Observatoire ARGA

INTERNATIONAL ANALYTICAL REPORT

Study No. 20

An International Model for an Independent AI Observatory

Public-Interest Monitoring, Performance Indicators, Risk Registers and Institutional Independence

Organisation: Observatoire ARGA

English

21 rue de l’Aviation, 64600 Anglet, France

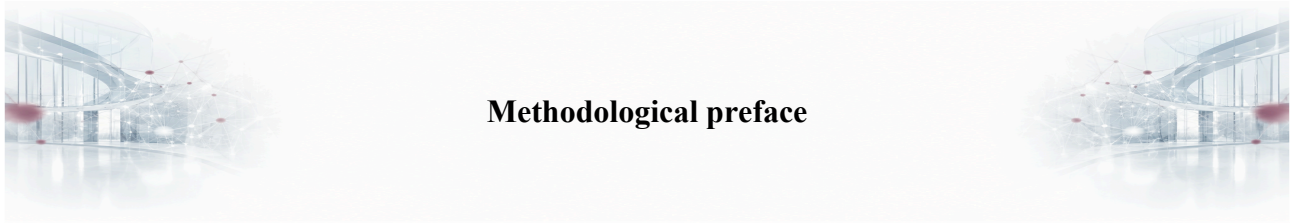
info@argaobservatory.org | www.argaobservatory.org

22 October 2026

Contents

Methodological preface — 3
Scientific grounds and research design — 4
1 Scope of the study and the research question — 5
2 Defining an independent AI Observatory — 7
3 International institutional reference points — 9
4 The legal nature of monitoring and its limits — 10
5 Objects of observation and the system register — 13
6 Organisational architecture — 14
7 Institutional independence — 16
8 Funding and safeguards against agenda capture — 18
9 Conflicts of interest and recusal — 19
10 Geographical coverage and comparability across jurisdictions — 20
11 A public-interest monitoring strategy — 22
12 Receiving reports and protecting reporting persons — 23
13 Event classification: incidents, hazards and signals — 24
14 Principles for a system of indicators — 26
15 Denominators, selection and observation bias — 27
Changes in the number of events with increased sources — 28
16 Collection, provenance and quality of evidence — 30
17 Information verification and the right of reply — 31
18 The risk register: structure and purpose — 32
19 Risk scales and prioritisation rules — 33
20 From monitoring to recommendations and rights protection — 35
21 Personal data protection and information security — 37
22 Vulnerable groups and participation of affected persons — 39
23 International data exchange and cross-border cases — 40
24 Algorithms in sanctions compliance — 41
25 International wanted-person alerts, identification and migration — 42
26 Financial monitoring and AML/KYC — 45
27 Public services and automated decision-making — 46
28 AI procurement and public-sector contracts — 47
29 Frontier AI and general-purpose models — 48
30 Democratic processes and information integrity — 50
31 Environmental, labour and societal impacts — 51
32 Engagement with reporting persons and civil society — 52
33 Publication procedures and categories of material — 53
34 Managing corrections, withdrawals and challenges — 55
35 Research quality control and peer review — 57
Analysis of agreement in independent coding — 58
36 Digital infrastructure and technical oversight — 60
37 Performance metrics for the Observatory itself — 61
38 Cooperation with international organisations — 62
39 The Observatoire ARGA pilot programme — 63
40 Financial and staffing feasibility — 65
41 Conclusions and the proposed international model — 66
Practical Annexes: Working Forms and Contingent Scenarios — 67
Appendix 1 Statute of the independent AI Observatory — 68
Appendix 2 Charter of Independence and Editorial Guarantees — 68
Appendix 3 Declaration of conflict of interest — 69
Appendix 4 Register of funding and related services — 70
Appendix 5 Matrix of competences and appointments — 71
Appendix 6 Passport of the monitored AI system — 71
Appendix 7 AI incident report record — 72
Appendix 8 AI-hazards and advance signals — 73
Appendix 9 Register of risks and control measures — 74
Appendix 10 Key indicator dictionary — 74

