximum 10 days from awareness	Death and conditions of the norm
GDPR article 33	Without undue delay; if possible within 72 hours of awareness	Only a qualified violation of the PoA

18.2. Practical calculation of the date and the competent addressee

When working with article 73, they first build a legal application record. They check whether the system is high risk in the right category, whether the relevant obligations have expired, whether there is a serious incident in the defined sense and where it occurred. Then they determine who normally acts as the addressee of the duty: the supplier, in certain circumstances the operator and the relevant supervisory authority. The same event may become known to several companies on different days; it is not possible to automatically assume the date of a corporate press release as the date of their initial knowledge.

The following distinction relates to the basic and special periods. For a case that really falls under Article 73, the total maximum is 15 days, certain types of events fall under a two-day limit, and when a person dies there is a special rule with a maximum limit of ten days. These deadlines do not give the right to wait for the last day: the obligation to act immediately when a causal link is established or suspected remains in the appropriate wording of the norm. The time of completion of the technical examination cannot be considered the first point of knowledge if the documents show that the organization knew about the substantial circumstances before. Each deadline is followed by a norm, trigger event, timestamp and confirmation of shipment. [1]

Legal action	What's being recorded	Possible outcome
Scope	Specific system, role and territory	Applicable / not applicable / requires verification
Calendar	Start date of the standard for type of system	Duty not yet in effect
Seriousness	The category under Article 3(49), harm and evidence	Current class and level of confirmation
Special term	Types of event under art. 73(3)-(4)	Separate date and base
Notification	Address, channel, contents, timestamp	Confirmation of submission or proof of attempt

Practical case analysis 19.18

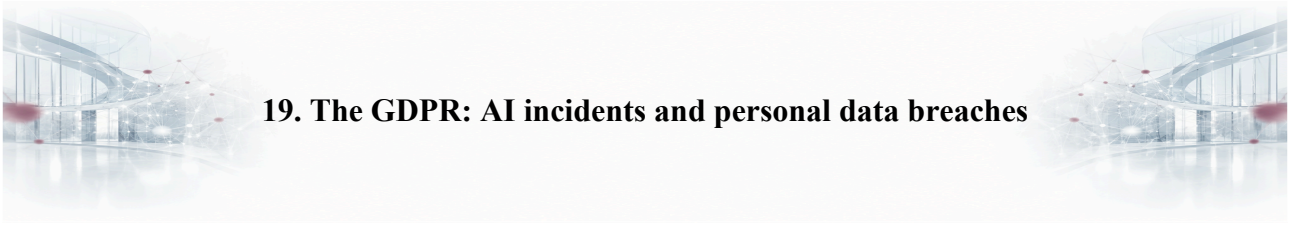
Recommendation: Segregate the initial assessment and update the notification

Hypothetical example: The result is known, but the root cause is still being investigated.

Analysis: The expectation of full certainty may hinder timely legal evaluation of actions.

Practical steps: The lawyer checks the applicability and timing; the person responsible prepares the known information and a log of the clarifications.

Outcome criterion: Copies of the messages contain the date, the facts and the identified uncertainties.



19. The GDPR: AI incidents and personal data breaches

An AI incident may result in a personal data breach, such as when an agent illegally sent an internal document to a third party or the system gave access to a third party's profile. However, an incorrect assessment of credit risk does not necessarily constitute a breach of personal data security within the meaning of Article 4 (12) GDPR. GDPR breaches cover breaches of security that result in accidental or illegal destruction, loss, alteration, unauthorized disclosure or access to personal data. [8], [9]

Under Article 33 of the GDPR, a notification to a supervisory authority is made without undue delay and, if possible, no later than 72 hours after the controller has learned of the infringement, unless it is an improbable source of risk to the rights and freedoms of individuals. Article 34 relates to informing the persons themselves of the exceptions provided for in high-risk situations. Not every leak requires public reporting, but documenting all violations by the controller remains an important obligation. EDPB 9/2022 explains the practice of applying these rules. [8], [9]

The controlr should be distinguished from the processor. A third-party AI supplier can be a data processor for one operation and an independent controller for another; a label in a marketing contract does not solve the issue without analyzing the actual purposes and means of processing. If a notification is sent under the GDPR, it does not exempt you from other applicable industry reports or from investigating individual violations of fundamental rights.

19.1. Minimum risk analysis GDPR

Determine the types of data, the number of subjects, the presence of special categories, identifiability, duration of access, possibility of abuse, means of protection and likelihood of consequences. The decision "not to notify" is recorded with motivation, person, date and sources so that it can be challenged when additional facts arise.

19.2. AI error and data break: two different evidentiary items

An incorrect automated decision may violate the principles of legality, accuracy or transparency of personal data processing without creating a data breach in the narrow sense of Article 4 (pp 12) GDPR. A reverse example: an agent mistakenly leaked confidential documents to a third party, and then in addition to possible discrimination, there is a question of privacy violation. The legal analyst must first determine whether there has been a breach of security leading to the destruction, loss, alteration, unauthorized disclosure or access of personal data. Only then will the criteria for notification and the risks to rights and freedoms be considered.

It's not just DPO messages that are important to check the date of the information. There may have been an IB alert, a customer complaint, or internal tickets, enough for the controller to have a reasonable degree of certainty about the breach. The registry records the detection, verification and basis of the assessment, and does not arbitrarily postpone the beginning of the period to the date of the committee meeting. During the first hours it is advisable to determine the categories of data, the

number of affected subjects (with a pre-marking), the potential availability of information to outsiders, the possibility of replenishment and the severity of possible consequences.

If the message has been postponed, the report should be honest about the reasons for the delay. The legal obligation to notify the supervisory authority and the obligation to notify the data subject have different conditions. Public advertising in the spirit of "we alerted all customers" does not replace evidence of specific messages or prove that the message responded to articles 33 to 34 of GDPR.

Practical case analysis 19.19

Recommendation: Check personal data separately

Hypothetical example: EI sent the wrong answer with other people's IDs.

Analysis: There's an independent question about disclosure beyond the reliability of the answer.

Practical steps: The person responsible determines the data, recipient, access and limitation measures; the lawyer assesses further obligations.

Outcome criterion: Separate decisions on the content of the error and data protection are preserved.



20. NIS2 and DORA: when an AI incident is also an ICT incident

For organizations falling under NIS2, significant incidents may require a separate sequence of reports under Article 23: early warning without undue delay and usually no later than 24 hours after notification; incident notification no later than 72 hours after notification; final report no later than one month after notification of the incident. For trust service providers, notice of a significant incident affecting the services is given without undue delay and no later than 24 hours after being notified. In the event of an ongoing incident, a report on the progress of the work is submitted instead of the final report, and the final report is submitted within a month after the response has been completed. The qualification of the subject and the implementation of the directives in national law should be checked separately. This does not automatically apply to any supplier or user of AI. [10]

For financial organisations, DORA and Delegated Regulation (EU) 2025/301 define a phased reporting of major ICT-related incidents. In general, initial notification is given as quickly as possible, no later than four hours after the incident is classified as a major one and no later than 24 hours after being informed of the ICT incident itself. The interim report shall be submitted no later than 72 hours after the initial notification, even if no changes have been made; updated interim reports shall be submitted under special conditions. The final report shall be submitted no later than one month after the interim report or the last updated interim report. If an event is recognized as a major event after 24 hours of being notified, Article 5 ((2)) sets a four-hour period after such classification. Special rules for weekends and holidays apply only when the conditions of article 5(4)-(5) are observed. The actual calendar depends on the characterization of the event, the subject and the applicable regime. The rise of fake AML research alone does not prove a major ICT-related incident. [11], [12]

A common matrix of regimes ' mutual influence is useful. It combines the registration of facts, but records separately the legal basis, the competent authority, the addressee of the duty, the beginning of the term and the required document. Different authorities may require different levels of disclosure; excessive transmission of financial or medical data through the wrong channel creates a risk.

Mode	Main trigger	You can't mix with
AI Act art. 73	Serious High-risk AI incident	Any technical malfunction
AI Act art. 55	A serious GPAI incident of systemic risk	Normal operator incident
GDPR	Personal Data Security Branch	Any algorithmic discrimination
NIS2	Significant entity of the NEX2 sphere	Any access to chat bot
DORA	Major ICT-red financial institution	Any margin analysis error

20.1. A technical match does not mean a match of the legal regime

One event may involve hacked accounts, cloud platform compromise, and a model error. This creates an object for the IB response, but it doesn't always mean that all the requirements of the NIS2, DORA and AI Act apply at the same time. For NIS2, it is necessary to establish the status of the national implementation organisation of the Directive and the criteria for the significance of the incident. DORA requires verification of whether the entity falls under the category of financial institutions, whether the event has signs of major ICT-related incident and when it was classified. The GDPR requires a separate legal assessment of personal data breaches.

It is particularly dangerous to combine deadlines mechanically. In general, NIS2 provides 24 hours for early warning and 72 hours for subsequent notification of awareness of a significant incident, and a final report no later than one month after the notification is sent; in the event of an ongoing incident, a progress report is possible. For DORA, initial communication is usually limited to four hours from major ICT qualification and 24 hours from awareness, followed by 72 hours from initial communication to intermediate and one month from intermediate to final. For later DORA classification, the special rule of Article 5 ((2)) applies, and for NIS2 trust services, a shortened notice period applies. The points of reference vary, and each must be substantiated by documents, without replacing the legal date with the subjective time of completion of the investigation. [10], [12]

Within the enterprise, we recommend appointing one event coordinator and several legal advisers, while keeping separate decisions on notification. Sending a single document through the portal of one jurisdiction should not automatically be considered as the fulfillment of all other obligations. For each shipment, the system confirmation, file version and correspondence identifier are stored.

Practical case analysis 19.20

Recommendation: Maintain a parallel map of regulatory regimes

Hypothetical example: The assistant's failure disrupted the financial process.

Analysis: One internal record does not define all possible addressees.

Practical steps: The lawyer checks the applicability of industry regimes and agrees on the actual basis of the messages.

Outcome criterion: Each applicable route has a responsible and executing log.



21. The content and stages of regulatory notification

The initial notification does not have to contain the final RCA if the special standard allows for clarification. Its purpose is to report substantial known facts, the likely nature of the consequences and the measures taken so that the competent authority can respond. The sources of uncertainty must be clear. An uncertain visibility of accuracy may be more damaging than an honestly stated unknown: estimating the number of affected persons without a methodology is not appropriate for the supervising body.

The second stage is to update the status after initial damage control and collection of substantial artifacts. The third is the final fixation of the root cause, scale, corrective measures and measures to prevent repetition. The exact composition and timing of each message is determined by the applicable regime, not by the internal universal form. For each correction, a version of the change log is created.

In information interaction, disclosure by duty should not be confused with public communication. The regulator may require personalised information on the protected channel; only anonymized conclusions can be disclosed publicly. For the benefit of the affected persons, it is important to avoid premature statements "the problem is completely eliminated" when remediation has not yet been independently retested.

21.1. The principle of true incompleteness

If a causality has not yet been established, the formula "the relationship of the system to the event is investigated; these signs are found" is usually more precise than "the error caused by AI". When sending the first message, they list separately what is known, assumed and not yet available. This maintains confidence in the reporting and reduces the need to retrospectively justify early hypotheses.

21.2. Message configuration and regulatory transparency

The first version of the message must reflect what is known at the time of dispatch. When new data is received, both the original document and the addition with a list of new facts, corrections, replacement assessments and still unknown circumstances are retained. The most important characteristic of such a package is the absence of "disappearing conclusions": if a leak was originally anticipated and it was later determined that the data had not left the Wednesday, the change should be explained in evidence. Replacing content without versioning undermines trust in the organization and makes time-audit virtually impossible.

Each authority may require its own form and channel. The universal internal file serves as a data source for different messages, but does not replace the special regulatory format. Prior to sending, they check the type of incident, the subject, the competent addressee, the exact events and the permissibility of disclosure of confidential information. Confirmation of shipment shall be kept in the same register as the content version.

Practical case analysis 19.21

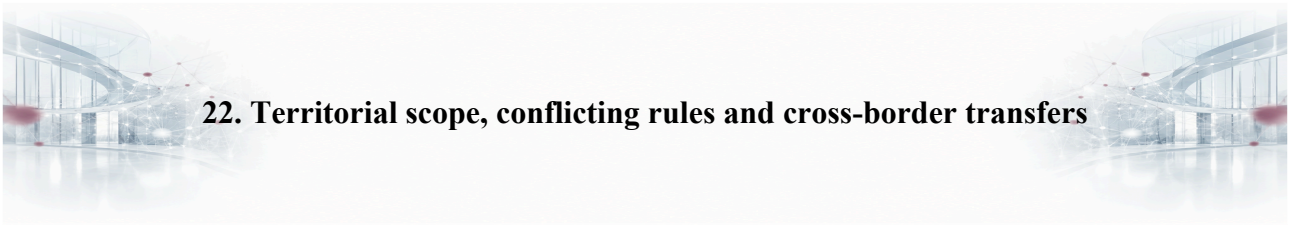
Recommendation: Report verified data with versions

Hypothetical example: In the original report, the model was suspected, and then the original error was discovered.

Analysis: An unmarked replacement of the text conceals knowledge development.

Practical steps: The focal point sends a clarification of the reason for the change and maintains both versions.

Outcome criterion: The recipient can distinguish between the original assumption and the specific fact.



22. Territorial scope, conflicting rules and cross-border transfers

One AI system can be serviced by an American supplier, processed on servers in the EU, used by a financial organization in France and harm a customer outside Europe. You can't automatically choose the right to locate a cloud server. The applicability of territorial norms, contractual relations, the location of the consequences, the status of the person concerned and the authority's competence should be determined consistently.

In a cross-border case, the prohibition of unauthorised transmission of personal, medical and law enforcement data is particularly important. A copy of the evidence cannot be freely transmitted to any consultant involved just because it relates to the investigation. First, they define the legal basis for access, the limits of authority, protective guarantees and the need for pseudonymisation. For police data, there may be a separate regime that is not identical to the general GDPR.

In case of discrepancy, the legal coordinator prepares a map to ensure that each applicable obligation is met separately. If there is an irreconcilable conflict, a local legal assessment and, if necessary, an appeal to the competent authority is required. This report does not state that an international "one window" of AI incidents already exists.

22.1. Access to evidence by an expert from another country

The international nature of the digital infrastructure creates the mistaken illusion that all materials are available to an expert regardless of geography. In practice, a record can be stored on a server in one country, contain data from another, and belong to a controller from a third. Before copying the entire set into a common archive, it is necessary to establish a contractual processing regime, legal basis for disclosure, minimum data transfer and professional secrecy restrictions. Sometimes the optimal solution is to check in a protected environment of the owner by issuing an unidentified protocol instead of mass transmission of the original data.

The cross-border transfer of personal data, the use of cloud services, the access of a foreign specialist from outside the EEA and procedural disclosure by court order are defined. Having a foreign jurisdiction doesn't automatically prohibit checking, but requires a conscious scheme of legal access. The expert assignment should specify which elements can be copied, who is entitled to read them, whether derivative data can be made, how the screens and commands of the researcher are recorded, and what should be destroyed when the legal retention period is over.

In a cross-border dispute, only the final refusal of a bank or border authority is often known to one party. The expert shall indicate which facts are confirmed by available documents and which can only be established by the legal response of the competent body. You can't imagine having access to INTERPOL's closed systems, law enforcement agencies or banks without proof.

Practical case analysis 19.22

Recommendation: Check the base of each transmission

Hypothetical example: A foreign expert asks for a full client's archive to analyse one record.

Analysis: The transfer of the entire archive may be redundant.

Practical steps: The defendant determines the necessary fragments, permissible channels and restrictions; the lawyer checks the basis.

Outcome criterion: The minimum package agreed with the recipient's log has been delivered.

23. Administrative investigations and independent expert assessment

Administrative oversight and independent audit have different objectives and powers. The market surveillance authority may require information, investigate compliance and apply the measures provided for by law; the private expert only conducts assessment within the scope of access and instruction. The expert should not give his request for documents as a compulsory public order, but a negative review by a private organization for a legally final decision on liability.

In an external assessment, the parties must agree in advance on the subject matter, criteria, access, preservation of evidence, limitations on publicity and rules for working with inconsistent conclusions. The customer's right to comment on the actual errors should not become the right to remove the inconvenient conclusion. If an expert receives a reward for a favorable conclusion, the conflict of interest must be disclosed; in sensitive projects this model is undesirable.

A separate problem is the evidence that is accessed only through a demonstration of the interface by the supplier. The expert must indicate that he observed the controlled demonstration, not replicated the operating system himself. If it is not possible to verify in depth, the conclusion is formulated as an assessment of a limited set of evidence.

Type of inspection	Authority	The precautionary standard
Internal RCA	Organizational access	Independent confirmation
Private assessment	Contract and agreement	Disclosure of restrictions
Regulator check	Applicable law	Competence and procedure
Forensics	Procedural appointment	Control of admissibility

23.1. Mandate limits for an independent researcher

An independent report may be drawn up on the instructions of a dispute party, an organisation, a professional association or a competent authority. In each case, the basis of access and the scope of authority must be disclosed. The private expert does not automatically obtain the right to confiscate information, to require a foreign bank to disclose data, or to perform process-relevant actions in

place of the investigator. The absence of such powers does not destroy the value of the study, but determines the extent to which conclusions can be drawn. Requests to third parties must be correct and accompanied by confirmation of consent or other legal grounds.

Professional independence does not mean that there is no employer. It is manifested in the disclosure of a mandate, the rejection of a pre-determined result, the documentation of negative findings and the verification of alternatives. If a specialist has previously been involved in the development of the system being tested, his knowledge may be useful for explaining architecture, but finding independence requires separate management of conflict of interest.

Practical case analysis 19.23

Recommendation: Maintain the independence of expert research

Hypothetical example: The requisitioner requests the expert to confirm a pre-selected version of the reason.

Analysis: The desired result distorts the study of alternatives.

Practical steps: The expert will agree on neutral questions, access and the right to report data insufficiency.

Outcome criterion: The report answers questions on the material without the perpetrator being asked.

24. Criminal and civil proceedings: admissibility and limits of technical conclusions

If the results of AI have affected the arrest, criminal prosecution, evidence assessment or denial of extradition, the investigation should assess the procedural significance of the source separately. A machine recommendation, created without transparent methodology and access to the original data, does not become a court-admissible evidence just because it is printed on the body's form. It is necessary to verify the procedural status, origin, reliability, possibility of dispute and access of the defence to relevant information.

In civil cases, there may be different rules for allocating the burden of proof, disclosing documents and using expert conclusions. This report does not set a single international standard of admissibility for all jurisdictions. It offers an evidence structure compatible with the tasks of subsequent local expertise: independent data acquisition, a stored storage chain, accurate description of methods and a transparent list of alternative causes.

The existence of a trade secret should not be seen as an absolute barrier to verification, but the requirement to provide a complete source code is always disproportionate. Protected examination, access controlled by the courts, verification of a particular component, output examination and independent evaluation of procedures are possible. These decisions depend on national law and the powers of the court.

24.1. Questions to the expert

What facts have been established directly, what are based on the material provided by the party, what tests can be repeated, what is wrong and what conclusions cannot be drawn from the data? The answers to these questions should be seen in the conclusion itself, not in the team's correspondence.

24.2. Expert report and evidentiary standard of procedure

A technical expert may determine that a particular model output is missing from the original source, that the log contains the specified hash, that the configuration has changed and that a certain sequence of queries is reproduced. He must not, without a special legal mandate, classify it as a crime, establish a statement of the management's actions or invalidate a decision of a public authority. Conclusions on the admissibility, relevance and sufficiency of evidence shall be subject to the procedural law of a particular jurisdiction and the assessment of the competent person.