Appendix 11 Statistical calculation passport — 75
Appendix 12. Source quality classification — 76
Appendix 13 Evidence chain log — 76
Appendix 14 Report of independent record verification — 77
Appendix 15 Procedure for the protection of a sensitive complainant — 78
Appendix 16 Standard for Request for Comment and Right of Reply — 79
Appendix 17 Method of correction, withdrawal and replacement of the report — 79
Appendix 18 The applicability matrix for jurisdiction — 80
Appendix 19 Country Profile AI Government — 81
Appendix 20 Passport of International Comparison Methodology — 82
Appendix 21 AML sectoral screening and sanctions methodology — 83
Appendix 22 Protocol for the verification of AI systems in international tracing — 84
Appendix 23 Human control and appeal matrix — 85
Appendix 24 Form of external review report — 85
Appendix 25 Editorial passport of publication — 86
Appendix 26 Checklist to the regulator or public authority — 87
Appendix 27 Protocol of Emergency Analytical Excise — 87
Appendix 28 Public Dashboard Visualization Rules — 88
Appendix 29 Pilot plan for 12 months — 89
Appendix 30 Scenario: False Sanctions Convergence — 90
Appendix 31 Scenario: Risk of misidentification at the border — 90
Appendix 32 Scenario: Social benefit and non-transparent sormage — 91
Appendix 33 Scenario: Updating the model without re-evaluation — 92
Appendix 34 Scenario: AI-incident journalistic surge — 92
Appendix 35 Passport of the annual report of the observatory — 94
Appendix 36 Consolidated decision-making matrix on launch — 95
Appendix 37 Data flow chart and links between research registers — 95
Appendix 38 Harmonization of classification by independent evaluators — 96
Appendix 39 Verification of time series and structural gaps — 97
Appendix 40 Expanded matrix for assessing public harm — 98
Appendix 41 Request to IE supplier for system verification — 98
Appendix 42: Sensitivity analysis of the indicator with an example — 99
Appendix 43 Procedure for handling a request for data removal or limitation — 100
Appendix 44 AI incident liability model — 100
Appendix 45 Plan for independent methodological advice — 101
Appendix 46 Programme for the Independent Safety Assessment of the Archives — 102
Appendix 47 Method of verification of institutional applications for cooperation — 102
Appendix 48 Instrument for control of inter-lingual legal terminology — 103
Appendix 49 Method of auditing internal errors AI Observerry — 104
Appendix 50 End-of-financing sustainability plan — 105
Index of Practical Recommendations — 105
References and Supporting Sources — 107



Methodological preface

The study is based on official documents available as of October 22, 2026, and does not contain the results of its own international field survey, which has not yet been conducted for the model described. The empirical quantities inside the conditional applications are intended solely to explain the methodology. Lack of information on the system or the country is not assessed as a lack of risk. Particular attention is paid to the distinction between applicable law, non-binding standards and author recommendations.

The Protocol distinguishes surveillance, interpretation and causal statement. Confirmed information about the restriction of an operation allows you to describe an event; communication with AI requires a separate basis; inference about the cause and prevalence requires additional design. The researcher is obliged to indicate the level of each result. This discipline determines the choice of sources, the structure of the record, the rules for calculation and the language of public reporting.

Scientific grounds and research design

The observatory is viewed as an observational study with a documented mechanism for detecting events. Selective methodology sets the field of possible generalization; content analysis ensures the reproducibility of categories; measurement theory determines the meaning of indicators; institutional research evaluates conditions of independence. Law and research ethics establish the admissibility of collection and disclosure. The proposed design does not imply a pre-existing representative base. [M1–M3]

Scientific direction	Method	Quantified product
Selective observation	Description of scope, channels and exemptions	Structure of the shell
Contact analysis	Independent coding prior to discussion	Table of agreement and resolution of differences
Measurement theory	Definition of unit, numerator and denominator	Passports of each indicator
Institutional study	Verification of decisions in conflict of interest	Legend of drafting decision and revision
Law and ethics of research	Evaluation of admissibility of access and publication	A well-founded regime of data and protection of the person

Observed aggregate and unit of account

The study distinguishes between the message, the event, the source and the person concerned. One event may have multiple messages and sources; one material may describe several events. The rules of association shall be established before the calculation of the indicators. When the data is lacking, the connection is maintained as a hypothesis, and the disputed association is verified in sensitivity analysis. This prevents an artificial increase in the number of cases through repeated publications.

The observed population is described through object, period, territory, languages and channels. If public decisions and appeals are collected, the result characterizes the registered body in this collection method. To estimate the frequency in all operations, you need an additional marker and a valid detection mechanism. The unknown possibility of inclusion is not automatically corrected by the increase in the base volume.