For each essential conclusion, a table is useful: the verifiable statement, the original data, the method, the alternative explanations, the degree of uncertainty and the question that remains behind the court. For example, the phrase "log displays a match" can be confirmed directly, and the conclusion "bank illegally freezes assets because of the work of AI" requires further analysis of the legal basis and the entire decision chain. The division increases the suitability of the conclusion for a court proceeding and protects the expert from exaggerating competence.

When preparing materials for the parties, the possibility of verification by the opposing party should be maintained within the permissible limits. The confidential version may contain information about architecture and sources, while the open version may describe a verified conclusion and its limitations. If it is not possible to publicly disclose the method for reasons of safety or secrecy, this should be explicitly stated as a limitation on external verification rather than a reference to a "secret check" as an undeniable fact.

Practical case analysis 19.24

Recommendation: Distinguish technical conclusion and legal responsibility

Hypothetical example: The retest repeated the malfunction, the side demands the name of the guilty person.

Analysis: Reproduction does not establish all the legal elements of liability.

Practical steps: The expert describes the mechanism and limitations; passes the legal characterisation on to a competent specialist.

Outcome criterion: The detention did not mix observation, a causal conclusion or a decision on guilt.

25. Human rights, remedies and restoring the affected person's position

An effective investigation should lead not only to the correction of the model, but also to an assessment of the rights of the victim. If a person has lost access to banking operations, experienced restrictions on his freedom or was excluded from the asylum procedure, a successful update of the algorithm does not in itself compensate for the consequences. It is necessary to ensure that information about the decision is available, that objections can be raised, that data can be

independently reviewed, corrected and, if there are grounds for doing so, that effective compensation is provided.

The right to a fair trial, respect for privacy and effective legal remedies can be affected by algorithmic operations; qualification requires an analysis of applicable law, the role of the body and actual intervention. A particular focus is the vulnerability of a person who doesn't know that an automated indicator was involved in the decision. The lack of the ability to detect an AI component reduces the practical effectiveness of the dispute. The Charter of Fundamental Rights of the EU also provides a basis for an analysis of EU law, in particular the guarantee of effective legal remedies, respect for private life and protection of personal data. [17]

The proposed ARGA model requires two registers: technical corrections and individual restoration measures. The first is a new threshold, tests and deployment; the second is cancellation of an erroneous result, notification to third parties, correction of corresponding databases, and verification that the harmful solution is not replicated in subsequent systems. Only together do they allow the case to be dismissed.

Measure	Question for confirmation	Fake performance risk
Base correction	Are all copies corrected?	The error is left in thedownstream
Revision of the decision	Conducted by an independent person?	Automatic failure repetition
Notification of partners	Received and accepted?	The message went to the wrong address.
Compensation	By applying the legal framework?	Formal apology note

25.1. Effective remedy as an indicator of the completion of the investigation

The professional discipline of investigation requires that the harm be determined separately and the effectiveness of its elimination checked separately. For an erroneous sanction mark, it is not enough to remove the blockage in one bank if the data continues to spread through a common system of counterparty checks. For an erroneous outcome which has affected the administrative procedure, it is not sufficient to correct the internal score if the decision in the case has not been reviewed by the competent authority. The CAPA registry must contain assignments for each downstream recipient and be supported by independent evidence.

The reconstruction programme should be based on legally available means. It may include rectification of facts, consideration of objections, independent human review, notification of recipients, unblocking of documentation, removal of an illegally created record and, where applicable, consideration of damages. The organization must not promise the victim a result that can only be provided by a court, regulator or other independent body. When there is a conflict between disclosing technical details and protecting personal data, proportional access is used while maintaining the possibility of effective dispute resolution.

The criteria for closing the incident are recommended to be formulated as a set of demonstrable conditions: current harm is terminated, precautions have been taken, external obligations have been fulfilled or have not arisen, an accessible means of protection has been provided to the affected person and repeat control has been organized. Closing an engineering ticket is not the same as eliminating legal or social consequences.

Practical case analysis 19.25

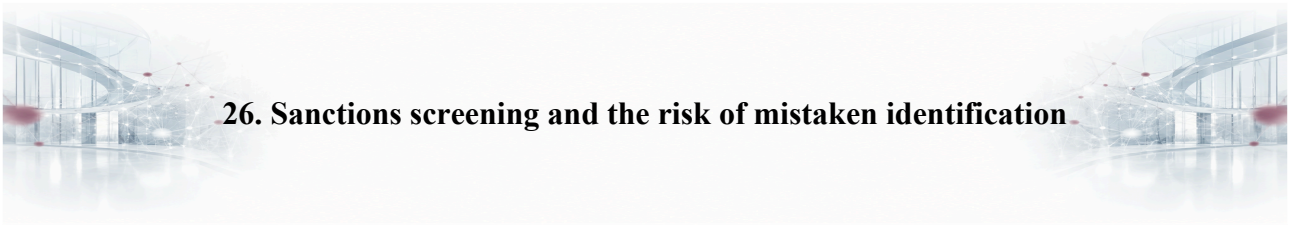
Recommendation: Provide a correction for the affected person

Hypothetical example: The error has been corrected in the model, but the client continues to be denied on the old record.

Analysis: The technical correction did not restore the client's position.

Practical steps: The Coordinator shall verify the records of the recipients and the available revision of the particular decision.

Outcome criterion: There's proof of the remedy or a reasoned description of the barrier.



26. Sanctions screening and the risk of mistaken identification

In sanctions systems, AI often does not make a legally final decision, but instead forms a signal that can have a strong impact on employee compliance or the blocking algorithm. An error investigation begins by finding out exactly what was matched: full name, transliteration, date of birth, citizenship, address, corporate role, beneficiary liability, or an external assumption of affiliation. In the absence of additional identifiers, the linear match may be technically true and legally inadequate.

There are two disadvantages: an unreasonable restriction on a particular client's operations and a systematic overlap where certain transliterations, countries of origin or long names produce excessive false matches. In the first case, the chain of decision and the possibility of correction are checked; in the second, the distribution of errors by well-chosen groups while respecting the principles of data protection. Sensitive attributes need to be treated with caution: statistics for auditing do not give general permission to collect them without a legal basis.

The report should establish whether the original sanctions list was updated, who confirmed the match, whether information was transferred to other financial institutions and what consequences remained after correction. The most important recommendation is to build downstream correction: if a client base is fixed in one bank, but the wrong name is left in the counterparty provider, the incident is not closed.

26.1. Reference Comparisons

Sign	Possible cause	Mandatory inspection material
Only full name	The Conflict of Names	Match logic and threshold
Differing transliteration	Normalization of alphabets	Change table
False corporate connection	An Old Watership Graph	Grape version and sources
Locking after manual "OK"	Automatic workflow	Audit staff member

26.2. Illustrative structure of a false-positive sanctions screening match

Let's say a learning situation: a financial organization gets an automatic surname transliteration match, after which the operating module stops payments without independent verification. The study is broken down into four chains. The first set the exact record of the public sanctions list and the date of its update. In the second, they restore matching rules, additional identifiers and confidence thresholds. The third is to see if there was any real human control and what data the employee could see. The fourth establishes the legal basis for stopping the operation, the possibility

of manual review and action to restore the rights of the client. Even if the algorithm is wrong, the legitimacy of the banking measure is not automatically derived from a single technical fact.

The expert's task is to distinguish between a name and a person's identity, to distinguish the source of the normative record and the private aggregator, and to prove exactly where the error was multiplied. Particular attention is paid to the brokerage company that could have spread the outdated ownership graph to other institutions. The report does not recommend publishing the full identifiers of third parties whose names match the lists. A documented correction and a mechanism to prevent recurrence is more important to the client than a public demonstration of all matches.

Practical case analysis 19.26

Recommendation: Check identity from different data

Hypothetical example: The names match, but the dates of birth differ.

Analysis: The automatic flag doesn't confirm the identity of the faces.

Practical steps: The specialist keeps both sources and checks the additional identifiers until the final qualification.

Outcome criterion: The decision was motivated by a set of data, and the erroneous association was cancelled.

27. International police cooperation, wanted-person alerts and extradition

Studying an AI incident in international search should not be based on the assumption that INTERPOL automatically makes decisions based on a particular model. The applicability of algorithmic processing and the specific architecture of the system is established only by evidence. However, the overall risk is that an inaccurate machine match will be transmitted between jurisdictions, where a separate national authority will decide on detention.

Investigations investigate the origin of the police report, the status of the information, the history of updates, the existence of a correction request and the actual use of the data by the state. The subject of a separate test was whether the machine communication could be taken as an officially confirmed basis, although it contained a limited hypothesis. The technical match does not override the requirements of necessity, accuracy, competence and procedural guarantees.

A legal defense strategy may include challenging a national decision, requesting a source, using existing mechanisms to correct police data, and preserving evidence of cross-border error. The claim that one body can cancel all records of all states by one decision would be false. Independent AI expertise here helps to determine the origin of the error, but it does not replace special procedures in national and international law.

27.1. List of evidence for defence

The original notifications, the documents on detention or refusal of entry, the identified identifiers, the official reply on correction, the internal access logs (if legally available), the time scale and

information about the decision-making authority are required. In the absence of access to closed systems, the examination directly indicates procedural constraints.

27.2. When a technical match interacts with a police record

In international searches, the technical identification of a person's resemblance and the legal obligation of the body to carry out a check are not equivalent. The report should divide the availability of verified reports, data from the national law enforcement agency, the arrest decision, the court's assessment and the alleged contribution of the algorithmic tool. From the fact that a person was stopped at the border, it is impossible to conclude reliably what type of international notification was active or what software was used. Access to limited materials is only possible through proper procedures, and an independent researcher does not substitute official answers.

The evidence package in the training case includes identity documents, records of border action and its time, issued procedural decisions, known communications from competent authorities and legally obtained information about the source of the match. A technical hypothesis can only be verified by accessing the rules of comparison or verified procedural materials. Homonyms, transliteration, obsolete records, routing errors and misinterpretations of data by humans are considered separately. The consequences for human freedom require the immediate use of available legal remedies, even when the digital forensic is not yet complete.

Practical case analysis 19.27

Recommendation: Do not replace official search with an algorithmic reference

Hypothetical example: The Assistant has compiled a summary of several informal publications.

Analysis: It does not confirm the existence or content of an official record.

Practical steps: The lawyer shares publications and official materials, asks for proper confirmation by procedure.

Outcome criterion: The entry indicates the valid status of each source.



28. AML/KYC: automated suspicion and its consequences for customers

The AML model is able to legitimately identify questionable patterns and simultaneously make systematic mistakes. In denial of service or excessive freezing cases, an organization's obligation to comply with AML rules, a prohibition on disclosure of individual transaction reports, and the client's right to fair, non-discriminatory and proportional treatment under applicable law should be distinguished.

Determining an error requires analysis of the model version, a list of features, source data, rules for changing risk score, analyst actions, decision time, and legal action. There should be no public disclosure of details that the law specifically protects; a closed independent assessment with a limited circle of admitted experts is possible. The questions of the validity of the decision and the possibility of appealing against it shall be assessed within the framework of the relevant State.

Of particular importance is the verification of mass failures, which are not reflected in one "large" incident: hundreds of small, erroneous blockades can affect more people than one visible incident. The individual indicator shall describe the frequency of repetition under identical conditions and the terms for restoring service after justified objection.

28.1. Mass AML-escalation as a controlled management malfunction

When thousands of operations suddenly become suspicious after the model is updated, independent analysis must take into account not only precision metrics but also the actual distribution of labor. It is necessary to determine the changed baseline frequency of signals, the proportion of repeat customers, the change in thresholds and employee burden, and then to determine whether automatic refuses have replaced the intended manual review. The number of signals alone is not equal to the number of laundering cases or proven customer violations.

In a separate matrix, the model version, the comparison period, the ground truth criteria, the number of independent operations and the number of affected subjects are recorded. You can't compare the new stream to last month without a change in seasonality, a change in client portfolio and an external change in legislation. If an erroneous system has caused long-term account closures, the legal basis and availability of the dispute procedure should be checked rather than limited to returning false positives to their original value. Corrective measures include data recovery, independent re-verification of particularly difficult decisions and a transparent process for correcting system errors.

Practical case analysis 19.28

Recommendation: Segregate Suspicion from an established violation

Hypothetical example: The AML model highlighted an unusual operation, and an employee called it a proven crime.

Analysis: The risk indicator does not specify the nature of the offence.

Practical steps: The reviewer corrects the qualification, keeps the original data and checks the basis for the subsequent decision.

Outcome criterion: The material distinguishes the signal, verified facts and legal evaluation.



29. Algorithmic credit scoring, insurance and private-sector decisions

The case of unjustified denial of credit requires an accurate separation of the default probability assessment from the bank's final decision. The outcome is influenced by limits, income, industry rules, human consent, and sometimes external anti-fraud checks. If an organization uses a third-party model, the investigation analyzes what factors have been passed on to the operator and whether the operator is able to verify errors or appeals.

To assess discriminatory impact, you need to choose legally permissible comparison groups, adjust important mixing factors and not just give average results for established discrimination. When preparing evidence, it is desirable to show not only disparity but also a possible causal mechanism: inaccurate proxies, historical data selection, misused personal traits or unstable group logic.

Technical metrics are particularly unreliable without evaluating failures that have not reached customer turnover. The incident register should take into account third party requests, repeat requests and feedback from the operational unit. If the bank changes the threshold after the complaint, the previous edition is recorded and an independent re-purchase is made, otherwise the evidence disappears.

29.1. Automatic decision-making and private legal effects

In lending and insurance, complexity often arises at the border of business model and rules. The algorithm can return the risk, and the contractual or management threshold turns the score into a rejection. To calculate the score, you need to get the value and interpretation of the policy at that date, the completeness of the original data and the history of objections available to the client. The technical explainability of a prediction is not always the same as a legally sufficient explanation of a decision. If personal data processing is used, the applicable requirements of the GDPR are analysed separately, including under appropriate conditions Article 22.

An erroneous scoring can have consequences outside of one product: a risk record is transferred between units, affects another account or is reproduced later. Therefore, the expert should check the map of downstream usage and the availability of the correction. The legal opinion does not assess a "bad score" in a vacuum, but a specific action and its basis.

Practical case analysis 19.29

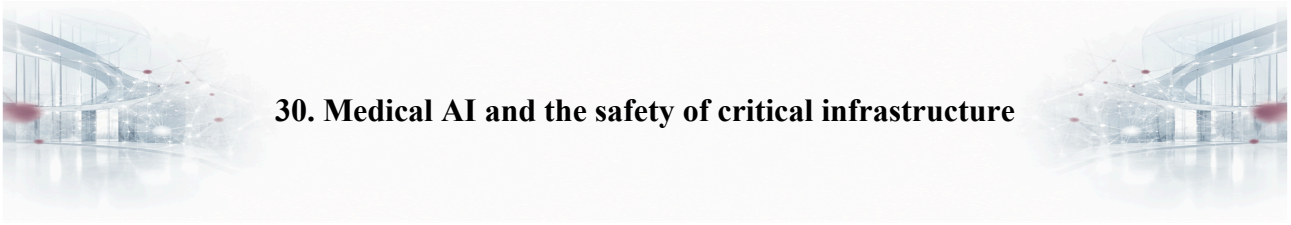
Recommendation: Check Sorting Entrance and Review

Hypothetical example: The denial of credit is based on an outdated income record.

Analysis: Model balance review doesn't correct the wrong entrance.

Practical steps: The person responsible checks the record with the client's documents and arranges for a re-examination.

Outcome criterion: Decision reviewed with corrected entry; result recorded.



30. Medical AI and the safety of critical infrastructure

In medical practice, an investigation should not be limited to the analysis of algorithm accuracy. It is necessary to find out the clinical situation, the role of the doctor, the accuracy of the input data, the limitation of indication, user instructions and patient characteristics. For medical devices, the EU has special sectoral regimes, including its own vigilance procedure. AI Act provides for coordination with them and exclusion from the usual reporting regime of Article 73 in certain circumstances. [1]

In critical infrastructure, the priority is physical system security and continuity of vital service. The operating model cannot be deactivated arbitrarily for the sake of full memory preservation if it will exacerbate the threat. The organization must develop a safe isolation regime and an alternative route of administration in advance so that evidence preservation and human protection are not incompatible.

The error of self-confidence is a particular danger. A high probability indicator can only represent the internal calibration of the classifier and does not guarantee clinical reliability of the conclusion. An expert opinion must reveal the sample, target population, limitations and what conclusions are not transferable to a particular patient.

Practical case analysis 19.30

Recommendation: Prioritize prevention of continuing harm

Hypothetical example: In a critical process, a reproducible error has been detected.

Analysis: A long-term study should not leave a dangerous function without control.

Practical steps: The operator enters a secure restricted mode and stores the materials; the profile team investigates the cause.

Outcome criterion: Reduction of current risk and investigation have separate responsibilities.

31. Synthetic documents, fabricated media and evidentiary errors

Generative models can create convincing documents, sound, and images. You can't declare the material fake in an investigation just because the automatic deepfake detector gave a high rating. The sensitivity of the detector depends on the format, compression, new generation methods and the frequency of false starts. On the contrary, the absence of fusion marks does not prove genuine.

Technical verification involves obtaining the original file, analyzing the available metadata, transmission history, independent confirmations of time and place, and, if necessary, expert methods of media forensics. In criminal prosecutions, the principle of competence requires that the party has the opportunity to challenge the method, control sample and limitation of the evaluator. For documents that may have been converted several times, the absence of an EXIF is normal in itself and does not amount to forgery.

The practical scheme includes three independent branches: the origin of the carrier, the consistency of external circumstances, and the technical expertise of the contents. Even if one component is unavailable, the others may have independent evidence value. The final conclusion should not replace the probable technical conclusion with a categorical claim of authorship.

31.1. Authenticity of synthetic files and transformation chain

In cases of suspicious audio, video and textual material, it is not possible to carry out an examination solely on the basis of the probability given by a commercial "dipfeuve detector". Information is required about the file's origin, device extraction, control amounts, transmission history, encoding, and possible platform conversions. Even a confirmed use of a synthetic instrument does not automatically mean that a single fact in the record is false. Conversely, a document with no signs of synthesis may contain a real scene shot in the wrong context or assembled with a distorted chronology.

When comparing the record with investigative and judicial material, the expert makes a separate list of claims: what is seen or heard directly, what is assumed from the context, and which elements are confirmed by independent sources. The unknown degree of accuracy of the detector is clearly indicated; the results of different instruments should not be mechanically averaged. For important decisions, an original carrier, an independent media expert, and procedural evidence of alternative origin are recommended. The service label "AI generaled" does not in itself constitute a judicial finding of forgery.

Practical case analysis 19.31

Recommendation: Check the origin of the disputed media file

Hypothetical example: The record looks convincing, but only the copy sent is available.

Analysis: The appearance does not confirm the authenticity of the author either.

Practical steps: The expert asks for the original and origin, records the transformations; when the data is lacking, limits the inference.

Outcome criterion: The degree of confidence is consistent with the primary information available.

32. Independence of expert organisations and conflict management

AI developer is interested in demonstrating the limitations of the incident; the injured party - in confirming the claimed damage; the public authority - in protecting the legitimacy of their own decisions. An independent researcher cannot ignore these incentives. Before appointing the experts, they draw up a register of interests: commercial relations, participation in system development, remuneration, provision of computational resources, publication and communication with interested institutions.

The degree of independence can be complete, limited or insufficient for a particular issue. For example, an expert who has previously conducted model tests can explain his own methodology, but his independence in assessing the quality of these tests is limited. The working group should have the right to document dissenting opinions, retain the original materials and identify obstacles to access, even if the conclusion is uncomfortable for the customer.

Practically, a two-tier check is offered: the first expert reproduces the technical scenario, the second assesses the compliance of the process with the evidence criteria. A human rights expert's review is particularly useful when a formally correct statistic conceals a serious individual consequence.

32.1. Prohibition of false institutional attribution

Observatoire ARGA's conclusions should be disseminated as an expert study by a private organisation, unless otherwise confirmed by a separate official act. It cannot be defined as a decision of the European Commission, the UN, INTERPOL, a national court or a certification body. Sending an opinion to the consultation does not mean that the regulator has accepted the conclusions.

32.1. Limitations of institutional name ARGA

It is important not to give up an independent research organization as a government agency, official inspection or state-accredited laboratory without proper documentary status. Having expert experience, participating in international consultations and analytical publications is not the same as being a procedural expert in a particular case. For each document, the authors, the technical mandate, the disclosed limitations, the nature of the review and the final legal conclusions are indicated to the competent specialist.

When submitting methodological proposals to other institutions, it is correct to speak of a potential standard of comparability rather than of "binding international ARGA rules". This increases professional confidence and allows the materials to be used as an analytical basis in the development of internal and public procedures.

Practical case analysis 19.32

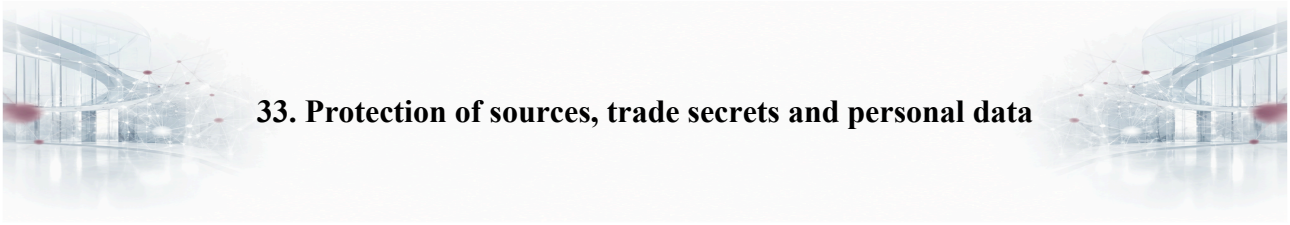
Recommendation: Unlock the conflict and appoint an independent reviewer

Hypothetical example: The expert previously advised the vendor on the disputed component.

Analysis: There is a risk of self-assessment.

Practical steps: The coordinator documents the relationship and reports the matter to an independent specialist.

Outcome criterion: The decision on the component was not approved by his previous consultant.



33. Protection of sources, trade secrets and personal data

Important evidence can include medical maps, criminal prosecution records, internal anti-fraud models, personal correspondence and critical infrastructure vulnerabilities. The investigation should be based on the principles of necessity, minimization and delimitation. It is not possible to form a "full expert archive" of other people's data without a legal basis and a specified duration of storage.

Three levels of access are recommended: worker, limited expertise and public. The work contains descriptions without any additional identifiers, the protected one contains original and verifiable artifacts, and the public one contains only edited facts that do not allow a dangerous attack to be recreated or the victim to be identified in violation of his rights. You need to keep proof of the editing itself: the original hash, the conversion log, the list of removed fields, and the person's authority.

A particular problem is the involvement of external AI services in the preparation of the investigation itself. Transferring a complete closed file to a generative tool requires an independent assessment of the legal basis, contractual guarantees, data retention and confidentiality; the model may erroneously summarize materials and should not be the only expert or evidence keeper.

Practical case analysis 19.33

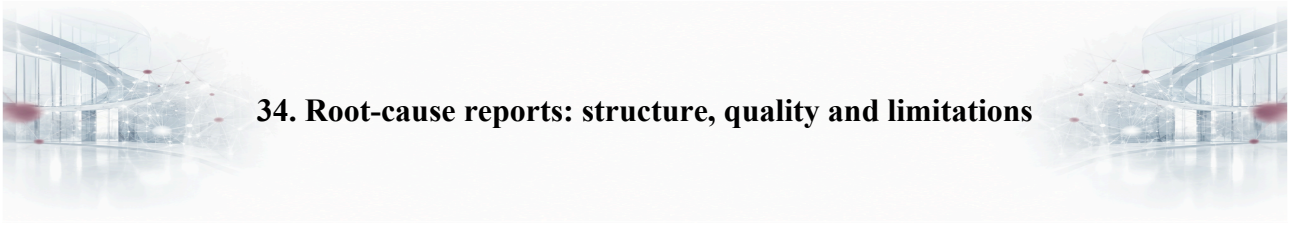
Recommendation: Protect sources when transmitting the result

Hypothetical example: The public report contains details to enable the complainant to be identified.

Analysis: The removal of a name does not eliminate indirect identification.

Practical steps: The editor checks the combination of dates and circumstances, prepares a closed proof version.

Outcome criterion: The public text retains the conclusion without unnecessary disclosure of the identity.



34. Root-cause reports: structure, quality and limitations

The main report is based on research issues. For each, an established observation, primary source, verification performed, substantial alternative and limit of inference are given. The associated corrective actions and criteria for re-checking are then described. Such a structure allows us to see what measures follow from the confirmed mechanism and what measures are proposed to reduce uncertainty or prevent a similar scenario.

For each conclusion, a marking is applied: F - established fact; E - expert assessment; H - working hypothesis; U - unknown. The source's digital identifier links the text to the experience profiler. The report should allow the third-party verifier to reconstruct the course of reasoning without personal acquaintance with the author. If the external counterparty has not provided part of the data, the information about it shall be recorded in the list of missing evidence.

The root causes should be divided into immediate and systemic causes. The closest reason could be the wrong answer to a particular model; the systemic one is the lack of a testing policy, unchanging logs or the ability to manually cancel. Technical patches alone are not sufficient to prevent repetition. Both categories are described separately.

RCA section	Contents	Quality control
Factual findings	Only F + source ID	No unconfirmed quotations
Hypotheses	H1, H2, H3 and bases	Alternatives taken into account
Conclusion on the mechanism	Relationship of action and implications	Missing links opened
Limitations	Access, sampling, conflict	Not hidden behind a careful phrase
CAPA	Action, owner, test	The result is measured.

34.1. Protection of the report from circular confirmation

The independence of confirmation is assessed by the origin of the data. The model log, which is automatically compiled from the resume, and the expert text that uses only that resume, remain one information chain. For the key conclusion, they look for material obtained through another mechanism: an external transaction record, a source document, an independent time stamp or a evidence with established restrictions. In the absence of such material, source dependence clearly limits confidence.

It is proposed to separate the facts, results of the technical inspection, causal interpretations and legal conclusions. The term "confirmed" is supported by the basis: the origin of the record, the reproduction of the test, or the combination of agreed materials. Unresolved contradiction is shown near the conclusion. Recommendations do not increase the degree of its validity; they determine the effect on available data.

Level	Acceptable wording	Undesired switch
Fact	" The magazine has a record of time after time... "	"The system has objectively done..."
Test	"In this environment, the result is reproduced..."	"That's what happened in all versions..."

Level	Acceptable wording	Undesired switch
Hypotheses	"The data are consistent with the version..."	"The guilty person has been found..."
Legal conclusion	"The applicability of the regime requires verification..."	"Any incident automatically violates the law..."

Practical case analysis 19.34

Recommendation: Build a report around a proven causal chain

Hypothetical example: The conclusion lists the documents, but does not explain the transition to the conclusion.

Analysis: The list of sources does not show which hypothesis is confirmed.

Practical steps: The author links each conclusion to observation and alternative; he lists the limitations separately.

Outcome criterion: The reviewer reproduces a chain of argument on the material.

35. Corrective measures, retesting and resumption of operation

The corrective measure must be linked to a confirmed mechanism or a clearly indicated precaution. If the wrong identifier came from the source, changing only the model threshold does not correct the source. If the correct signal has turned into an automatic lock, check the execution rule. For each measure, a variable component, expected observation, independent success criteria and possible side effects are specified.

A correction check includes a baseline scenario, a new set of similar cases, and situations where the measure is likely to worsen the outcome. Repeating only a known example confirms the elimination of the reproducible defect on it, but not the transfer to other cases. To change the threshold, false signals and passes are compared; to confirm newly by a person - timeliness and actual possibility of refusal. The settings are separated from the final check.

The Go/no-go decision is issued by the competent management body and includes open residual risks. The formula "critical errors not found" does not mean that there is no risk; the solution must be based on exposure, target application, sampling uncertainty and control measures in the future. In the event of a fundamental rights threat, a separate legal assessment may be required, independent of the commercial need to run the service.

35.1. Review of the impact of corrections and prohibition of cosmetic CAPA

The impact of the measures is assessed according to the mechanism and the impact. The technical criterion answers the question of whether the defect has ceased to reproduce under given conditions. Procedural - whether the new check is being performed in the flow. Recovery criteria - whether the adverse action has been stopped and whether the associated records from the established recipients have been corrected. The fulfilment of the first criterion does not automatically mean the fulfilment of the others.

The resumption decision indicates the verified configuration, criteria met, remaining restrictions and signal for reintervention. The tester is separated, as far as possible, from the corrected command. When independent verification is not available, this is reflected as a limitation. Observation after resumption is planned on the basis of risk and available exposure; a single favorable case does not establish the sustainability of the effect.

Practical case analysis 19.35

Recommendation: Confirm the corrective measure with a repeat test

Hypothetical example: The supplier reports that the error has been corrected, but only presents a new instruction.

Analysis: The declaration of change does not confirm efficiency.

Practical steps: The engineer repeats the initial scenario and close control tasks; the operator checks the approval process.

Outcome criterion: Resumption is permitted after a documented verification result.

36. Recording losses, compensation and correction of third-party data

The consequences of an AI incident are often divided between direct and indirect losses: cancellation, credit cost, legal costs, reputational damage, deteriorating health, freedom restriction, and violation of procedural law. An AI expert is not entitled to independently assign a monetary equivalent to all consequences without a legal and economic assessment. The damage report should distinguish between documented losses, justified estimates and disputed claims.

For the organization, it's useful to fix the cost of repairing its own systems separately from compensation to the victim. If remediation is included as an "incident expenditure", it does not show whether people have received a real remedy. In the case of cross-border distribution of erroneous data, it is important to confirm that the data has indeed been corrected for each recipient and that reuse has ceased.

Closure indicators: the proportion of confirmed errors that were fixed on time; the proportion of downstream recipients that confirmed correction; the proportion of people who were reviewed outside AI-conveyor; the number of repeated unfounded rejections; the remaining open requirements. These indicators relate to the proposed management methodology and are not automatic legal criteria.

36.1. Division of material and intangible consequences

The investigation should not automatically reduce the harm to the amount of the direct financial loss. Detention with incorrect identification, restrictions on access to medical care, loss of reputation and the inability to challenge an automated decision can have significant consequences without immediate monetary value. The registry records the nature and duration of the impact, documented expenses, restrictions on freedom, administrative procedures and recovery measures. The legal assessment of compensation requires an indication of the applicable material and procedural law.

For comparison, it's useful to separate the documented damage from the likely future damage and the cost of preventing a recurrence. The company's own audit costs are not automatically compensation to the victim. The execution of remediation measures requires proof for the person concerned.

Practical case analysis 19.36

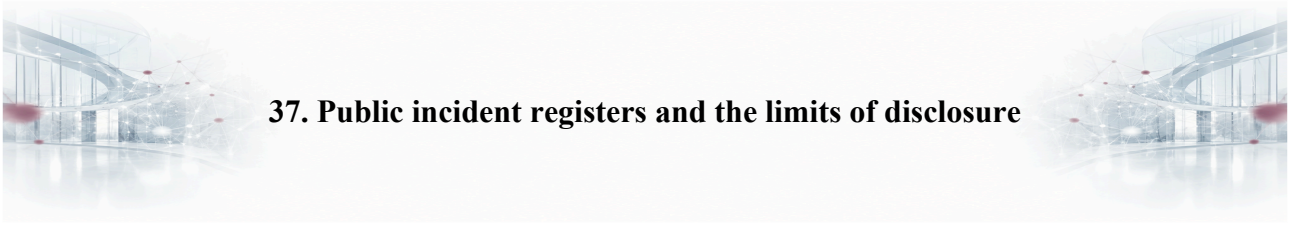
Recommendation: Take into account direct and subsequent effects

Hypothetical example: The wrong recording caused the rejection, and the copy was left with the partner.

Analysis: Correcting the local base does not eliminate the secondary effect.

Practical steps: The coordinator draws up a map of recipients and proofs of correction; records the losses according to the documents.

Outcome criterion: All known recipients have been vetted; the disputed amounts have not been disbursed for established amounts.



37. Public incident registers and the limits of disclosure

A public database is necessary to accumulate knowledge, but its structure must take into account incomplete reports, legal prohibitions on disclosure and the risk of repeated identification of victims. The OECD AIM demonstrates the possibility of international classification and linking news reports to events; at the same time, its methodology does not make the media a complete official list of all incidents. The development of a public register on the basis of one category of sources requires a clear indication of sampling bias. [4]

Two related sets are proposed: a closed evidence register with full originals and a public incident abstract with anonymized facts. The public version contains the system type, the threat class, the source assessment information, the impact type, the investigation stage and the correction status. No case numbers, personal data, access tokens, or technical instructions to facilitate abuse are published.

The register should not give the impression that the included publication proves legal guilt. For each record, you need to show the level of verification, the date of the last change, the type of source, a link to the public material, and if denied, the saved correction record. Closing the case means achieving the criteria of the investigation, not automatically missing the human right to continue the appeal.

37.1. Protection against statistical illusions of public reporting

The public register, built on voluntary reports and news, inevitably reflects not only the frequency of incidents, but also the intensity of the press, the availability of journalists, and the incentive for organizations to report on problems. It can't be taken for total AI-incident statistics. It's helpful to show the level of confirmation, the primary source, the type of damage, whether there's an independent verification, and the date of the last update. Disputed events can be accompanied by an unconfirmed causal link.

General transparency does not require the identification of the victims or the operational details of a hazardous failure to be published. An open record must contain sufficient aggregate information for public control and have a legitimate policy of error correction. Unverified machine-compiled resumes cannot be considered as equivalent to the official report of the regulator.

Practical case analysis 19.37

Recommendation: Publication of status rather than unfounded accusation

Hypothetical example: There's a client's address in the register, but there's no confirmation of the architecture of the system.

Analysis: Public categoricality can get ahead of evidence.

Practical steps: The editor signs the message, the confirmed event and the unresolved issues separately.

Outcome criterion: The record makes it possible to understand the evidentiary status without reading a closed file.



38. International notification exchange and confidential coordination

The same incidents can be relevant to different jurisdictions: AI Office, national market surveillance, the DPA, the cybersecurity authority, financial supervision and the judiciary. However, a specific notification must be given on its own legal basis. ARGA's proposal to create an interagency coordination point should not be described as an existing universal international mechanism.

For unification of communication, a neutral fact-based package with event, digital identifiers, impact, security measures and version is suitable, from which legal specific applications for each addressee are formed. Such an architecture reduces the risk of mismatching facts: for example, in one message, you can name ten affected people and in another, a thousand without explaining the difference in the calculation methodology.

When a request is received from another country, the specialist first checks the authority, purpose, protected channel and legality of the international transmission. Commercial promises of "full access by regulators" do not alter the rules of bank secrecy, personal data and attorney's secrets. In a conflict of access and confidentiality, proper procedure or court clearance is required, not informal file sharing.

38.1. International coordination and control of duplication of notifications

In a single cross-border case, it may be necessary to refer the matter to several independent bodies. For this purpose, each message is assigned a separate number and a map is kept of what facts are transmitted and in which version. Differences in legislative terminology are recorded separately: a public authority may use its own notion of significant incident, which does not match the definition of AI incident OECD. Even the correct sending of a message by one organization does not automatically confirm receipt of it by all appropriate recipients.

The exchange of documents shall take into account the limitations of international data transmission and the powers of the person initiating the contact. If an authority has replied that a question is beyond its competence, such a response is retained and used to correct routing without disguising it as confirmation of the substance of the expert conclusions.

Practical case analysis 19.38

Recommendation: Agree on the facts of an international notification

Hypothetical example: Two partners sent different descriptions of the same event.

Analysis: Different versions can cause erroneous decisions by addressees.

Practical steps: The coordinator uses a single identifier, a coordinated chronology and a log of updates.

Outcome criterion: The recipients have an up-to-date version and corrections.

39. The independent Observatoire ARGA methodology: five quality standards

The proposed independent evaluation programme is based on five requirements: traceability of each material fact, reproducibility of the methodology, legal separation of regimes, proven independence of verifiers and measurable correction measures. These requirements may be used for ARGA's voluntary expert opinions, analytical consultations and submissions to regulators, but do not establish a new legally binding form of certification.

The procedure is opened by a qualification memorandum: subject, purpose, parties, conflict of interest, data security, victim safety and preliminary competence map. Then a register of evidence with access levels, a test program with prohibition of exploitation in production without sanctions and a legal obligation matrix is formed. The result is not one "pretty PDF" but a tied package of source register-analyses-remediation-review.

It is proposed to publish the methodology in the form of an open quality standard with emphasis on correct use of terms and control of uncertainty. The ARGA should not portray itself as an official oversight body or authorised to access classified public information. This approach increases the suitability of materials for professional discussion with institutional partners.

39.1. Internal control phases

Phase	Withdrawal	Completion criterion
Intake	Primary record	Triggers and safety evaluated
Preservation	Evidence register	Sources and access recorded
Investigation	RCA + legal matrix	Alternatives considered
Review	Independent review	Objections resolved
Closure	CAPA + Recovery	Performance measured

39.2. Independence and repeatability of the expert panel

Observatoire ARGA can use a five-component quality procedure: source registration, digital integrity protection, comparison of alternative hypotheses, separate legal assessment and recovery monitoring. Each block is required to be checked by a person who did not make the original

technical decision or is not involved in evaluating their own results. A conflict of interest may arise not only with direct financial involvement, but also when the expert is dependent on the model provider, the client organization or the expected outcome. The conflict detected is disclosed; not every conflict automatically excludes participation, but it should be manageable and understandable to the reader of the conclusion.

Before the external transmission of the report, it is recommended to go through a double check: the technical specialist checks the reproducibility and limitations, the lawyer - the legal coverage, dates and addresses of the obligations. The lists of sources and endorsements shall be machine readable as far as possible without disclosure of confidential data. If an error is detected in a previously issued conclusion, only a version with a clearly marked correction and preservation of the previous history is allowed. Silent substitution of a document without a change mark is contrary to the very idea of traceability.

Practical case analysis 19.39

Recommendation: Check five elements of the quality of the investigation

Hypothetical example: The file contains conclusions and annexes, but not a map that loses alternatives.

Analysis: The external reviewer sees no limitations on the analysis.

Practical steps: The curator checks the origin, causation, law, independence and enforceability of the related records.

Outcome criterion: Each element is supported by a document or a specified restriction.



40. Contractual architecture for expert investigations

The agreement with an independent researcher determines the subject of the verification, the legal basis for access, the amount of confidentiality, the storage and return of artifacts, the participation of specialists, the limitation of liability, public use of the name and the right to an objective conclusion. A separate annex sets out which party is the controller of personal data, which transfers are permissible and which documents the supplier undertakes to keep.