Research issues and verified expectations

The pilot's first question is, can independent researchers include and classify new materials according to the same rule? Second, is the descriptive conclusion sustainable for rational decisions about sources, duplicates, and uncertain cases? Third, does the pattern of events recorded change when channels are expanded? The latter question does not in itself determine a change in real public risk.

To evaluate an improved instruction, independent results before and after its development are compared on different, comparable sets. The expected improvement is to reduce substantial differences without losing substantially important categories. Simply enlarging the classifier can increase consistency while eliminating the necessary distinction. Therefore, both the reproducibility and suitability of the categories for the research question are evaluated.

Independent marking and verification of the source

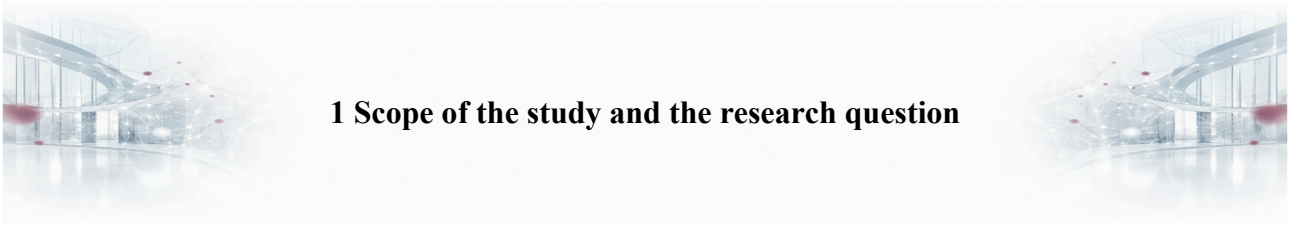
The instruction sets out the unit of analysis, permissible categories, rules for working with insufficient data and examples of border cases. Evaluators first work independently; then the reasons for the discrepancies are dealt with by keeping the original marks. Krippendorff's work justifies the need for meaningful verification of content-analysis reliability. [M1]. The choice of a consent coefficient depends on the scale, number of assessors and passes; it does not replace the content breakdown.

The verification of the source is separate from the consistency. Two evaluators are able to classify the wrong message in the same way. Therefore, the record contains the basis for confirmation, the status of the original material, the date and a substantial contradiction. The absence of independent validation remains a limitation even with the coders' full consent. In resolving differences, a new ground is established, not just the final vote.

Comparability and sensitivity of findings

For temporary or international comparison, the same meaning of categories, registration period, channel composition and availability of materials are checked. The result is recalculated on a stable subset of sources; new channels and unresolved cases are shown separately. The OECD and JRC recommendations on indicators are used as a basis to test the impact of selection of parameters and not to hide the instability of aggregate assessment. [M2]

The success criteria for a pilot is a reproducible procedure with understandable limitations and executable correction. Reducing real harm is a separate research outcome: it requires observation of consequences and a reasoned comparison. The calculations below illustrate the methodology; the results of international field research are not stated.



1 Scope of the study and the research question

Institutional challenge

The subject of the study is the ability of an independent observatory to form comparable and verifiable data on socially significant AI applications. Verifiability means the ability of another researcher to recover the source, the selection procedure, and the basis for classification. Comparability means the match of the meanings of categories and observation conditions in the comparing groups. The presence of a single website or a common form of record does not in itself establish these properties.

Question under review

The main research question is: what combination of sources, selection rules, and independent coding makes it possible to draw consistent descriptive conclusions with limited resources? The question of reducing public harm is raised separately and requires data on consequences and a reasoned comparison. The report develops a methodology for monitoring and a program for its pilot testing; it does not report on the proven effectiveness of the existing global network.

Results and limits

The report proposes a draft charter, a methodology for selecting cases, a classifier of sources, a system of indicators and a risk register. This is a model for further discussion and testing, not a statement that such bodies, processes or official powers have already been established at ARGA. The tools target cross-border cases and sectors where illegal automation can affect freedom of movement, access to financial services, property and procedural protection. Each proposal is tested on three criteria: public utility, reproducibility and cost of independent control.

Registration of research and conclusion criteria

The registration note records the research question, the type of analysis, the sources and the criteria for sufficiency. For causative research, hypotheses and verifiable alternatives are formulated; for descriptive work - rules of observation and classification; for legal analysis - jurisdiction, the editing of norms and method of interpretation. Not every documentary review needs a causal hypothesis. In the case of automated banking blocking, the restriction itself does not establish the application of a specific model: in the absence of such information, the conclusion can only refer to a confirmed procedure.