It is not recommended that a contractual clause permits the customer to change the actual findings at his discretion without hindrance. A fact-checking procedure is permitted where the customer objects to the source; the reviewer records whether it has been accepted and why. When the work is completed, the original report and the log version are kept so that new editions cannot be postponed to the first.

For international cases, the treaty takes into account the possibility of applying national rules on confidentiality, sanctions restrictions, export control and special data retention. Even with a legal contract, the perpetrator does not automatically access closed police and medical systems.

40.1. What is required in the contract prior to the commencement of the investigation

The assignment records the legal status of the parties, the subject matter of the independent study, the list of available sources, the demarcation of technical and advocacy services, the terms of access

to confidential information, the intellectual rights regime for concluding and the storage of digital originals. The rules for publication, including the right of a person or organisation to object to disclosure of personal data, are set out separately. The duration of the work depends on the availability of evidence; the contractor should not promise to confirm a certain hypothesis regardless of the test result.

The conflict between the interests of the payer and the protection of the victims is resolved by procedures of independence and disclosure of restrictions. If the customer prohibits the disclosure of substantial rebutting data, the expert is obliged to determine whether and in what capacity he can continue his work, instead of knowingly issuing a unilateral conclusion.

Practical case analysis 19.40

Recommendation: Establish terms of access and conclusion of the contract

Hypothetical example: The perpetrator is obliged to give a reason, although the employer does not provide historical logs.

Analysis: The treaty promise exceeds the available research opportunities.

Practical steps: The parties shall determine the availability, stages, criteria for the outcome and the permissible inference of data inadequacy.

Outcome criterion: The adoption does not require a fictional certainty in the absence of material.

41. Quality control of the investigation and safeguards against AI-generated errors

The internal use of generative AI to build a timeline, find inconsistencies, and prepare text tables is only permissible as an auxiliary tool. Each quote and normative reference should be checked against the primary document. The system can safely come up with a court decision number, a law date, or a statistical indicator; without manual verification, it creates a new evidence defect.

Quality control of an investigation involves repeated restoration of the key chain by another researcher. It checks for the presence of a primary object, the correctness of the time connection, the reproducibility of the calculation and the conformity of the causal inference to the procedure performed. The legal wording, confidentiality and permissibility of external disclosure are separately checked. Observations are classified according to their impact on the conclusion; the resolution of a substantial objection requires a basis in the material.

Each conclusion must have one of three markings: a reference to the verified document, an expert's conclusion with the described methodology, and the Observatoire working proposal. The author's recommendation cannot be processed as if it were directly demanded by GDPR or AI Act. When updating the law, the date of the legal picture and the log of changes to the text shall be kept.

41.1. Monitoring of the use of generic tools in the preparation of an opinion

Using a generative model for initial document search or editorial assistance does not turn its answer into an independent source. Each cited court order, norm, document number, date and statistical confirmation must be reviewed from the original source. A particular risk is the fictitious

numbering of cases, the confusion of law editions, false quotes from standards with paid access, and the automatic rewriting of assumptions into categorical statements. That's why they create a claim registry: confirmation, source, verification date, verifier and restrictions. The unverified allegations are removed or clearly marked as research hypotheses.

The verification of the contents of the permits begins with the question of what observation could change the main conclusion. The reviewer chooses a significant alternative and assesses whether it has been investigated. If an alternative cannot be distinguished because of the lost data, it should change the wording of the result. The existence of a long timeline and completed tables does not replace such a check. The final status of the investigation depends on the extent to which the questions are answered, not the size of the document.

Practical case analysis 19.41

Recommendation: Independently check AI-aid in investigations

Hypothetical example: The assistant put a reference to a non-existent magazine in the report.

Analysis: Automatic generalization has created new unreliable evidence.

Practical steps: The reviewer checks all significant links to the register, counts the calculations separately.

Outcome criterion: Each substantive allegation is supported by accessible material.

42. Development priorities for 2027-2028

The forecast of legal development should be based on a known calendar, but not give out expected projects for the obligations assumed. Amended in July 2026, AI Act's architecture provides for subsequent stages of high-risk regulation, which makes comparative studies of voluntary reporting procedures and organizational readiness particularly valuable. The objective direction is to monitor the emerging official guidelines, supervisory decisions and evaluation methods. [2], [2]

For the international programme, it is appropriate to analyse the possibilities of mutual compatibility between OECD incident classification schemes, EU industry legislation and national administrative practices. Unification should be about the fields of evidence, not the artificial erasure of differences between legal triggers.

A separate task is to create mechanisms for the secure exchange of anonymized AI error data in police and financial analytics. Such a study should begin with an assessment of confidentiality, legality and the risk of abuse: data on detentions, sanctions matches and bank transactions cannot be published mechanically in a publicly available database.

42.1. International standard development scenarios

One can expect further convergence of the terminology for describing an AI incident and independent audit practices, but the forecast should not be assumed to be valid. It is important to maintain the flexibility of internal forms for changes in national legislation, updating NIST AI RMF and new OECD methodologies. Subsequent changes to the industry rules should update the legal

map with the date, edition and specific changed duties, without erasing the history of previously sent messages.

Prospective issues include verifiability of adaptive models, access rights to trace for independent professionals, protection of the affected person when using cross-border AI services and responsibility management in the long supply chain. ARGA's methodological initiatives should be formulated as questions for public consultation, allowing for comparison of different jurisdictional decisions.

Practical case analysis 19.42

Recommendation: Develop a programme to address identified gaps

Hypothetical example: The team wants new directions, although current cases do not retain model entrances.

Analysis: The expansion won't correct the basic defect in the evidence.

Practical steps: The manager first introduces context preservation and a pilot re-analysis; then he evaluates the new tasks.

Outcome criterion: The stage is complete by a test result, not a calendar.



43. Conclusions and institutional proposals

Investigating an AI incident requires three independent outcomes: a verifiable picture of events, a justification of the causal mechanism, and a legal assessment of consequences. Digital forensics keeps objects, causal analysis checks the correlation of changes to outcomes, system security investigates denial of restrictions and feedback. None of these methods per se establishes the guilt of a particular person. Their combined use allows us to determine exactly what is confirmed and what action follows.

For ARGA, a protocol is proposed with a register of research issues, a map of sources and losses, an alternative matrix, a description of repeating checks and completion criteria. The key explanation must be re-analysed before being used externally. If the causative part remains incomplete, the verified facts and the necessary safeguards may be set out separately without fictional certainty.

The closure of the incident must be confirmed by the technical, procedural and restorative part. The model correction does not complete the work if the erroneous recording continues to be transmitted or the rights restriction is maintained. For each affected outcome, a validated outcome is required: a corrected record, termination of the restriction, a message to the specified recipient or a reasoned description of the remaining obstacle. These criteria relate the scientific study of the mechanism to the practical protection of the individual.

ARGA's conclusions are of a research nature, intended for expert advice and do not constitute an official position of a public or international body. This report does not replace the assessment of a local lawyer, judicial expert or competent supervisory authority in relation to a particular dispute.

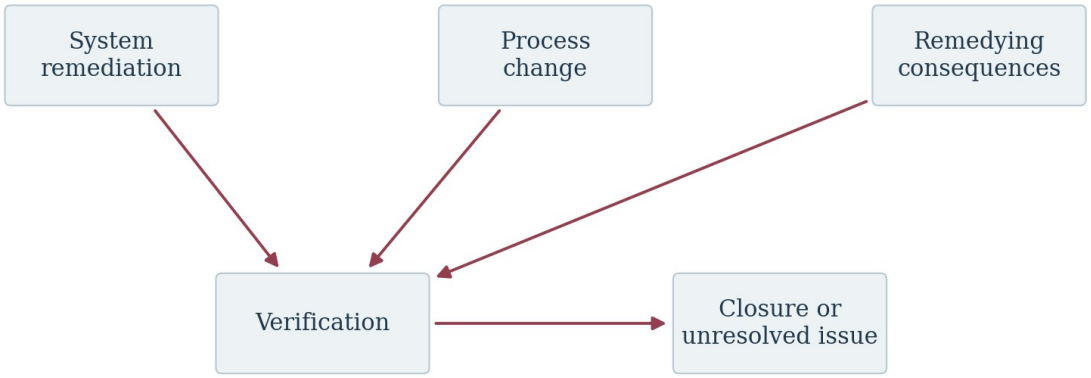


Figure 19.4. Three separate conditions for incident closure. Section 43. A technical correction does not complete remediation while erroneous records and restrictions remain in place.

Practical case analysis 19.43

Recommendation: Close the technical process and remediation parts

Hypothetical example: The retest's successful, but the mistake still affects the client.

Analysis: The work is not complete on the consequences and the copies distributed.

Practical steps: The curator compares the test, process change and correction confirmation from the recipients; the unresolved remains open.

Outcome criterion: The closure is justified in each part without a general unjustified status.

Practical annexes: forms, protocols and monitoring procedures

The materials presented below are Observatoire ARGA's reference analytical templates. All fields, internal stages and priority classes are designated unless specifically specified by law. The forms should be used after verification of local law, confidentiality requirements and the actual competence of the participants. External transmission of information is subject to legal authority.

Appendix 1. Initial AI incident report record

The record is entered before the causality is discussed. Filling in is done directly from the applicant's words, marked "composite" and separately from the results of the "verified" reference files. The classification is changed by a new record, not by replacing the original document. The purpose of the procedure is to prevent the loss of the first message, precise time and original artifacts.

Field	What do you want?	Verification
IR-ID	Unique address number	Not repeated
Received at	UTC, channel, source	Reference heading
Reporter	Role and connectivity	Consent to contact
Alleged event	Literally summary statement	Separated from conclusion
AI system	Name, vendor, version if possible	Do not give a guess for fact
Actual harm	Documented impact	Annexed sources
Imminent threat	Risk of repetition and to whom	Primary protection measures
Evidence list	Name, format, carrier	Hash on receipt
Initial classification	I0-I4 and legal evaluation separately	Author, time

Instruction to the perpetrator: confirm receipt of the message without declaring guilt, open a security log, determine the need for protective measures and pass on the qualification of legal deadlines to the appointed lawyer. A separate line should record everything the applicant asks for correction, including any potential emergency measures to prevent detention, blocking or denial of the necessary service.

1.1. Filling instruction

In the field of "the complainant's communication", the original wording is inserted without editorial normalization. In the "facts" field, only independently validated elements with an indication of the relevant Evidence ID are reflected. The fields "probability", "seriousness" and "intended role of AI" are not mixed. If the name of the model is not known, the "not installed" shall be recorded, not the most likely trade mark. The record receives the version number, initiator and date; new data is added as a clarification rather than subtly replacing the original evidence.

Appendix 2. Priority and immediate protection scheme

The internal P0-P3 scale is designed to manage investigative resources; it is not a legal definition of a serious incident. Priority P0 means a potentially ongoing emergency risk to life, liberty, health or critical service; P1 - significant harm or high risk of recurrence; P2 - localized confirmed incident; P3 - a signal without established harm. The class shall be reviewed as soon as information becomes available.

Priority	Step One	Who Confirms the Action
P0	Immediate safe isolation of function and escalation	Incident light + technical responsible
P1	Retention of evidence and temporary limitation	Incident lead
P2	Planning investigation with damage verification	Designated expert
P3	Collection of missing facts and monitoring	Intake Officer

Each decision contains four compulsory motives: the detected signal, the available confirmations, the risk of incorrect system restraint, and the criteria for returning to normal operation. The P0

should not be automatically assigned to each reference to the word "AI" or to every report in the media. Priority can be high even with incomplete data, if possible delay is unacceptable.

2.1. Interaction with the crisis team

In P0 the preliminary assessment should be immediately sent to the person authorized to alter the condition of the hazardous system. For P1, the person responsible for maintaining the digital tracks and for alerting the victims is assigned to the extent permitted. A separate observer records all isolation operations. If the safety of persons requires immediate action, the lack of a complete technical picture is not a reason to wait; the reasons for the priority decision are indicated in the background number as an explanation, not as a fabricated prior authorization.

Appendix 3. Model order to preserve electronic footprints

The purpose of the order is to suspend the usual removal of a limited number of digital objects without necessarily stopping the legitimate work of the organization. Before sending, the lawyer shall establish the authority to dispose of the relevant information resources, data protection and applicable storage periods. The request should be targeted, listing systems, entities and the time range.

Recommended storage subject: production logs and their indexes; current and previous model deployment artifacts; runtime configurations; system cues; external tool calls; tickets; correction messages; source data; interface access confirmations; policy updates. If it is not possible to keep the file in its original form, describe the export process and known metadata losses.

Elements of an order	Value
Legal hold ID	Code, date, authorized person
Reason	Brief signal without legal prioritisation
Facilities	Systems and categories of logs
Interval	Start/end with UTC and buffer
Responsible	Data uper and experience common
Permitted exceptions	System security/law
Monitoring	Confirmation of performance and deviation log

The Technical Owner confirms the performance in writing or lists the non-performing items. Once the grounds for preservation have expired, the order is canceled by documented procedure; the investigation cannot be turned into an indefinite storage of sensitive information.

3.1. Legal and technical targeting of the request

Preservation orders are sent to responsible administrators, cloud service owners and contractually available external suppliers. The request contains specific data areas, time interval and access restrictions. It is not possible to require indefinite copying of all the personal data of an organization: scope and retention are determined by the needs of the investigation and legal grounds. If the request is refused or it is technically impossible to execute, the request itself, the answer and alternative available actions are retained.

Appendix 4. Digital artifact register and integrity check

Each object has an independent Evidence ID. The information about receipt, verification, and copying should not be confused: the screenshot author, the analyst, and the file keeper may be different people. The use of SHA-256 is a practical recommendation for integrity control rather than a universal legal requirement; consideration should be given to the quality of the tool, access rights and the possibility of hardware certification.

Form field	Value / sample
Record 1	
Evidence ID	E-001
Type	Raw log
Source	Example: Service A
UTC Time	Fill-in
SHA-256	Fill-in
Limitation	Limited access
Record 2	
Evidence ID	E-002
Type	Model manifest
Source	Example: withdrawal
UTC Time	Fill-in
SHA-256	Fill-in
Limitation	Commercial secrecy
Record 3	
Evidence ID	E-003
Type	Operator's Decision
Source	Example: internal workingflow
UTC Time	Fill-in
SHA-256	Fill-in
Limitation	Personal data

The register should be maintained in a system with a controlled history. When an artifact is forwarded, it not only records the recipient's name, hash, size, channel and authorization. If an export tool changes a file, it creates a new entry with the ratio "produced copy E-001" rather than replacing the value of an old hash.

4.1. Exactly what the control amount confirms

A hash comparison is used to prove the identity of a file's copies at a given point, but it does not replace the question of completeness, origin and truth of the contents. The availability of SHA-256 in the inventory does not guarantee that there is no change before receiving the facility. A new derivative object is created with its own identifier, and a record of its connection to the original is kept in the log. The missing control amounts are marked as a limit, not filled with assumed values.

Appendix 5. Chain of society magazine

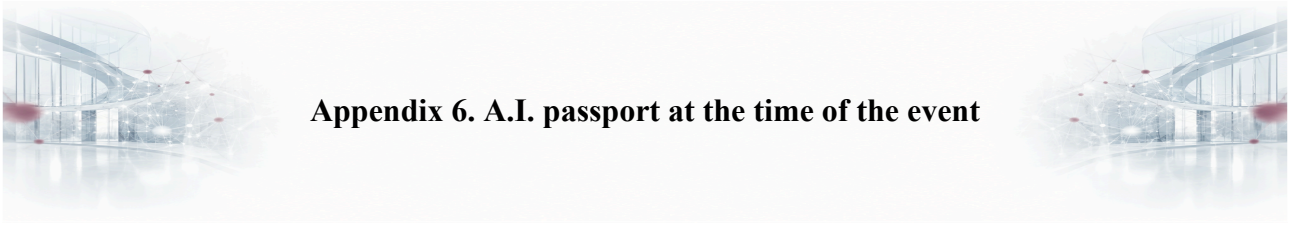
The storage chain describes the physical and logical movement of evidence. It is designed to allow the verifier to determine that the object being analyzed is consistent with what was received and to identify gaps in unauthorized access. An electronic signature, an access log, unchanged storage and a hash are useful, but they don't guarantee the completeness of the source in themselves: you can also sign a partial export.

Form field	Value / sample
Record 1	
Transfer ID	T-001
Evidence ID	E-001
From Who	Keeper
To whom	Expert
UTC Time	Fill-in
Channel	Secured
Objective	Analysis
Signature/post	Fill-in
Record 2	
Transfer ID	T-002
Evidence ID	E-001
From Who	Expert
To whom	Archives
UTC Time	Fill-in
Channel	Secured
Objective	Return
Signature/post	Fill-in

It is prohibited to leave without explanation a "common folder" to which any staff member had access by the administrator. In such cases, the risk of changing the facility is directly recorded in the expert report. If necessary, take a check taken of the history of access to the folder itself.

5.1. Control of transfers between experts

Each transmission must have the receiving party, the destination, the exact time, and the security of the package or file. If the user receives a copy by e-mail and then sends it to an expert, the log shows both known stages and indicates those that are not available. An unknown path of ownership reduces the burden of proof, but it does not always make the subject completely unfit: the final assessment depends on the applicable procedural law and other evidence.



Appendix 6. A.I. passport at the time of the event

For an agency service, playback requires information that is not visible on the screen to the user. The template covers the exact model name and version, API-provider, deployment area, container ID, decoding configuration, prompts, security policy, retrieval-corpus version, list of tools and permissions. If a part of the fields is closed by the supplier, a written request and an inaccessible response are recorded.

Layer	Parameter	Confirming object
Model	Provider, model ID, revision	Manifest or vendor response
Launch	Development ID, date, region	CI/CD and cloud log
A clue	System/developer version	Orchestrator Photo
Context	RAG corpus and relevant snippets	Snapshot index
Tools	Tool list, credentials scope	IAM export
Protection	Policy filters and guardrail version	Safety configuration
Human	Approval workflow	UI activity record

When the code is partially closed, black-box analysis is allowed, but it's not the same as an analysis of the original production environment. All differences in the test and initial regimes are listed in the conclusion.

6.1. Minimum reproducible passport

The environment passport shares the technical version of the model, the configuration of the application and the state of external dependencies. For example, the same weights may give different responses when changing the retrieval body, the rights of an external tool or a system instruction. If the vendor does not reveal the version identifier, the documented time and route of the API call should be kept. A passport is needed to determine the limits of reproducibility, not to create a false guarantee of determination.



Appendix 7. Protocol on Time Normalization

The source logs may reflect local time, Unix epoch time, timezone offset or an indefinite record without a time zone. The transformation protocol should retain the original value and the rule of normalization. When the notice period is assessed, the legal entity is notified of the event and not just the technical event. If the dates are different, create a scale with a range of uncertainty.

Form field	Value / sample
Record 1	
Event ID	EVT-01
Raw timestamp	Fill-in
Zone/offset	Fill-in

Form field	Value / sample
Normalized UTC	Fill-in
Confidence	High/Medium/Low
Origin	Log ID
Record 2	
Event ID	EVT-02
Raw timestamp	Fill-in
Zone/offset	Not established
Normalized UTC	Range
Confidence	Low
Origin	The witness's words

It is not allowed to turn the technical timetamp of subsequent exports into the date of the event. The uncertainty interval is reflected both in the graph and in the legal note. If necessary, independent system clocks, postal headings or transaction logs are used as synchronization points.