To verify the effectiveness of the Observatory, it is suggested to keep a log of research promises: what was supposed to be established that was actually verified and which sources were not available. It protects against subsequent rewording of the purpose to the result. All findings for external use are subject to the addressee's test: who is able to act on the recommendation and what evidence is objectively needed. If the addressee is required to have an official record or access to undisclosed logs, the publication should explain the scope of the investigation itself, not the inspection already performed.

Control issues and minimum evidence

Verification Element	Evidence criterion
Observed result	Tested publication with attached methodology
Not a result	Unjustified rating of organizations without data
Institutional limit	Lack of power and automatic access to closed systems

Key sources for comparison: [1], [2], [7], [20].

Practical case analysis 20.01

Recommendation: Ask a research question before collecting

Hypothetical example: The team gathers all the news about AI and wants to assess the extent of the damage.

Analysis: The news building does not automatically form a representative sample.

Practical steps: The curator limits the subject, period, languages and measured outcome; the goal is fixed before analysis.

Outcome criterion: The conclusion responds to the question and does not go beyond what is being observed.



2 Defining an independent AI Observatory

Three functions

An Observatory is a continuously operating analytical structure that collects and verifies information about the socially significant consequences of AI. The first function is documentation: recording objects, events and evidence with origin and date. The second is comparison: calculating metrics with the same definitions, periods and constraints. The third is a public explanation: reports, methodological notes, registers of open data and corrections to previously published conclusions. It's not enough to run a news site for quality control: news becomes an observation only after classification, elimination of duplicates, verification of connection with AI and description of uncertainty.

Derogation from other institutions

The scientific commission can assess the state of knowledge, the state supervisory body - to demand the execution of the legal norm, the certification body - to certify the established conformity, the judicial expert - to give a conclusion on the assignment in a specific process. The Observatory has the right to interact with all of them, but does not confer their status or attribute to its findings a legal force that they do not have. It is particularly important to distinguish between the words "monitoring", "audit", "conformance assessment" and "investigation", depending on the authority, responsibility and expectations of the addressee.

Working unit of observation

The basic unit is offered a "system - context of application - event - evidence" record. The same product can be used with different data, thresholds and measures of human control, which means different assessment objects. The record must contain the cut date, geography, decision owner, system version, observed effects, confirmation level and response status of the affected organization. International reports aggregate the records without erasing their unique proof power.

Status of findings and prohibition of false attribution of authority

The Classification of Products of the Observatory allows for three levels of approval. At the level of observation, an available fact is recorded: a contract has been issued, a complaint has been filed, an act has been issued, a specific inconsistency has been identified in the open documentation. At the analytical level, a reasoned conclusion is drawn about the risk or lawfulness that is not required to have legal force. At the recommendatory level, it is proposed to modify the practice or research protocol. The error begins when the third level is presented as a mandatory order or the first as a final assessment of guilt.

The public record should therefore start with one sentence on the status of approval. For an external reader, the language "certified", "investigated by an international body" or "officially verified" is particularly dangerous if the relevant procedure and authority were not available. A two-signature rule is proposed for a text containing a mention of a particular state, bank or international structure: the methodical editor checks the basis of the conclusion, the legal editor - possible impression about the author's status and powers. It is an inexpensive measure to prevent a significant proportion of reputational and legal risks.

Control issues and minimum evidence

Verification Element	Evidence criterion
Documentation	Data origin record
Comparison	Common definition of numerator and denominator
Communication	Public version with reservations
Status	Private research organization

Key sources for comparison: [1], [5], [6], [20].

Practical case analysis 20.02

Recommendation: Segregate Observation and Power Decision

Hypothetical example: The Observatory's publication is called an official finding of a violation.

Analysis: Research status does not create the power of a court or a regulator.

Practical steps: The editor describes the observations and recommendations; the legal conclusions are limited by competence and materials.

Outcome criterion: The public text accurately transfers the status of the study.