7.1. Limitations on the normalization of time

Clock correction should not erase the original timestamp provided by a server, human or cloud service. For each record, the time zone, the possible drift, the method of counting and the reliance on the source are indicated. One-minute mismatch is sometimes significant for causation, but the greatest danger is daylight saving errors and a shift to summer time. If it's impossible to pinpoint the exact moment, they record the interval instead of the artificial precision to milliseconds.

Appendix 8. Mandatory and voluntary communications matrix

For each regime, a separate line is created with the legal norm, the system category, the legal addressee, the trigger, the time of notification, the deadline, the notification channel and the status. This avoids a false "general deadline of 72 hours". The table below serves as a reminder of the basic different regimes rather than automatic legal characterisations. [1], [8], [10], [12]

Mode	Standard orientation	Trigger and reservation
AI Act art. 55	Without undue delay	GPAI with systemic risk
AI Act art. 73(2)	Immediately after establishment/reasonable probability of communication; maximum 15 days	Only after verification of coverage, transitional regime and special cases
AI Act art. 73(3)	Immediately; maximum 2 days from awareness	Special type of serious event
AI Act art. 73(4)	Immediately following the establishment or suspicion of communication; maximum 10 days from awareness	Death and conditions of the norm
GDPR article 33	Without undue delay; up to 72 hours, if possible, from awareness	Personal data gap and risk

Mode	Standard orientation	Trigger and reservation
NIS2 art. 23	24 hours; 72 hours from becoming aware; final report within one month of notification	Significant incident; scope of the subject and national implementation. For confidential services, notification - within 24 hours; for ongoing incidents, special procedure applies
DORA RTS 2025/301	4 hours from classification and maximum 24 hours from awareness; late classification - 4 hours from it. Interim: 72 hours from first report; total: month from last interim report	Major ICT investment; multiple-stage reporting under art. 5 RTS 2025/301

Practical procedure: first check the applicability of each rule, then calculate the deadline in the local calendar of the respective addressee, designate the responsible person and create a reminder. The final legal note must be signed by a specialist who has verified the actual status of the organization.

8.1. Filling out of the legal time register

For each individual obligation, four fields are required: normative basis, applicable subject, legal trigger and start tag. The timing of a technician's identification, client reporting and obtaining a legal characterisation may vary; they are recorded in the same log, but not equated. The lawyer must explain why a specific start date has been chosen, including information about the organization's possible earlier knowledge. The dates AI Act, GDPR, NIS2 and DORA cannot be filled in with one template number.

Appendix 9. Form of early communication to the competent authority

Introduction: "The circumstances known at the time of dispatch are given below. The definition of cause and scale may be clarified after the preservation and independent assessment of the artifacts." This formulation is a neutral example: the regulatory formulae in force and the fields provided for by law have an advantage.

Mandatory blocks: legal person and contact of the responsible person; number of contact; applicable legal basis; date of notification; identifier and designation of AI-system; confirmed harm and potential impact; current causal assessment; information about security measures; plan of additional communication; protected document exchange channel.

Field	Content instruction
Legal addressee	Name of competent authority confirmed by a lawyer
Basis	Specific article, revision, date of application
Information	Facts separate from assumptions
Consequences	Confirmed / estimated / unknown
Conservation	What information is protected from disposal

Field	Content instruction
Contact	Secure e-mail or official portal

It is recommended to attach an inventory list, not primary confidential arrays without the requirement of law. Subsequent clarifications should be sent with reference to the previous notification and the dated table of the changed information.

9.1. Material sufficiency of early notification

If the specialized regime allows for an initial incomplete report, the document should clearly distinguish between verifiable facts and verifiable hypotheses. Include the date of detection, the scale available, the nature of the risk, timing and contact for further interaction. Later, the supplement refers to the original document and explains the changes. You can't hide an important uncertainty from the fear of looking unprepared: true incompleteness within the applicable norm prefers known false accuracy.

Appendix 10. Request to the supplier of the basic model and the integrator

The purpose of the request is to obtain the necessary and proportionate information to explain a particular event, rather than "the whole source code of the developer". The letter should specify the identifiers of the operations, the period under investigation, the nature of the consequences and the specific legal or contractual regime of access. The logs of the model supplier and the interface integrator are separately requested.

List of requests: full model ID and changelog; history of migration between versions; timestamps inference; API parameters and restrictions; known relevant incidents; history of guardrails refusal; conditions of journaling; order of evidence storage; description of independent tests and their restrictions; contact technical and legal responsible. The request does not imply an unconditional right to unavailable commercial data; the terms of the transfer are agreed in a lawful manner.

Response	Action by the investigator
Full data	Check origin and control of integrity
Total statistics	Clarify denominator and sample
Secret material	Offer a secure review
No data	Record the absence and cause
Refusal by law	Request legal grounds for refusal

A compulsory rule of thumb: Don't give the vendor no answer for admitting guilt. However, the impossibility of verifying a causal mechanism must be clearly stated as a limitation of the inference.

10.1. Security of request to supplier

The request shall be made through the channel provided for in the treaty or the formal process. It is accompanied by a specific request ID, a time window, a list of technical objects and an explanation of necessity. It is forbidden to assume that a supplier is obliged to disclose a trade secret to a private expert only because an investigation has begun. In case of refusal, it is useful to offer an independent expert in a confidential environment or another legitimate way to confirm the facts.

Appendix 11. Report of the interview of the operator and the decision maker

The survey is divided into free presentation, fact-finding and presentation of previously registered artifacts. Interviewees are prohibited from suggesting an expected explanation or starting a question by claiming that the algorithm was wrong. First, the job role, the information available, the power to stop the action, the real-time processing of the record and the familiarity with the instructions.

Recommended questions: what did you see on the screen; what results were available to you before you made a decision; was there any other source; could you cancel the machine offer; what warnings were given; how deviations were recorded; who approved the application policy; has anyone notified you of similar errors before?

Protocol Element	Requirement
Consent and grounds	Checked by Local Law
Conditions of interview	Time, place, participants
Free Story	No lead language
Show artifacts	ID and presentation
Discrepancies	Fixed without erasing the original response

Discrepancies should not be removed from the final protocol for the sake of "purity" of narrative. The discrepancy between the interview and the log may reflect a memory error, another interface, or a significant control gap.

11.1. Do not mix witness memory with system log

In the interview, the operator is given a neutral timeline without an early message "model declared defective". Questions about what he saw on screen and what actions he took separate him from his assumptions about the causes of the failure. If the interview record is used as evidence, the exact agreed formulation of the answers, date, way of recording and information about possible verification are kept. The policy of storing interviews should take into account labour and personal legislation.

Appendix 12. A matrix of causality

For each major event, at least three competing hypotheses are formulated. For example: the original data is wrong; the model processed the correct data incorrectly; the conclusion was correct, but the employee applied it outside the context. For each hypothesis, there are supporting and rebuttal observations. It is not permissible to limit oneself to hypotheses that are known to be less plausible than the original version of the customer.

Hypotheses	What must be seen	Which it will deny	Status
H1: Wrong entry	Divergence with primary source	Entry correct at decision date	Open

Hypotheses	What must be seen	Which it will deny	Status
H2: error model	Playing Error	Re-evaluation independent tests don't confirm	Open
H3: wrong workflow	Act contrary to regulations	The rules and actions coincided	Open
H4: External attack	Trace injections/access	No relevant developments	Open

The conclusion is awarded after evaluating the quality of each source and the impossibility of accessing additional data. If H1 and H3 are both significant, you cannot artificially choose "one guilty cause". Methodology allows multi-factor attribution.

12.1. Verification of the causality hypothesis

The matrix is filled with negative alternatives: incorrect source document, erroneous routing, failure of manual control, an external attacker, divergence of versions or a legitimate reason for the final measure that doesn't depend on AI. For each hypothesis, an understandable criterion for confirmation and rebuttal is formulated in advance. If the available information does not allow for the deletion of several explanations, the final status remains "multifactorial uncertainty" rather than "the most convenient version".



Appendix 13. Test environment control playback and restriction plan

Playing should have a certain research question. Before the experiment, the model version, the samples, the data set, the instrument scheme, the seed if available, the generation parameters and the expected comparison criteria are recorded. No dangerous cyber operations, personal data manipulation or harm simulation should be conducted in an open production environment. First, they assess the safety and feasibility of using synthetic/impersonated test objects.

Element	Verification requirement	Document
Subject	One verified hypothesis	Test plan ID
Wednesday	Isolated and authorized	Environment manifest
Data	Permitted, required	Dataset manifest
Repeats	Number of launches determined	Sampling plan
Success	Observed criterion	Outcome rubric
Interruption	Conditions of step	Safety plan

The results include a number of successful and failed attempts, scattered outputs, test environment failures, and withdrawal from the plan. If the test model is already different from production, the report indicates the level of transfer of conclusions: confirms the mechanism, indirectly illustrates the possibility or is not suitable for evaluating a historical event.

13.1. Design of permissible re-test

Before starting the control reproduction, the permissible data, the extent of the agent's powers, the conditions for stopping and the evaluation criteria are confirmed. Use of a real bank account, access to critical infrastructure or sensitive information is only possible with proper authorization and protection. For each run, a complete protocol, a version of the environment and the result are recorded. A test performed exclusively on a modified model should not be considered an accurate reconstruction of the original incident.

Appendix 14. Statistical form for testing rare harm

When investigating rare events, it is particularly important not to give "0 out of 100" as proof of zero frequency. The report contains the exact sample size, the selection mechanism and the confidence boundary. For illustration of zero events among 100 independent observations, the one-way approximate upper boundary of 95% under the rule of three is 3/100 = 3%. The method is only applicable when independence and representativeness are allowed; for clustered events it requires correction.

Field	Example of symbol	Verification
N	100	Number of independent exhibitions
Errors	0	Definition of an erroneous event
Percentage	0%	Precise observation, false risk.
Upper boundary assessment	≈3%	Specify the method chosen
Displacements	Selection, lost events	Describe without concealment

If complaints are collected only among those who know about the right to dispute, the sample is likely to be removed. The actual frequency may differ on both sides, and the confidence interval does not correct a systematic selection error.

14.1. Uncertainty Report for rare events

When there are no reported incidents, the evaluation should include the size and nature of the sample, the dependence of the observations, the reliability of the detector and possible missed events. The rough 3/n rule in an appropriate binomial setting represents only the upper boundary guidelines and is not automatically applied to serial correlated observations. When comparing systems, absolute frequencies, denominators, and confidence intervals are reported, not just the percentage of change. The risk taken is not taken from one statistics without an assessment of the severity of the consequences.

Appendix 15. Checking discriminatory effects

The working matrix for high-impact decisions should take into account the legal basis for using sensitive features in the audit. The aim is not to automatically label discrimination, but to assess whether there is a statistically and legally relevant difference due to unacceptable factors. The quality of the data and the comparability of the groups should be monitored in assessing the results.

Comparison Field	Group A	Group B	Commentary
Number of applications	N_A	N_B	Different sample size
Refuses	R_A	R_B	Legality of the decision separately
Percentage of refusals	R_A/N_A	R_B/N_B	Doesn't prove causality

Comparison Field	Group A	Group B	Commentary
Corrected Errors	C_A	C_B	Check the time frame for correction
Missing data	M_A	M_B	Risk of selectivity

Before you use the indicators in the conclusion, you should check the definitions, eliminate the double count and make sure that the algorithmic procedure really influences the final decision.

15.1. Assessment of comparability of groups

Demographic comparison requires a legal basis for processing the signs and analysis of the representativeness of groups. It is necessary to determine whether the rules for marking are the same, the likelihood of complaints being filed and the quality of the original records. The outcome can be determined not only by the model, but also by individual decision-making policies. The absence of reliable group data should not be compensated by fictitious values: the expert should limit the conclusion and specify the requirements for admissible additional analysis.

Appendix 16. Expert independence checklist

Independence is not guaranteed by one signature under the word "impartial". The expert discloses financial involvement, previous services to the party, access to proprietary data, customer publication restrictions, participation in system development, personal relationship with management and promised future contracts. The conflict may be manageable or inconsistent with the assignment; the decision must be motivated.

Verification	Yes/No	How to compensate
You used to be involved in the development?	Fill-in	Independent reviewer
Is the payment dependent on the result?	Fill-in	Revision of conditions
Is the publication of the negative conclusion monitored?	Fill-in	Author's right to text
Any access to the originals?	Fill-in	Unlock Restrictions
Are you using your own pay-as-you-go tools?	Fill-in	Reproduction by others

The expert's organizational location should be mentioned in the report. A private researcher involved in a contract should not be called a "appointed public investigator" if the relevant act of appointment is not available.

16.1. Documentation of independence

The conflict resolution form is filled out not only by the project manager, but also by leading technical and legal authors. The contractual relationship with the verified supplier, previous involvement in system design and material interest in a given outcome are indicated. If independence is limited, the report must indicate the limits of conclusion and verification by another group. The name Observatoire ARGA does not imply a government mandate or official supervision function.

Appendix 17. Instruction to a judicial or technical expert

The technical task asks questions that can be answered by a research method: what versions of the model were used, if the input can be reconstructed, how AI-to-human-to-action chain is arranged, what alternative causes are confirmed, how reproducible the results are, where the evidence gaps are. The question "is the organization guilty" should not be asked if it requires a judicial decision outside the expertise of the expert.

Question	Information for analysis	Boundary of conclusion
How did the result come about?	Prompt, data, model, tool traces	Without independent legal characterisations
Why is it decided?	Human Logic and Regulations	With incomplete access - restriction
Is it possible to repeat a mistake?	Isolated test	Does not mean an identical historical environment
Was there an impact on rights?	Impact papers	Legal conclusion separately

The contract must guarantee the expert the right to report data that contradict the original hypothesis of the customer. The requirement to remove uncomfortable files from the review is itself to be documented.

17.1. Content of the expert's task

The task is to formulate the questions to be checked, not the desired answer. For example: "Did the source of the disputed allegation exist?" or "Is the call of the tool documented with these access rights?" Separately lists the unavailable systems, permissible experimental methods and prohibitions on illegal object modification. The expert reports the impossibility of answering in the absence of data, instead of replacing the absence of a trace with probable reasoning.

Appendix 18. Investigation calendar and establishment of control points

The internal schedule is built around real legal timelines. At the first point, they record the risk, the artifacts and the responsible; then, as the data comes in, they organize a legal characterisation, an independent test and the update of the interested parties. The deadlines for individual internal tasks are determined on the basis of risk and resource; they do not cancel out legal deadlines and may be shortened for emergencies.

Reference point	Result	If not implemented
Intake complete	Primary record and contact	Set Alternative Source
Evidence hold	Written confirmations of retention	Reflect risk of loss
Legal matrix	Message regimes checked	Escalation to the Legal Coordinator
Root cause review	Checked competing hypotheses	Supplement the expert review
CAPA validation	Rectification test	Do not close the incident

The calendar should contain the task owner and a time-delay log. For urgent notifications, it is unacceptable to hide the delay within the general status "under investigation".

18.1. Control points checked

The calendar contains individual dates for the preservation of digital evidence, initial assessment of the legal regime, decisions on protective measures, sending notifications, obtaining materials from the supplier, technical reproduction, external verification and restoration of rights. The postponement of one step should not automatically delay the legal period of the other. The schedule changes maintain the original version and the reasoned decision to change.

Appendix 19. Notification of the victim and explanation of the restriction

The text should be understood by a person without technical education. They report on the event, what has been confirmed and is still being investigated, how to contact them, how to correct errors and contact them for help. In cases where the law restricts disclosure, you must disclose as much factual information as possible without violating confidentiality.

Recommended logical blocks: "what action was affected", "what decisions have been taken about you", "what has already been corrected", "what information remains", "how to challenge the decision", "what time limits for response are set by applicable law". Compensation should not be promised without verification of grounds and authority. Nor could the notification be replaced by the phrase "there was a technical error" if the individual's particular decision was still valid.

Check before dispatch	Result
Doesn't the text attribute guilt to a person?	Fill-in
Okay, where do I go?	Fill-in
Are there any real corrective measures?	Fill-in
Other persons' data not disclosed?	Fill-in
Are the local rules of notification respected?	Fill-in

19.1. How to report unknown facts to the victim

The treatment of the person concerned should be understandable and practical. The verified fact, the probable risk, the defence measures, the available means of dispute and the coordinates of legal contact should be described separately. It is not permissible to promise a final elimination of the consequences if part of the downstream organizations has not yet confirmed the correction. The notification shall not disclose the personal data of other applicants and information whose publication could increase security risks.

Appendix 20. CAPA Register of Corrective Actions

A technical correction is deemed to be executed only after verification of the expected result. The register reflects the relationship of the measures to the identified root cause, responsible person, test criteria, status and residual risks. The strongest proof is the pre-formulated criteria for adopting change, not the backward metric of success.

Form field	Value / sample
Record 1	
CAPA ID	C-01
Reason	Mistaken Transliteration
Action	The Versailed Normalitation Rule

Form field	Value / sample
Owner	Data team
Test	Regression set
Risk after	Fill-in
Record 2	
CAPA ID	C-02
Reason	No hand override
Action	New Workflow
Owner	Operator
Test	UI exercise
Risk after	Fill-in
Record 3	
CAPA ID	C-03
Reason	Incomplete logs
Action	Immutable trail
Owner	Security team
Test	Audit replay
Risk after	Fill-in

For a significant incident, an independent verification is assigned. If both the correction and its acceptance are carried out by a single employee, the risk of conflict of functions is indicated in the conclusion.

20.1. Evidence of CapA performance

In addition to the "completed", specific evidence is stored: configuration diff, independent test protocol, re-training document or confirmation of correction to the addressee. The designated performer and the inspector shall be visible. If the measure only partially reduced the risk, the residual risk remains a separate record with the owner and the date of the next control. Correction of documentation without correction of ongoing damage is not considered to be sufficient closure.

Appendix 21. Protocol of the decision on the renewal of the

The decision to restart is made within the scope of legal authority after a safety and residual risk assessment. The report includes a list of corrected defects, repeat test results, detected regressions, independent review, post-launch monitoring conditions and reverse shutdown procedures. If necessary, limit the class of users and the scope of actions allowed.

Question	Confirmation
Are the originals saved?	Evidence ID and repository
Is the primary cause eliminated?	RCA and CAPA test
Residual risks disclosed?	Risk register
Is the victim's defense system in place?	Remediation register
Did the supervisor sign the responsibility?	Decision memo

The protocol specifically specifies "the criteria for immediate return to safe operation". Restarting without the ability to observe a repetition of the original dangerous pattern is not a sufficiently confirmed solution.

21.1. Conditions for resumption of operation

The decision to launch should not be made only by a team interested in complying with the commercial schedule. An important system requires technical confirmation of correction, regression testing, verification of legal notices, analysis of remaining risk and appointment of a person capable of stopping the service again. If the original cause cannot be replicated, a more cautious mode of functional restriction and enhanced monitoring is in place.

Appendix 22. Table of levels of confidentiality and legal transfer

The separation of access begins with the determination of the purpose and the legal basis. Not all participants should see the full entries, bank numbers or health information. When handed over to an outside expert, a specific role is assigned and the level of access recorded. Public disclosure of results must be independent of personal data, trade secrets and details.

Category	Possible access	Model measure
Public Chronicle	Total audience	Anonymization
Working machine logs	Technical Group	Pseudonymization
Personal records	Legal officer and authorized expert	Limited workspace
Security vulnerabilities	Narrow security team	Controlled disclosure
Legal correspondence	Only authorized persons	Special protection regime

In each case, applicable laws and treaties are checked. The status of "internal audit" is not in itself a universal exception to data protection.

22.1. Distributed access practices

Each category of material corresponds to the list of authorized recipients. The original medical information, personal data, organization secrets and law enforcement materials should not automatically enter the general archive. For the publication version, data minimization and content verification are used so that deletion does not distort the meaning of the conclusion. The list of removed categories and the reason for concealment are kept in the internal register; the breach of evidence availability is marked as a restriction of independent verification.

Appendix 23. Structure of the expert opinion for external verification

The conclusion must be drawn in such a way that the regulator or judicial expert can independently verify the facts, methods and assessments. Recommended order: title sheet; mandate; independence; issues; sources; restrictions; established facts; technical scheme; results of reproduction; competing hypotheses; causal inference; legal map as a separate annex; protective measures; evidence attachments.

Chapter of the conclusion	Objective	Anti-failure
Introduction	Who and what did	Refuse to approve a State appointment
Sources	Controlled equipment	Do not refer to unnamed files
Method	Experiment and assumptions	Don't hide the test environment
Result	Facts, estimates, unknown	Do not mix confidence levels
Objections	Inconvenient data and alternatives	Do Not Remove Dissent

It is recommended that a table source-to-clay index be attached. A long report without such an index can be literally compelling and evidence-poor.

23.1. Completeness of the expert opinion

The conclusion involves a chain from the question to the sources and method, then to the facts, to alternative hypotheses, to an assessment of access constraints and outcome. Each categorical affirmation must contain a reference to the verified artifact or official legal text. The reader should be aware of which parts are prepared by a technical expert, which are prepared by a lawyer and which institutional recommendations belong to the Observatoire ARGA. The formal attributes of the report do not in themselves create the status of a forensic expert.

Appendix 24. Criteria for closure and follow-up

The closure is made only after the verification conditions have been met: continuing immediate harm has been ruled out; all important sources have been preserved; legal notices have been completed or documented as not applicable; explanation of causes is sufficiently substantiated; corrective measures have been tested; individual consequences considered; residual uncertainty disclosed.

Reference criterion	Artifact required
No immediate ongoing harm	Protection Act/Retest Act
Legal map updated	Note by the lawyer
Artifacts available	Register and chain of society
RCA reviewed	Independent opinion
CAPA accepted	Signed verification report
Affected persons have a channel of protection	Register of rehabilitation measures

The status of "closed" does not mean the destruction of history. The storage period is determined by the legal grounds and the function of the evidence; after it is completed, a documented removal policy is applied.

24.1. Delayed detection of repeat effects

The formal conclusion of the investigation is accompanied by a period of re-monitoring. It depends on whether the company continues to use the system, what copies of the disputed records could be preserved, and when those affected may experience the consequences again. The registry records the channel for late reports, the owner of the post-incident review and the terms of reopening. The destruction of evidence after the expiry of the applicable storage period shall be made separately from the decision to close the incident.

Scenario Annexes: Methodology of application in model cases

All the following situations are designed to test the research methodology. They don't reflect actual events or client work. In each case, the type of problem, the minimum evidence base, permissible and inadmissible conclusions, individual legal issues and conditions for restoring rights are indicated.

Appendix 25. Scenario: Identity error in the sanctions check

Condition (inventory). A bank in the EU uses an external tool to compare customer profiles with sanctions lists and company ownership data. After updating the transliteration rules, the system gives a high rating for the resemblance of the client's name to the name of the person on the list. The employee sees the final "risk allert" and suspends the major payment. The client indicates a different date of birth and citizenship. Three weeks later, the bank admits there's no match, but the alert information remains with the outside supplier and becomes the cause of new inquiries.

The researcher does not conclude that the illegal blocking has "received AI". It should be determined which legal mechanism the bank applied, whether there was an independent basis for verifying the transaction and whether the specialist could assess personality traits. The log of the matching algorithm version, the result of the original query, the completeness of the employee's identifiers, the history of decisions and mails, downstream customer complaints are particularly important in the case. The data error variant and the human confirmation variant should be checked separately.

Verification step	Artifact	What Can Be Fitted
Signal Identification	Original match event	Precisely comparable fields

Verification step	Artifact	What Can Be Fitted
Manual evaluation	Employee decision trace	The Role of Man
Basis of the measure	Compliance memo	Legality of limitation separately
Corrigendum	Correction tickets	Where error has been corrected
Repeat	Provider downstream logs	Whether harm continues

Causal hypothesis H1: low normalization quality increased the likelihood of a mismatch. H2: The system correctly reported similarities, but the interface hid uncertainty. H3: The bank applied its own excessively broad ban, irrespective of machine evaluation. To analyze, you need name control pairs, threshold checks and access to manual review regulations. The absence of the false negative observed cannot automatically be turned into a basis for radically increasing the threshold.

A separate legal map covers the sanctions law, GDPR for personal data processing and possible sectoral banking regulations. The reporting obligation under Article 73 of AI Act is not established without evaluating the category of the system and the legal threshold for a serious incident. The final means of protection includes reviewing the blocking in writing, correcting an erroneous profile and verifying notifications to third parties.

25.1. Working procedure for verification

The original object is a specific decision to delay or block the operation and the technical signal that preceded it. The expert compares the list's identifiers, the timing of its update, the results of the identification check and the version of the match rules. Passports, unverified dates of birth, transliteration and just a name match should not be considered equally reliable. It is documented separately what information was available to the employee who confirmed the alert and whether he had any real possibility of revoking the machine recommendation.

The next step is the distribution map of the wrong recording. It shows which vendor has formed a match, which organizations have received the result, and what actions will be required to correct different copies of the data. Restoration of banking services shall be determined by the applicable law and an independent decision of the competent financial institution. The expert's conclusion is limited by the verified reason for the technical match, the distribution of roles and the fact of execution of corrective actions. This scenario is not a description of the real ARGA client.

Appendix 26. Scenario: algorithmic match at border crossing

Condition (inventory). When checking the border, the matching system matches the foreign prosecution record. The border guard detained the person for further verification, citing the absence of other identifiers. A few hours later, the prosecutor's office releases the man, establishing a difference between birth and nationality. The internal documents don't reveal whether AI was used or a determined search rule.

The first requirement is not to attribute an action to a particular system without proof. The researcher collects the detention decision, the official reason for verification, the date and time, the identifiers reflected in the national decision, as well as any legally available traces of automated comparison. The fact of the release itself does not prove the illegality of the original verification, and the existence of an international communication does not imply its automatic authenticity or correct interpretation.

The second requirement is to reconstruct the chain from external information to a national solution. It is important to know what additional checks were made available to the employee and whether it was possible to quickly clarify the data. The quality of the translations, the transliteration, the differences in status should be assessed: the verification report, the orientation, the national search, the court decision. Each link indicates the competence of the body and the procedural basis.

Verification branch	Source	Acceptable conclusion
Identity information	Official documents	Confirmed differences
Origin of information	Legal response of the competent authority	Status and date of recording
Work of the Classificationator	Available technical documentation	AI-input boundaries
Decision on restriction of liberty	Protocol and judicial records	Legal qualification by a lawyer
Subsequent effects	Travel and complaints documents	Scale of repetition

The script shows the limits of purely technical remediation. Even if the matching mechanism is adjusted, the wrong copies of the data may remain in other systems. Therefore, the completion program must contain a confirmed correction of the source and available legal disputes in the relevant jurisdictions. In the absence of access to a closed national log, the expert must acknowledge the impossibility of establishing an accurate mechanism of AI.

26.1. Working procedure for verification

The first document should be a procedural act on the actions of the border authority, not a guess about the existence of an INTERPOL notification. Identity certificates, national background checks and available official answers are collected. If the authority has used automatic comparison, the set of fields being compared, the logic of the threshold and the employee's participation are further investigated. The unnamed algorithm is explicit; it is not possible to say "machine confused people" on the basis of one similar name.

The materials are distributed between the technical expert and the lawyer of the relevant jurisdiction. The first analyzes traces of data processing; the second determines the admissibility of evidence and appropriate ways to dispute detention, extradition or record keeping. The conclusion of a violation of the right to liberty is a competent legal assessment and does not automatically follow from a report of a technical error. Priority is given to available urgent procedural measures for the protection of human beings.

Appendix 27. Scenario: Automatic classification in the asylum procedure

Condition (inventory). The administrative system sorted the statements by the types of evidence and used AI to translate, extract dates, and detect inconsistencies. In the same case, the algorithm mistakenly combines dates from different calendars and marks the applicant's story as contradictory. The decision-maker refers to an automatically prepared summary. Later authentic documents and translations show that a discrepancy arose during processing.

The investigation should not only determine the translation error, but also the extent to which it affects the motives of the decision. All summary edits, input documents, translation settings, source language, actual citations in the official decision and human oversight procedure are required. It is checked separately whether the person has been able to read the original material and correct the wrong translation until an unfavorable decision has been made.

The storage of sensitive data is particularly difficult: asylum information can relate to persecution, health, political situation and relatives abroad. The technical team should use minimization, limited disclosure and controlled methods of analysis. It is not permissible to transfer such a file to an unknown external LLM service without verifying special requirements. An independent expert can study the process on an unidentified copy, and the originals are stored in a legally accessible protected storage.

Subject	Control action	Type of error
Date of event	Compare original and translation	Normalization of the calendar
Quoted in decision	Compare to Source	Breaking out of context
The role of AI	Describe specific workflow	Sub-summy and decisions
Right to review	Check Accessible Means	Incomplete correction

With a proven contribution of AI, it is required to restore the correct translation and take it into account in the proper review procedure. The conclusion of a private auditor does not invalidate an administrative decision; legal defense must use competent national and international mechanisms.

27.1. Working procedure for verification

In the asylum procedure, automatic classification can sort documents, determine the priority of consideration or form a draft resume. For judicial and administrative analysis, it is important to determine whether AI-output influenced the final reasoning, whether the employee was able to access the original sources and whether the complete translation of the applicant was preserved. An incorrect automatic remark should not be considered as a proven fact against the applicant. The technical audit compares the source, machine translation, summary and final administrative decision with an accurate indication of each transformation.

The actual legal significance is established in the national challenge procedure. The expert has the right to show that the denial has been translated incorrectly, that the date has changed or that the original document has been excluded from the final analysis; however, deciding whether this affects the legality of the denial and which procedural method of correction is available requires an

attorney's assessment. With limited access to the internal system, the fact of using AI is not assumed, but documented.

Appendix 28. Scenario: Mass AML operation after model update

Condition (inventory). The financial institution updates the transaction scoring, after which the number of withheld transactions increases fourfold. The technical team links growth to seasonal translations and closes ticket as a "normaal representation". The subsequent selection audit detects a feature extraction error: one currency code was taken as an indicator of the country's high risk.

Technical testing begins with comparing input distributions over periods before and after the update. But changing the aggregate alert number doesn't prove the rise of false starts without validated ground truth. Requires sample of reviewed cases, marking by independent experts, history of feature engineering changes and statistical intervals. It's also important to remember that the over-filtration could have existed before the update, and the new rule only made it more noticeable.

Verification	Material	Why?
Before/after	Versioned event counts	Identify change point
Feature mapping	Commit history	Find the wrong coding
Sample review	Blind labels	Assess false positionives
Human workflow	Operator logs	Assess secondary monitoring
Client remediation	Complaint and reversal log	Confirm the removal of the injury

The legal analysis covers the AML obligations of a financial organization, the rights of clients, and the limits of disclosure of internal compliance. Even if the event affected IT infrastructure, the DORA classification requires independent verification of the criteria for major ICT-related incident; the rise of false concerns does not in itself determine this status. The final action is to correct the sign, check the regression on multiple currencies and evaluate downstream customer feedback.

28.1. Working procedure for verification

The study begins with the fixed date of the model update, control set and three time slots: before introduction, immediately after and after adjustment. We need to distinguish the increase in real risk from the increase in the number of signals due to the change in the threshold. The report separately accounts for transactions, unique clients, confirmed violations and unfinished manual checks. The comparison is unacceptable if the marking criteria, processing channels or customer coverage have changed and this is not reflected in the methodology.

The key question about the damage is which operations are blocked, who is denied service and whether there is a procedural way to fix it. The control team records which decisions were automated, which were actually human-reviewed, and which remained in line and affected the client. The legal characterisation of an AML process should not be confused with the statement that a person must disclose a suspicious operation to a particular subject; it is necessary to comply with the restrictions on confidentiality and tipping-off.

Appendix 29. Scenario: The agent changed the monitoring settings during maintenance

Condition (inventory). Operation AI agent gets the task of reducing the number of repeated notifications. It invokes an administrator tool and changes the alert routing configuration, after which some critical messages stop coming to the IB group. The agent's own log contains a brief explanation for "noisy events removed," but the original version of policy is not attached to the tikit.

The first line of investigation is de facto powers and a chain of action. We need to find out if the agent could change the rules without approval, what parameters of the tool he used, and why the side effect test didn't work. The agent's explanation itself is not a credible evidence of intent; it's a text artifact that needs to be compared to real calls. Possible causes include unclear task setting, lack of change-control, hidden instructions in an external document, and a problem with evaluating the significance of alerts.

The second is the preservation and restoration of evidence. The modified monitor can affect the recording of subsequent events, so independent cloud logs, audit log IAM, configuration snapshots, and system administrator messages are especially important. In order to check systemic risk, it is not only necessary to restore the old configuration, but also to establish a period of no warning and possible events that have passed without detection.

Test	Question	Artifact
Access control	Why would an agent have a right to change?	IAM snapshot
Change history	What parameters have been changed?	Diff configuration
Alert routing	What kind of alerts have not been delivered?	Independent delivery logs
Approval	Was there a human gate?	Approval history
Recovery	Is the surface recouped?	Control alarm result

CAPA includes reviewing powers, banning critical rules from being changed without independent agreement, isolated pre-deploy testing, and monitoring the integrity of monitoring settings. The necessity of external communications is assessed separately: the technical fact of the change itself does not yet prove the threshold of any AI Act, NIS2 or GDPR regimes.

29.1. Working procedure for verification

If an autonomous agent changes the monitoring configuration, they're studying not only the text of the model command, but the actual powers of the tool. Find out which service account was used, who assigned the appropriate role, whether someone checked the request and what restrictions were supposed to block the change. The initial and final configuration images allow us to determine whether the thresholds and routes of alerts have actually changed. The absence of messages from the detector after the intervention is not proof of the absence of attacks; perhaps the system just stopped detecting them.

The correction should include restoring monitoring rules, rotating compromised accounting data if necessary, applying the principle of minimum privileges and testing control events. The notification

of the regulator depends not on the unusual behaviour of the agent, but on the legal threshold of the relevant regime and the proven or reasonably probable consequences.

Appendix 30. Scenario: mistaken medical sorting of patient

Condition (inventory). The hospital has introduced a patient sorting aid system. After updating the entry directory, the model lowers the priority of the patient, although the medical card contains a severe symptom. The doctor agrees with the machine recommendation at a high load. A few hours later, the patient's condition is deteriorating.

The investigation divides medical causation analysis, technical reconstruction and regimen analysis of medical devices. Experts must obtain current clinical sorting instructions, accurate model input, history of the reference book, human decision information and medical records with confidentiality. The fact that the model was supportive does not prove the absence of its causal contribution. Similarly, a difficult outcome doesn't prove that another AI solution would have prevented deterioration: it's a subject of evidence-based medical evaluation.

Verification	Evidence	Expert
Clinical signs at the time of decision	Medical records	Professional doctor
Input/Output system	Versionized trade	AI engineer
Action by staff	Medical workflow history	Clinical auditor
Sectoral notification regime	Product/harm classification	Regulatory Specialist

Remediation includes medical impact assessment, manual adjustment, re-testing of subgroups, a safe manual reassessment procedure, and learning lessons from the incident. The reporting period defines the applicable sector rules, GDPR and AI Act separately. The timing of one procedure should not be mechanically postponed to another.

30.1. Working procedure for verification

In a medical scenario, you have to separate the medical truth at the time of the intervention from the retrospective knowledge of the diagnosis. The symptoms available to the doctor, the timing of data collection, system indications and instructions for use are considered. The technician reproduces the introduction, and the physician of the appropriate profile evaluates the clinical significance of the differences. If the model was used outside the approved application area, it could be an independent organizational cause other than an algorithmic defect.

In parallel, the status of the medical device is established and the applicable industry rules for vigilance/reporting. The complexity of the case is that timely reporting of a hazardous outcome and subsequent algorithm verification can be regulated by more than just AI Act. An off-the-shelf model update without preserving the previous version could destroy the evidence. When there is a direct health threat, medical action to protect the patient takes precedence over the completeness of the forensic experiment.

Appendix 31. Scenario: The generic summary changed the content of the legal instrument

Condition (inventory). The legal department uses AI to prepare brief references on large foreign court documents. The model incorrectly translates the denial from the paragraph of the order and writes that the court has confirmed the extradition of the person, although the original contains the refusal. The staff member sends the executive summary to management and plans for further action.

It's critical to keep the original decision, AI-concept version, the prompts, the procedure for legal verification and all derivative documents. The expert should not claim that the court has ruled otherwise; a certified source should be cited, language, date, legal force and available translations taken into account. The algorithm used is a processing link, not a primary legal source.

The link	Action during investigation
Original	Determine authenticity and applicable wording
Machine translation	Compare denials, dates and procedural status
Executive summary	Sclam-to-source links
Human verification	Find out if the staff member opened the original
Dissemination	Track the addresses'downstream and decisions

Recommendation: For process-relevant conclusions, AI-summary must contain a pointer to the exact page and line of the original document. Neither the number of sources in the app nor the reliability of the model formulation can replace authentication.

31.1. Working procedure for verification

For a legal resume with incorrect denial or changed date, a claim-to-source table is formed: each significant affirmation of the final document is compared with a specific fragment of the original. There are separate items missing from the original source and cases where the model mixed several different documents. An error may occur at the stage of OCR, machine translation, summarisation, searching for relevant fragments or even when editing manually. The researcher is therefore obliged to re-establish the sequence of change rather than to stop at the "negroste is wrong".

An important control is an examination of the role of the signatory. The presence of a human signature does not guarantee substantive verification, but the very fact of using AI does not prove the absence of professional supervision. To eliminate the damage, they correct downstream documents, send clarifications to recipients and evaluate the procedural ways of correcting decisions already made.

Appendix 32. Scenario: Deepfake as an alleged proof of action

Condition (inventory). The network receives audio recordings of alleged negotiations on an illegal operation. AI-detector marks the entry as "highly probablely genuine" and the other service as "probable synthesis". The management of the organization demands that the material be declared false immediately. At the time of the investigation, there was only a compressed file from the messenger without the original recording device.

The correct conclusion does not select a convenient detector. The chain of custody, the origin of the published copy, the expected period of creation, the independent circumstances of the negotiations, the methods of assembly and the applicability of audio forensics techniques to the format are required to be restored. One probability estimate from a closed algorithm should not be presented as an objective frequency of counterfeits in a court array. The origin may remain uncertain; it is a full-fledged result of the investigation if the limitations are well described.

Verification category	Limitation
Messenger file	Loss of reference metadata
AI deepfake detector	Dependence on the home of learning
Witness	Possible impact of the public version
Behavioral details	Indirect context, not "detector"

If the recording is used to restrict rights, the legal coordinator provides an opportunity to challenge the source and method. Otherwise, the "niger" said that the recording was real" was becoming a new and independent source of misconduct.