3 International institutional reference points

United Nations

UN General Assembly Resolution 79/325 created the Independent International Scientific Group on AI and the Global Dialogue on AI Management. The scientific group consists of experts acting in a personal capacity and prepares scientific assessments; the dialogue provides for multilateral discussion. These mechanisms are not a global regulator. For the independent Observatory, their principles of international representativeness, transparent mandate and separation of evidence from political decisions are useful. Interaction is only possible through the channels provided and does not imply institutional membership or recognition of a particular organization. [1], [2]

UNESCO and OECD

UNESCO's Global Observatory for Ethics and AI Management contains methodological tools and country profiles; RAM is designed to assess the readiness of state systems, EIA - to assess individual AI-systems. OECD.AI brings together policy instruments and incident/threat monitoring. It's a different scale of measurement: state maturity, product risk, and event documentation can't be summed up in one index. The proposed model uses these approaches as independent methodological guidelines and maintains direct links to the original sources. [3]–[5], [7]

Council of Europe and national authorities

HUDERIA helps assess the risks to human rights, democracy and the rule of law and is not an independent mandatory norm. The Framework Convention of the Council of Europe CETS 225 exists as a treaty instrument; its bindingness for a particular country is determined by the entry into force and participation of that country, not by the fact that the convention is published. National regulators, courts and ombudsmen remain subject to legal procedures. The Observatory can identify the problem and provide substantiated information, but the final conclusion about the breach is made by the competent authority. [9], [10]

Map of the functions of international institutions

A comparison of international mechanisms requires a function map, not a logo table. The UN scientific group has a research function, the multilateral Dialogue has a discussion function, UNESCO has a regulatory and methodological support, OECD.AI has a comparison of policies and events, state regulators have a national law application. For each possible contact, describe the expected type of input from the Observatory and the official channel. For example, the provision of aggregated data can be the subject of research exchange, and a written comment at a consultation is a form of participation without the right to use the institution's name as a sign of institutional approval.

The evidence base for the existence of a partnership is kept separately from the case register: official programme of the event, confirmation of receipt of a comment, contract or published joint material. The word "cooperation" is allowed only with an indication of the actual act. If an observer on his own initiative issued a review based on the UNESCO methodology, the correct description is "the principles and tools used" rather than "the study was prepared jointly with UNESCO".

Control issues and minimum evidence

Verification Element	Evidence criterion
UN Group	Independent scientific assessment
UNESCO RAM	Country institutional readiness
OECD AIM	Events, incidents and risks
HUDERIA	Context-dependent assessment of rights

Key sources for comparison: [1], [2], [3], [4], [5], [7], [9], [10], [21], [22].

Practical case analysis 20.03

Recommendation: Transfer institutional experience with resources

Hypothetical example: A small team wants to copy all the functions of a major international portal.

Analysis: Borrowing the structure does not provide funding and quality.

Practical steps: The manager selects a limited function, accessible sources and necessary review.

Outcome criterion: The pilot will implement the available composition without unconfirmed statements of scale.



4 The legal nature of monitoring and its limits

Applicable law

The right to conduct public research does not eliminate data protection, intellectual property, privacy, defamation and security requirements. In the EU, the analysis of personal data requires an applicable legal basis and compliance with the GDPR. The research objective is not a universal exemption from these responsibilities. The use of information requests, court documents and open registers should take into account the limitations of secondary use and the possibility of mistaken identity. In other jurisdictions, legal grounds are checked separately; the transfer of the European approach without reservation creates false confidence. [13]

Status of findings

The conclusion "there is a sign of algorithmic risk on available material" differs from the assertion that "the organization broke the law". The latter involves proper legal analysis and, in disputed situations, competent procedure. Observatory publications should contain standard confidence formulas: the fact is documented; the circumstance is probable; the hypothesis requires confirmation; the information is contradictory; the data is insufficient. The granting of status should depend on evidence, not on public resonation. It shall be indicated separately whether the assessment is of a legal opinion, scientific analysis or technical test.

Communication with AI Act

The EU AI Act sets out obligations for certain entities and systems, and an independent observer does not become a supervisory body on the basis of the publication of this act. After the 2026 changes, the terms of use of different blocks of norms vary; for high-risk systems, separate transitional stages are provided. The Observatory's checklist therefore records the applicable date, system category and role of the chain participant. It is forbidden to use a single calendar date as evidence of any breach of AI Act. In doubtful cases, the applicability map is published with reference to the current EUR-Lex version.

Legal review before information is collected

For each publication that touches on a potential violation of the right, a brief four-part legal map is offered: (1) established factual situation; (2) specific norm with date and circle of obliged persons; (3) question of applicability of this norm to established facts; (4) remaining restrictions and competent addressee. If any action is not confirmed, the editorial conclusion should not qualify the conduct as a violation. This is especially important for the rapidly changing AI Act, where different classes of systems and entities have different stages of application.