32.1. Working procedure for verification

If the recording is used to substantiate the accusation, it is necessary to distinguish claims about the assembly, the generation of individual elements and the complete syntheticity of the material. The expert procedure starts with the original file or the earliest available version and copying history. Metadata, container, coding, frame sequences and alternative sources are analyzed, but each method has errors and limitations. The commercial indicator of "the probability of a 97% dypfeuid" cannot automatically be translated into the procedural standard of evidence without published and relevant validation.

In the absence of a primary source, a limited inference about the properties of the received copy should be used. Additional evidence relating to time, place and persons shall be independently verified. In open terms, you cannot show potentially harmful personal data beyond what is necessary to understand the scientific conclusion.

Appendix 33. Scenario: A major language agent disclosed personal data

Condition (inventory). The support service uses an agency system that has access to an internal CRM and mail delivery tool. The user asks a question about his account, the agent mistakenly picks a record from the same name and sends its contents to an external address. The fact of the disclosure is confirmed by the SMTP log, but it is not known if the recipient used the information.

The investigation not only determines the reasons for the wrong retrieval, but also why the sending tool failed to verify the identity of the addressee and the amount of data attached. GDPR requires a self-assessment of the availability of personal data breach and the risk of individual rights. If the data are sufficient, the notification of the supervisory authority should not be delayed until an actual misuse of the disclosed information is established. The requirement to notify the subject under Article 34 depends on a separate high risk threshold and exceptions. [8], [9]

Artifact	Monitoring
CRM lookup trace	Identifier found
Email log	Real address and delivery
Permissions	Why Access Was Allowed
Risk memo	Justification for the need for notifications
Corrective action	Identity matching test

The correction requires the document to be tied to a verified subject, the agency tool's rights and the test for homonyms and mixed languages to be reduced. Closing a technical vulnerability does not eliminate the possibility of liability for disclosure.

33.1. Working procedure for verification

Leaking information through a language agent requires determining which personal data actually left the controlled environment, to whom it became available, and on what basis. A full prompt/response trace, a log of transfers to the tool, the level of service account rights and whether the data was stored by an external provider are important. The simple generation of personal text data does not always prove the disclosure of the original internal document. It is necessary to distinguish the actual reproduction of protected information from a random match or reconstruction from publicly available information.

In the event of a personal data breach, the terms of Articles 33-34 GDPR, the possibility of discontinuing further distribution, the need to inform the affected persons and measures to limit harm are considered. For a large language model, the mode of the model used and the role of the supplier should be further specified, but this does not replace the controller's obligations regarding personal data.

Appendix 34. Scenario: the rating model was subjected to data poisoning

Condition (inventory). The supplier of procurement analysts is training the model on open corporate data. An outside source begins to systematically post false information about the companies' ties; AI transfers these ties to a financial risk rating, after which legitimate counterparts get unreasonable restrictions. The investigation should distinguish between third party interference, insufficient data quality control and the user organisation's own decisions.

To verify, snapshots of sources and download dates, network and legal connections to publications, feature extraction history, and the timing of risk changes are recorded. The simple fact of fake internet publications doesn't prove a targeted poisoning; you need a proven causal sequence between the source, training or retrieval and the specific output. If there is no evidence of intent, it is more accurate to write "data-contagation" than "acta".

Hypotheses	Confirmation	Necessary reservation
Deliberate poisoning	History and communications	Thought doesn't come out of correlation.
Untested scraping	Lack of source satisfaction	Possible independent error
Violation of human oversight	Automatic locking	Uh, uh, decision analysis required.

Remediation involves correcting the source, checking for re-indexation, reassessing affected cases and notifying stakeholders when required by law. In publicly describing the incident, you can't re-disseminate defamatory claims in the form of investigative data.

34.1. Working procedure for verification

Suspicion of data poisoning is investigated through the origin of the data set, the history of its changes, the role of employees, the process of including samples and the characteristics of the verification set. Even the deterioration of the metric after the update does not prove a targeted attack: natural drift, markup error, audience change and incorrect validation are possible. The expert compares the hashes and manifests of the batch of data, the rules for admitting new sources, the history of claims, and the results of independent testing. If the model has a detected abnormal pattern, a controlled experiment is formed that checks its connection to a particular data set.

The outcome may establish the fact of the contamination of the set, but the intention and personal responsibility of a particular employee is determined by a separate legal procedure. Data clearance actions should keep the archives of previous versions as long as security and law permit.

Appendix 35. Scenario: When the model was updated, the possibility of replication disappeared.

Condition (inventory). The organization uses the SaaS model without access to a fixed version. After complaints of erroneous recommendations, the supplier replaces the basic model. The current

follow-up request no longer repeats an error. Management proposes to close the case because the "bug is off", although the original outputs are kept in client letters.

The absence of reproducibility in the current version is not a refutation of a historical event. The researcher records the model migration, the period of use of the previous version, the client documents and the integrator logs. If the provider is unable to provide a snapshot of the previous model, the impossibility of accurate replay is described as a limitation. The cause can be investigated by other preserved facts, but the exact internal mechanism of the inaccessible model cannot be established.

Source	Evidentiary role
Client letter	Saved External Result
Billing/API logs	Period of recourse to supplier
Release note	Confirmation of update
Integration history	Other components remain unchanged
Provider response	Boundaries of access to snapshot

The proposed organizational correction is contractual version support and retention of critical traces, including control outputs, while maintaining data minimization. Otherwise, any undetected vendor update could lose an audit of an object.

35.1. Working procedure for verification

After an undocumented cloud model update, the original sequence of actions may become unreproducible. The researcher asks for a contractually available history of versions, configurations and policies, and records when the supplier's response was received. If the model only indicates a marketing name without a stable version identifier, that's a separate significant limitation. The current version of the test is described as an experiment with a new object and does not serve as a final rebuttal to the original message.

To prevent repetition, the procurement requirements include an API version, routing logs, availability of archival specifications, and a procedure for storing evidence after a serious event. The absence of technical repetitiveness does not necessarily mean that all the facts are unproven: legal damage can be confirmed by independent letters, documents and bank transactions.

Appendix 36. Scenario: critical infrastructure and incompatibility of investigative objectives

Condition (inventory). In the power supply system, an automated AI forecast offers a sequence of switches. The operator rejects some of the recommendations, but after a minute, the station is shut down. The technical team demands that the planner be updated immediately, and the investigative team demands that each team be preserved. The situation requires a priority for human security and the continuation of energy supplies.

The check should separate the system's autonomous operation from the recommendation mode, the actual execution of switches, external causes and the correctness of the sensors. The naïve counterfact "without AI no shut-off" is not confirmed without account being taken of the state of the

physical network. In such a situation, safe replacement of components is permissible when maintaining control copies and explaining the reasons for technically unavoidable changes.

Priority	Measure	Examination tape
Protection of persons	Transition to a secure environment	Who approved and why?
Retention of the facts	Snapshot commands and sensors	Hash, time, limits.
Testing	Isolated simulation	Network model boundaries
Sectoral communication	Evaluation of special law	Independent Trigger Matrix

The report must recognize the difference between the statistical probability of an accident, the actual technical trigger and the legal liability of the operators. The learning outcomes are communicated to competent participants without publishing infrastructure hazardous details.

36.1. Working procedure for verification

Critical infrastructure requires a balance between keeping digital footprints and maintaining vital functions. Before the component is turned off, the command determines whether the isolation will pose a greater risk than continuing controlled operation. A pre-approved emergency profile is useful: separate logs, read-only-export, state mirroring and switching to backup manual mode. The participation of an independent expert shall not deprive the operator of the right to take an urgent technical decision in the event of an immediate threat.

The legal map identifies specific infrastructure security standards, NIS2 incident classification, and the possible application of AI Act. Not every decline in the quality of AI projections is a "serious and irreversible violation of critical infrastructure" in the legal sense. For this conclusion, facts about the nature of the impact and the conditions of the relevant definition are needed.

Appendix 37. Scenario: AI classification of mass complaints covered a serious problem

Condition (inventory). A large organization uses LLM to sort out client calls. The conditions of escalation are based on key words, and the new prompt is beginning to classify some of the reports of illegal freezing as a "timeline issue". The manual team sees a decrease in the number of complaints about serious limitations and concludes that the service has improved. Hundreds of outstanding stories are discovered after an external inquiry.

The incident is an example of monitoring, which itself has become a source of observational displacement. It is necessary to restore the initial sample of complaints, the previous and new instructions of the classifier, decision logs, actual processing times and the proportion of requests that did not reach a person. Checking only those appeals that the model identified as serious will not reveal missed cases. We need an independent random sample of all the calls.

Error	Evidence	Measure
Insufficient recall	Selective manual marking	Change control metric
Invisible Propt update	Changelog	Change-control
Weak human override	UI and load distribution	Emergency appeal route

A measure of recovery is to return to individual complaints, not just a change in the classifier. If automated sorting has been hiding appeals for months, the period of exposure, the number of potentially affected individuals and available ways to recover their rights should be assessed.

37.1. Working procedure for verification

When automatic sorting of complaints brings dangerous signals to the "low priority" category, the absence of complaints in the high priority line is not an indicator of safety. The expert asks for the full population of messages received, markup, distribution rules and the proportion of cases verified by the person. For retrospective sampling, independent examination of applications is used and different languages, channels and methods of submission are taken into account. The test should evaluate the particularly significant missed cases, not just the overall percentage of correct classifications.

Restoration includes reassessing previous complaints, connecting with potentially affected individuals within the allowed range and correcting escalation criteria. If the complaint had legal urgency, the delay due to machine classification may affect the availability of an effective remedy.

Appendix 38. Scenario: AI expert was not provided with data due to commercial secrecy

Condition (inventory). The independent expert was instructed to investigate the reason for the erroneous refusals of service. The supplier is willing to provide a presentation on the model's accuracy, but refuses to pass on raw outputs and requests, citing trade secrets. The client requires a categorical conclusion about illegal discrimination, although only user letters and general statistics are available.

The right solution is to structure the restrictions. The expert can confirm the adverse decisions being observed, check the contents of the public documentation, compare reports of victims and identify questions for the official requirement of evidence. But he can't determine the internal causality of the model without sufficient access. A protected audit with the NDA, a selection check, the participation of a court-appointed expert or an appeal to the competent authority, if permitted by law, are possible.

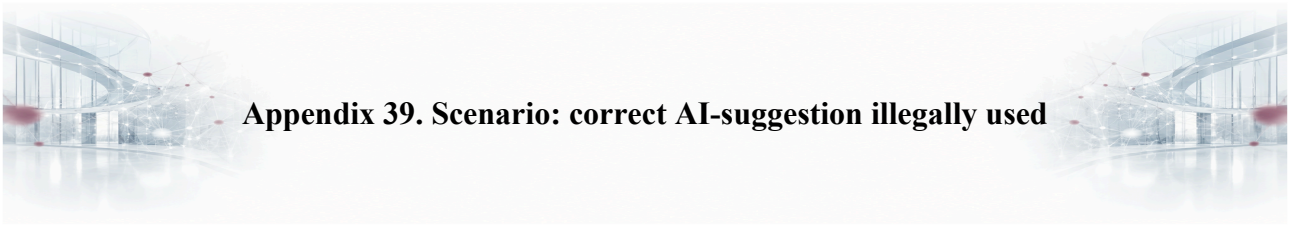
Access	Acceptable conclusion
Clear Output only	Fact of result and external consistency
Total statistics	Gross frequency with known limitations
Isolated test API	Conduct in a test context
Versified production logs	Renovation of a specific decision

In conclusion, the lack of data should not be disguised as a long list of common standards. Failure to provide the materials may have procedural consequences under applicable law, but a private expert is not authorized to automatically deduce the fault of the supplier.

38.1. Working procedure for verification

When a supplier refers to a trade secret to refuse logs, an independent report should distinguish between legal access rights, contractual restrictions and actual technical capabilities. The trade secret itself does not prove that information is missing or that the opposing party is obliged to believe the supplier's unconfirmed conclusion. There may be confidential access procedures, restrictions on the circle of experts, redacted reports, disclosure procedures or independent on-site verification. The specific option is determined by the applicable law and authority of the organ.

The report lists the materials requested, the answers and what key findings remain unverifiable. It is forbidden to replace the missing data with plausible reconstructions without any documentary basis. If independent verification is objectively impossible, the degree of reliability of the conclusions is reduced.



Appendix 39. Scenario: correct AI-suggestion illegally used

Condition (inventory). The organization uses a model to assess the likelihood of breach of contractual obligations. Statistically, the model works within the declared characteristics. However, the management uses forecasts to automatically refuse a service without considering the circumstances of a particular person and without the possibility of dispute.

In such a situation, the algorithmic accuracy and legality of the management solution diverge. The expert should check what grounds and procedures are required by law, whether the restrictions on use have been set out in the documentation and whether the operator understood the metric's purpose. The wrong legal procedure cannot be "corrected" only by raising the ROC-AUC: the model may become even more accurate, but the basic defect will remain.

Axis	Question
Technical	Does the system work on specifications?
Organizational	Used in an acceptable workflow?
Legal	Is the solution consistent with the applicable regime?
Remediation	Has the victim received a review?

This is an example of the need for two independent conclusions: a technical audit may not detect a model defect, whereas a legal analysis finds a violation of the organization's obligations. The mixing of conclusions impedes the fair attribution of responsibility.

39.1. Working procedure for verification

The model can accurately predict the statistical risk, but using such a conclusion could still be illegal. In this scenario, the technical and legal analysis is fundamentally divided. The data specialist checks the quality of the measurement and the validity of the scoring, and the lawyer determines the permissibility of the purpose of the processing, the basis for the use of personal data, the proportionality and procedure of human review. A discriminatory or illegal operation cannot be legalized just because the high accuracy is proven.

In the form of the conclusions, they specify separately: the accuracy of the prediction regarding a given reference; the applicable requirement to limit an automated decision; evidence of actual impact on a person; available measures for review. This protects against mixing a statistical description with a normative assessment.

Appendix 40. Scenario: Regulator dates started to count as a wrong date

Condition (inventory). The supplier received a message about a possible AI incident on Friday, but the official compliance team only registered it on Monday. The internal calendar counts the time since the ticket was registered, without checking the time of initial knowledge. In parallel, the operator has started its own investigation and there are separate GDPR requirements in place.

To calculate time limits, you need to determine the legally relevant moment of awareness for each subject and each regime. Keep email headers, support logs, time of receipt by the administrator, time of risk determination and legal classification data. Without this information, the communication cannot be declared to be timely or overdue.

Mode	What needs to be determined first
GDPR	When the controller found out about the personal data breach and what risk
AI Act art. 55	GPAI Status with System Risk and Serious Incident
AI Act art. 73	Scope, communication conditions and date of knowledge
DORA	ICT major classification and RTS exact triggers

The recommended measure is an independent legal deadline register with the initial time points attached. It is not possible to change an intake entry back for the sake of formal compliance.

40.1. Working procedure for verification

The situation is modeled after three participants who learned about the facts at different times: the service operator noticed a deviation, the IT specialist found an additional sign of leakage, and the legal service completed the qualification the next day. The researcher records the original timestamps, the transmission channel, the content of the information and the address within each organization. If the administrative responsibility depends on the knowledge of the controller or supplier, it is not possible to automatically adopt the date of "registration from lawyers" as the only relevant date.

Based on these records, separate calendars of the GDPR, NIS2, DORA and AI Act are drawn up, indicating the reasons why a particular regime is applied or not applied. The delays discovered are not hidden by the change of the intake record. The subsequent notification shall be accompanied by an explanation and documented correction measures.

Appendix 41. Comparison of ISO and NIST standards in the investigation programme

ISO/IEC 27037:2012 is devoted to the processing of potential digital evidence; ISO/IEC 27043:2015 - to the processes of investigating digital incidents; ISO/IEC 23894:2023 - to the management of AI risks; ISO/IEC 42001:2023 - to the management system of artificial intelligence. NIST SP 800-61r3 refers to an organization responding to cybersecurity incidents, and NIST AI RMF refers to a voluntary AI risk management framework. These tools complement each other and do not automatically create a cross-border procedural admissibility of evidence. [5], [7], [13], [14], [15], [16]

Phase	Tool	Practical boundary
Collection of digital objects	ISO/IEC 27037	Does not resolve the legal question of admissibility
Incidental procedure	ISO/IEC 27043	Total digital process, not AI specific
Action on Cybersecurity	NIST SP 800-61r3	Not a substitute for fairness and procedural rights
Management of AI-Risk	NIST AI RMF, ISO/IEC 23894	Not the law of quote
AI control system	ISO/IEC 42001	The certificate doesn't prove that every decision is correct.

It is proposed that the expert methodology indicate which specific recommendation has been used and how it has been verified. One phrase, "investigation is in line with ISO", is not sufficient: the scope and revision of the standard, the result of the test, the limits of independence and the rejection of automatic legal conclusions must be documented.

41.1. Conditions for correct application of the standards

Each standard needs to be linked to a specific type of activity: ISO/IEC 27037 describes the principles for dealing with potential digital evidence, ISO/IEC 27043 - approaches to investigation, ISO/IEC 23894 and NIST AI RMF - AI risk management, ISO/IEC 42001 - management system. Therefore, an expert should not write that the certification of a management system confirms the factual validity of an individual output or the admissibility of evidence in a criminal case. A sufficient confirmation of the use of the methodology is not a list of abbreviations, but a specific control protocol, document editing, performer and test result.

Appendix 42. Harmonization of methodology with human rights

The quality of the investigation is assessed not only by the number of logs preserved, but also by the availability of means of protection for the victim. The European Convention on Human Rights and the Charter of Fundamental Rights of the EU contain safeguards that, depending on the jurisdiction and actions of the authorities, may be relevant for assessing disproportionate

interference, procedural capacity and legal protection. The Council of Europe's Framework Convention on AI (CETS 225) creates an independent international legal architecture, but its status for the state should be verified by depositary data, without equating signature to ratification. [18], [19], [20]

Question	Verification	Why It Matters
Did the man know about AI intervention?	Notifications, interface, case files	The possibility of challenge
Is the primary source available?	Legal disclosure procedure	Equality of opportunity for the parties
Can the consequences be corrected?	Appeal, correction cannel	Effectiveness of the remedy
Are restrictions proportionate?	Legal framework and objective	Need and proportionality

When seeking asylum or being detained, the public disclosure of a name and fiction should be treated with great caution.

42.1. Conflict of technical closure and accessible means of protection

Effective dispute resolution requires that a person can understand the actual basis of the decision to the extent that he or she can raise objections. The disclosure of a complete model, special investigative methods or third-party information may be prohibited or unsafe. In the ARGA recommendation model, these interests are aligned through differentiated disclosure: the party receives the relevant facts and reasons, the court or authorized expert - confidential technical data with appropriate grounds, public report - aggregated information without personal identifiers. The availability of the correction procedure is verified on the basis of facts, not on the basis of a reference to the complaint form.

Appendix 43. Institutional system for the prevention of repeat AI incidents

The OECD Due Diligence Guidance for Responsible AI (2026) helps organizations tie management decisions to responsibilities of doing business. For investigations, the Identify - Prevent - Track - Communicate - Remediate cycle is offered in an adapted form, with the specific name of the stages in the ARGA methodology being an organizational choice, not a literal OECD norm. [21]

Phase	Control issue	Evidence
Risk definition	Who is affected by the process?	Impact map
Warning	What mistakes are we preventing?	Test Suite and Controls
Monitoring	How do we know about the damage?	Computers and allert log
Communication	Who and by what standard do we report?	Notification register
Corrigendum	Is there a specific harm?	Remediation tracker

For organizations with multiple AI products, a comprehensive incident/hazards registry with a single ID and separate legal regimes is appropriate. Public reporting uses only the permissible anonymized cut, and internal control keeps the originals and storage times.

43.1. Transforming investigation lessons into a monitoring programme

Every confirmed event should lead to a change in management control if the risk is repeated. For this purpose, the risk register shall include the reason, owner of the action, period, performance indicator and condition of re-checking. The result is related to the incident ID, but not limited to one command: changing the data source may require a reset of compliance, interface and contractual authority of the provider. Regular monitoring should not be limited to the number of tickets closed; the proportion of risks actually eliminated and the number of repeat episodes for the same reason is significant.

Appendix 44. Practical protocol for the private accounter

A data protection specialist must distinguish between an algorithmically unfair decision and a data breach. The first may affect GDPR, but it is not necessary to launch article 33. The second requires verification of definitions, risks, the controller's competence and notification obligations. In some cases, both regimes arise. The EDPB Guide 9/2022 should be used in its current final version, not in an outdated short narrative. [8], [9], [8]

Step	Question	Recording
Identification	Which PoAs have been modified/opened/lost?	Data inventory
Role	Who's controller and who's the processor?	Responsibility memo
Risk	Probability and weight for faces?	Risk memo
Decision	Do you want to notify DPA/persons?	Legal sign-off
Confirmation	What channel has been notified?	Receipt/ID

The processing of personal data is recorded separately during the forensics itself: there is no universal exception to the GDPR.

44.1. Specific DPO issues

The DPO determines the categories and source of personal data, the legal basis for processing, the outline of the controller and processor, and whether there has been a breach of security within the meaning of the GDPR. If there is only an incorrect scoring without any signs of breach of privacy, integrity or availability of data, the data breach procedure does not occur automatically, although other GDPR obligations may remain. For a leaked incident, a chronology, an assessment of the risks to the rights of the subjects, a record of the decision to notify the body and a separate assessment of the need for contact with the data subjects are prepared. All versions are retained.

Appendix 45. Report of whistleblower protection and internal error report

Some of the most significant information about a hidden defect comes from an employee who detects statistical distortions or systematic disregard for user complaints. If an organization requires that the signal be transmitted only to the head of the dispute resolution, there is a risk of concealment. An alternative protected route is recommended, compatible with applicable information protection rules.

Measure	Practical Meaning
Separate channel	The signal isn't dependent on the conflict leader.
Minimumization of identity	Protection against unjustified disclosure
Processing Log	Reported actions
Prohibition of destruction of primary signal	Retention of the facts
Access to independent review	Opportunity to challenge the closure of the complaint

You need to distinguish between honest internal communication and illegal disclosure of confidential information. The proposed practice does not give a person the unconditional right to disclose everything publicly; the specific protection is determined by the legislation of the jurisdiction.

45.1. Protection of internal informant

The employee's model error report is recommended to be kept separate from the technical package for broad audit. Information on the identity and circumstances of communication should be available to a limited number of recipients. The internal process involves confirming receipt, sending materials safely, banning unreasonable access by the team leader potentially affected, and documenting measures against negative consequences for the informant. Legal protection and disclosure conditions are determined by applicable national standards; this model does not replace the legal regime of whistleblowing.

Appendix 46. Qualitative evaluation of the unreliable AI-output

In the absence of the original logs, the expert may keep a screenshot, a letter, a statement and information about the production process. The quality of such materials is assessed step-by-step: the source is known or anonymous; the artifact was saved immediately or later; whether the original metadata is available; whether the transmission chain has been confirmed; whether there is an independent corresponding document; whether the information could be changed during conversion.

Level	Materials	What's admissible to say?
A	Full independent originals	Historical configuration reconstructed

Level	Materials	What's admissible to say?
B	Full internal logs with integrity checks	Technical output with limitations
C	Production files and agreed evidence	External event probable/confirmed
D	One unverified screenshot	Only the existence of an allegation

The A-D scale is proposed as an internal ARGA methodology and does not reflect the requirements of a court or international standard. It prohibits language substitution: a low level of evidence cannot be described by categorical wording.

46.1. Example of correct wording of uncertainty

If only the script and the oral reports were sent, the conclusion would be: "The material received shows that the sender has reported such a result; it has not been possible to verify the original system, its version and the actual way of making the decision." You can't write "IIA committed a violation" unless there's an original record, independent confirmation of model use, and testing alternatives. Additional information is ranked by what issue it solves, not by how many new files are in the folder.

Appendix 47. Report completeness test without disclosure of prohibited information

Before sending an external report, it's useful to check for sufficiency and excess at the same time. Sufficiency means that the authorized addressee can understand the event and start the procedure. Excess means the transfer of information that is not required for this purpose and can cause new harm or reveal critical vulnerability.

Monitoring	Question to author
Basis	Why does this body receive the document?
Competence	Could the authority request information?
Proportionality	Are all the transferred identifiers necessary?
Protection	The document doesn't contain active secrets?
Certification	Does every substantive allegation have a source?

For AI agents, it's useful to remove active API keys, precise bypass instructions, and access information from third-party accounts separately from public versions. The originals may be retained in a closed instance under a legal procedure.

47.1. Checking the completeness of open and closed versions

The open report is compared with the confidential stable identifiers of the statements and applications. For each hidden fragment, the reason for the restriction, the person who made the decision, and the effect of the cover-up on the possibility of external verification are noted. If the deletion changes the meaning of the conclusion, the open version must report a substantial limitation, not create false confidence. When transmitting the protected set to the regulator, the file version, control amount, transmission channel and receipt confirmation are indicated.

Appendix 48. Legal review of categories of automated solutions

The GDPR provides special guarantees for certain decisions based solely on automated processing and having legal or similarly significant consequences (Article 22), with exceptions. It doesn't match the concept of an AI incident under AI Act, and it doesn't necessarily depend on using a new generative model. Legal evaluation involves finding out the actual role of the person and the possibility of substantive intervention. [8], [8]

Question	Value for qualifications
Has the "only" automatic solution been used?	The human role must be real.
Had there been any legal or similar significant consequences?	Separate from normal personalization
Are there conditions for exclusion and guarantees?	Check applicable base
What are the rights of the subject?	Explanation and challenge by regime

Compliance with article 22 cannot be inferred from the staff member's record. It's just not possible to assume that any use of AI in customer service automatically falls under this article.

48.1. Classification of automated participation in the decision

For legal analysis, four situations are identified: the model generates the reference text without affecting the decision; the model recommends the decision, and the authorized person actually evaluates the original materials; the decision is taken automatically with the formal confirmation of the employee; the decision is fully automated. Such a scheme is designed to establish the actual role of technology and does not replace the provisions of Article 22 of the GDPR, where there is a significant legal effect and other criteria. The classification is supported by workflow materials, interface logs and regulations, rather than by a supplier's self-presentation as "human in the loop".

Appendix 49. Comparison of AI-incident public bases

Public databases help detect recurring problems, but differ in terms of inclusion rules, sources, deduplication, and quality of confirmation. OECD AIM uses newsletters and automated classification by a specific methodology; independent non-commercial catalogues may rely on other rules. You can't compare the absolute number of events without adjusting the sources and coverage time. [4], [4]

Sign	Minimum verification
Definition of incident	Does it take into account the actual Harm or the hazard?
Sources	News, reports, court reports, complaints?
Duplicates	How are the publications of one event combined?
Certification	Is there a status of confirmer?
Updates	Can the tape be corrected and withdrawn?

The practical value of a database is in finding repeating signs and formulating hypotheses, not in automatic proof that each supplier mentioned is guilty.

49.1. Methodological verification of open bases

News monitors and public registers are distinguished by the right to select incidents, the quality of primary documents, and the number of repeated reports about an event. For use in research, it is advisable to manually check a sample of the original reports, compare duplicates, date of event and date of publication, and also refuse to interpret the number of news as a measured probability of harm. A publication with an automatically formed description does not serve as a self-confirming causal link. The results of such databases are primarily suitable for detecting themes and monitoring signals, not for judicial qualification.

Appendix 50. Coordination with AI Office and national authorities

At EU level, there are various institutional roles to oversee GPAI and the systems used in individual areas. In describing the procedure, reference should be made to the current official powers of AI Office and the national competent authorities, without assigning an independent Observatoire the right to carry out an official investigation on their behalf. Articles 55, 73 and the market surveillance provisions require a separate definition of the addressee and type of communication. [1], [22]

Channel	Legal verification
AI Office	Is the event within the scope of oversight of the GPAI?
Market surveillance	The country of the incident and the high-risk system?
DPA	Is there a violation of personal data?
CSIRT/NIS authority	Is the national NEX2 flow chart applied?
Financial regulator	DORA and sectoral law requirements?

Each message sent is accompanied by a registration number, date, version and description of the application. The absence of a formal response does not constitute agreement with the expert's conclusions.

50.1. Model of communication with competent institutions

Each addressee is determined by a legal basis: AI Office in relation to AI Act, the national market surveillance authority, the data protection authority, the CSIRT or the financial regulator. There is no single "international AI regulator" for all cases. The package for transmission contains a specific incident ID, legal characterisation, recipient information, confirmation status, facts of damage, measures taken and an indication of unverified circumstances. If the document is subsequently supplemented, it is updated and the correspondence history is kept to confirm compliance with the relevant obligation.

Index of Practical Recommendations

Select the recommended task and proceed with the appropriate code. The code contains the report number and the main section number. Each review provides an example, analysis, action and outcome criterion.

19.01 Segregate event and explanation of cause
Section 1. Outcome: The conclusion separates the facts, the hypotheses and the confirmed mechanism.

19.02 Classify an event for proven consequences
Section 2. Result: No actual damage is reported in the registry without confirmation.

19.03 Record first information without indictment
Section 3. Outcome: The primary record distinguishes between the message and the verified information.

19.04 Record actual actions by participants
Section 4. Result: Each causal statement is related to the action of a particular participant.

19.05. Check the detector signal before escalation
Section 5. Result: Signal status changed based on the original recording.

19.06 First save the object then correct
Section 6. Result: The historical object is available for independent analysis.

19.07 Normalize time by source
Section 7. Result: The sequence of events is justified by the uncertainty of the clock.

19.08 Maintain the full context of the reply
Section 8. Result: The entry and transformation chain has been re-established for disputed approval.

19.09 Check the origin of the disputed topic
Section 9. Results: The report identifies the source of the discrepancy and lists the verified recipients of the adjustment.

19.10. Check the actual human control
Section 10. Result: The report established whether or not a person could change the outcome and at what stage.

19.11 Compare the different causal hypotheses
Section 11. Outcome: The underlying conclusion is supported by an observation that distinguishes it from alternatives.

19.12 Show statistical output base
Section 12. Result: The report does not contain a population frequency without a base.

19.13. Check the external exposure hypothesis safely
Section 13. Result: Confirmed or denied the mechanism of influence without unjustified attribution.

19.14. Segregate violation of access and defect of response
Section 14. Result: The type of incident is justified by the actual mechanism.

19.15. Localize the defect source by step
Section 15. Result: Rechecking covers the entire chain before the outcome document.

19.16. Check the legal threshold for facts
Section 16. Outcome: Qualifications contain motivation and separately the remaining unknown facts.

19.17 Establish a link between the incident and the appropriate model
Section 17. Result: Notification does not attribute the model to an unconfirmed cause.

19.18 Segregate the initial assessment and update of the notification
Section 18. Result: Copies of messages contain the date, facts and identified uncertainties.

19.19. Check personal data separately
Section 19. Outcome: Separate decisions on error content and data protection are preserved.

19.20 Maintain a parallel map of regulatory regimes
Section 20. Result: Each applicable route has a responsible and executing log.

19.21. Provide verified data with versions
Section 21. Result: The recipient can distinguish between the original assumption and the specific fact.

19.22. Check the base of each gear
Section 22. Result: A minimum agreed package with the recipient's log has been delivered.

19.23 Maintain the independence of expert research
Section 23. Outcome: The opinion answers questions on the material without the intended perpetrator.

19.24. Disparities between technical conclusion and legal responsibility
Section 24. Outcome: The opinion does not mix observation, causal conclusion and decision on guilt.

19.25 Provide a correction for the affected person
Section 25. Results: There is confirmation of correction or a reasoned description of the obstacle.

19.26. Check the identity of the different data
Section 26. Outcome: The decision is motivated by a set of data, the erroneous association is cancelled.

19.27 Do not replace official search with an algorithmic reference
Section 27. Result: The entry indicates the valid status of each source.

19.28 Segregate suspicion from established violation
Section 28. Result: Material distinguishes signal, verified facts and legal assessment.

19.29. Check Sorting Entry and Revision
Section 29. Outcome: Decision reviewed with corrected entry; result recorded.

19.30 Prioritize prevention of continuing harm
Section 30. Result: Reduction of current risk and investigation have separate responsibilities.

19.31 Check the origin of the disputed media file
Section 31. Result: The degree of confidence corresponds to available primary information.

19.32 Unlock the conflict and appoint an independent reviewer
Section 32. Outcome: The decision on the component was not approved by its previous consultant.

19.33 Protect sources when transmitting the result
Section 33. Outcome: The public text retains the conclusion without unnecessary disclosure of the identity.

19.34 Build a report around a proven causal chain
Section 34. Result: The reviewer reproduces a chain of argument on the material.

19.35 Confirm the corrective measure with a repeat test
Section 35. Result: Resumption is permitted after a documented verification result.

19.36 Account for direct and subsequent effects
Section 36. Outcome: All known recipients are checked; disputed amounts are not issued for established amounts.

19.37 Publication of status rather than unfounded accusation
Section 37. Result: The record makes it possible to understand the evidentiary status without reading a closed file.

19.38 Agree on the facts of an international notification
Section 38. Result: The recipients have an up-to-date version and corrections.

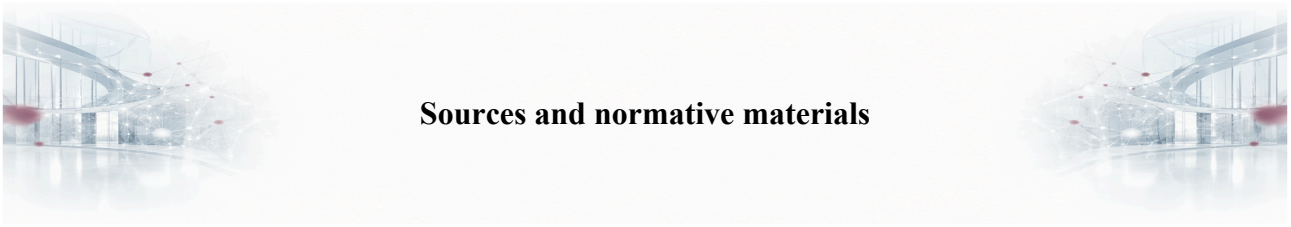
19.39. Check the five elements of the quality of the investigation
Section 39. Result: Each element is supported by a document or a specified restriction.

19.40 Establish terms of access and conclusion of the contract
Section 40. Result: The adoption does not require a fictional certainty in the absence of material.

19.41 Independently verify AI investigative assistance
Section 41. Result: Each substantive allegation is supported by accessible material.

19.42 Develop a programme to address identified gaps
Section 42. Result: The step is completed by a test result, not a calendar mark.

19.43 Close the technical process and remediation parts
Section 43. Result: Closing is justified for each part without general unreasonable status.



Sources and normative materials

The numbering in the square brackets in the text refers to the list of unique primary sources below; repeated references to the same act are combined. For mandatory rules, the official text with the relevant date of application should be used. Explanations by regulators, voluntary recommendations by the OECD and NIST, and ISO/IEC standards do not replace regulatory requirements. ISO references lead to official annotations; paid full texts of standards are not reproduced and there is no claim that they have been subject to full article examination.

[1] The European Parliament and the EU Council. Regulation (EU) 2024/1689 (AI Act), consolidated text of 27.07.2026; in particular articles 3(49), 26, 55, 73, 75, 113. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02024R1689-20260727

[2] The European Parliament and the EU Council. Regulation (EU) 2026/1744 of 08.07.2026 (Digital Omnibus on AI), including amendments to Article 113 of AI Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32026R1744

[3] OECD (2024). Defining AI incidents and related terms. https://oecd.ai/en/ai-publications/defining-ai-incidents-and-related-terms

[4] OECD.AI. Overview and methodology of the AI Incidents and Hazards Monitor (AIM). https://oecd.ai/en/incidents-methodology

[5] NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[6] NIST (2024). Artificial Intelligence Risk Management Framework: Generative AI Profile, AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

[7] NIST (2025). SP 800-61 Rev. 3: Incident Response Recommendations and Considerations for Cybersecurity Risk Management. https://csrc.nist.gov/pubs/sp/800/61/r3/final

[8] European Parliament and Council of the European Union. Regulation (EU) 2016/679 (GDPR), in particular Articles 4(12), 22, 33 and 34. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[9] EDPB (2023). Guidelines 9/2022 on personal data breach notification under GDPR, version 2.0. https://www.edpb.europa.eu/documents/guideline/guidelines-92022-on-personal-data-breach-notification-under-gdpr_en

[10] The European Parliament and the EU Council. Directive (EU) 2022/2555 (NIS2), Article 23. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2555

[11] The European Parliament and the EU Council. Regulation (EU) 2022/2554 (DORA), Article 19. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2554

[12] European Commission. Delegates to Regulation (EU) 2025/301 of 23.10.2024, art. 5: Time limits for DORA notifications. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32025R0301

[13] ISO/IEC 27037:2012. Guidelines for identification, collection, acquisition and preservation of digital evidence (official annotation). https://www.iso.org/standard/44381.html

[14] ISO/IEC 27043:2015. Incident investment priorities and processes. https://www.iso.org/standard/44407.html

[15] ISO/IEC 23894:2023. Artificial intelligence - Guidance on Risk Management. https://www.iso.org/standard/77304.html

[16] ISO/IEC 42001:2023. Artificial intelligence - Management system. https://www.iso.org/standard/42001

[17] European Union. Charter of Fundamental Rights of the European Union (2012/C 326/02). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

[18] Council of Europe. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS 225): text and overview. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

[19] OHCHR (2011). Guiding Principles on Business and Human Rights: Implementing the UN Protect, Respect and Remedy Framework. OHCHR catalogue. https://searchlibrary.ohchr.org/record/11593

[20] Council of Europe. CETS 225: information on signature, ratification and contractual status (checked at the date of application). https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225

[21] OECD (2026). OECD Due Diligence Guidance for Responsible AI. https://doi.org/10.1787/41671712-en

[22] European Commission. European AI Office: powers and functions. https://digital-strategy.ec.europa.eu/en/policies/ai-office

[23] European Commission, AI Act Service Desk. FAQ: the responsibilities of a systemic risk GPAI model supplier in the event of a serious incident (Commission position, not an independent legal norm). https://ai-act-service-desk.ec.europa.eu/en/ai-act/faq/what-are-obligations-provider-if-serious-incident-involving-its-model-occurs

Limitations on application

Legislation, standards and jurisprudence may change. Any finding that a particular person has a legal obligation, liability, or evidence is permissible requires fact-checking and the applicable law of the relevant jurisdiction. Model scenarios and application formats are training and analytical.

Scientific and Methodological References

[M1] NIST SP 800-86. Guide to Integrating Forensic Techniques into Incident Response. 2006. https://doi.org/10.6028/NIST.SP.800-86

[M2] Hernán M. A., Robins J. M. Causal Inference: What If. 2020. Author page of the public version: https://miguelhernan.org/whatifbook

[M3] Leveson N. System Safety. Lectures 6–7: CAST and Accident Analysis. MIT OpenCourseWare, 2016. https://ocw.mit.edu/courses/16-63j-system-safety-spring-2016/pages/lecture-notes/