International observation should also check the rules for transboundary information transmission and professional secrecy. Even socially relevant material cannot be published without analysis together with the applicant's passport or confidential bank documents. For a small team, a separate log, "Found processing - target - time - recipients", is useful, which is approved before the online complaint form is opened. The existence of an informative legal footnote does not correct the illegality of the original collection.

Control issues and minimum evidence

Verification Element	Evidence criterion
Personal data set	Verification of base and minimization
Charge of violation	Legal analysis and response procedure
AI Act	System category, role and date of application
International publication	Assessment of the legal limitations of each jurisdiction

Key sources for comparison: [11], [12], [13], [23].

Practical case analysis 20.04

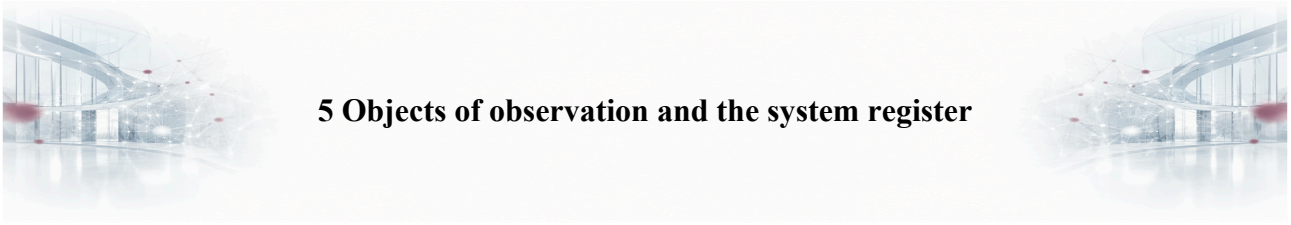
Recommendation: Check the legal basis for monitoring

Hypothetical example: The public registry plans to include sensitive applicants ' documents.

Analysis: The receipt of a document does not imply the right to disclose it to all.

Practical steps: The controller determines the basis for processing, the composition of the open data and the closed storage regime.

Outcome criterion: The public record contains only validly disclosed data.



5 Objects of observation and the system register

What to include

It is not the "IIA at all" that becomes the object of monitoring, but the detailed application with a socially relevant function. Priority is given to systems that affect access to justice, immigration status, banking relationships, social payments, employment, education, health and public safety. The register may also contain general purpose infrastructure models, but their evaluation is different from the assessment of a particular solution. For each record, you need a functional designation, the owner of the implementation, the type of model, the status of use, the degree of autonomy, the regulatory category and the geographical coverage.

Boundaries of inventory

The existence of a system in a registry does not imply illegality and should not become a public blacklist. The sources of confirmation can be official communications, procurement documents, impact assessment reports, court documents and technical publications. The company's press release is a statement from the company, but not an independent verification of actual use. When the applicability of AI is not confirmed, the record is marked as preliminary and not included in the statistics of the implemented systems.

Update

Over the life cycle, the system can change algorithms, providers, parameters, user categories, data volume and the legal role of the organization. The register shall maintain versions with the date of entry into force of the amendments and the history of the sources. Otherwise, two reports for different years will compare different technical products under the same trade name. The archiving system ensures reversibility of publications: the user can restore the basis on which the record had a specific status in the previous quarter.

Unit of account and version of monitored system

To avoid arbitrary expansion of the registry, an AI system passport is subject to a three-confirmation entry control procedure. The actual function of the system; the source of its exploitation; the connection with a particular context and subject is required. Public information about the model code is not required, but then the level of technical detail is fairly marked as unknown. For systems with one brand and multiple deployments, a parent product record and subsidiary use records are created. Otherwise, a change in a company's decision is mistakenly attributed to all the organizations that bought its software.

Each record has two independent dates: the "observation period" and the "last time of verification". The programme could be discontinued months before the publication of the study; leaving "useable" status would be negligent. The history of discontinued and replaced systems is preserved because it allows us to see how institutions respond to incidents. In the generalization, the number of systems has been confirmed to be operational on the selected date.

Control issues and minimum evidence

Verification Element	Evidence criterion
Identifier	Stable ID plus version history

Verification Element	Evidence criterion
Scope	Designation and affected group
Applicability	Real exploitation, not advertising
Time	Date of data, date of verification and date of publication

Key sources for comparison: [5], [6], [11], [15].

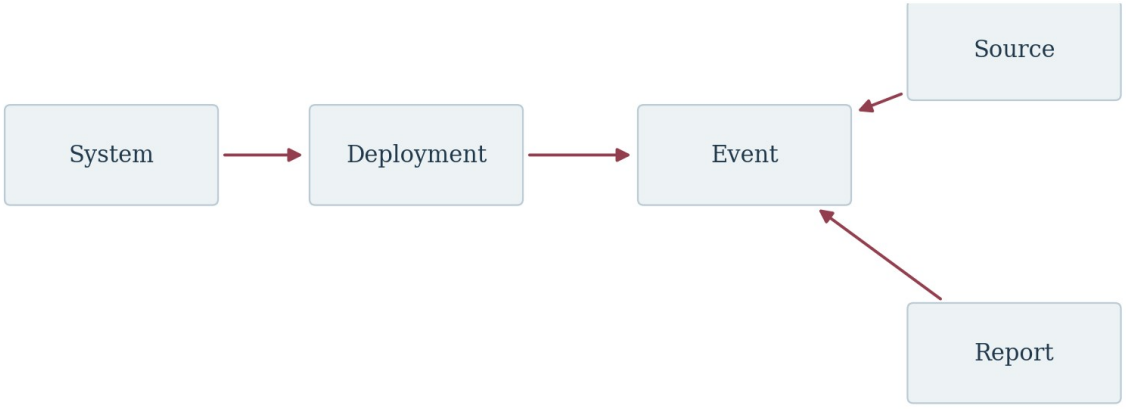


Figure 20.1. Relationships between the objects of observation. Model in Section 5. Reports, events, sources and affected persons are recorded separately; the number of publications is not the number of independent confirmations.

Practical case analysis 20.05

Recommendation: Separating the system and its application

Hypothetical example: One model is used by the bank and the migration intermediary.

Analysis: Developments involve different processes and actors.

Practical steps: The analyst creates separate records of patterns, deployments and events, connects them with identifiers.

Outcome criterion: The event is related to the actual operator and designation.

6 Organisational architecture

Division of roles

Four functionally separated links are proposed for the Observatory. The Scientific and Methodological Council approves the classifications and checks the independence of the methodologies. The Analytical Secretariat collects and encodes data. The Ethics and Data Protection Committee considers sensitive cases, security and the balance of rights. The editorial board decides to publish after checking the quality and procedure of the response. The representatives of a single project should not simultaneously appoint experts, conduct the collection of disputed data and authorize final conclusions alone.

Small launch

Full architecture does not necessarily require significant staffing in the first phase. When resources are limited, role sharing is achieved through written authority, independent external review and decision logs. For example, an analyst creates a record, a second expert conducts a blind review, a lawyer checks the risk of disclosure, and the editor approves the text. Even in a small team, it is unacceptable for a person who receives payment from a verified organization to independently approve the assessment of that organization.

Responsibility

Each decision must have the owner and the execution date: the date of registration of the incoming message, the date of appointment of the reviewer, the date of notification of the affected person, the date of closure of the verification. For substantial conflicts, there is an external independent arbitrator of the procedure. The powers of the council do not include rewriting scientific conclusions on reputational grounds; its task is to ensure correct methodology and ethics. The board's objection is documented and, if necessary, disclosed in the form of disagreement with the methodology.

Distribution of powers and right to stop publication

The allocation of powers should be anchored in a RACI-like matrix, where the key actions are identified as the executor responsible for the approval, the consulted specialist and the notifier. In the registration of an application, one responsible person is needed; in the legal evaluation, a profile competence; in the risk publication for a particular organization, an independent reviewer. If an employee performs two roles, the log shows this combination and compensating control. Recognition of limited human resources does not serve as a reason to cancel the conflict check.

The powers of institutional bodies should include the right to stop publication on documented grounds of security or a gross misrepresentation. However, the advice is not entitled to replace the scientific opinion with a "more convenient" one. The internal regulation should distinguish between stopping to eliminate a methodological error and trying to pressure the author. The minutes of the meeting shall state the reasons, the period for re-checking and the person responsible for returning the material to work.

Control issues and minimum evidence

Verification Element	Evidence criterion
Council	Methodology and independence
Secretariat	Data collection and storage
Ethics Committee	Risks for research participants
Edition	Right to publication and correction

Key sources for comparison: [1], [3], [14], [16].

Practical case analysis 20.06

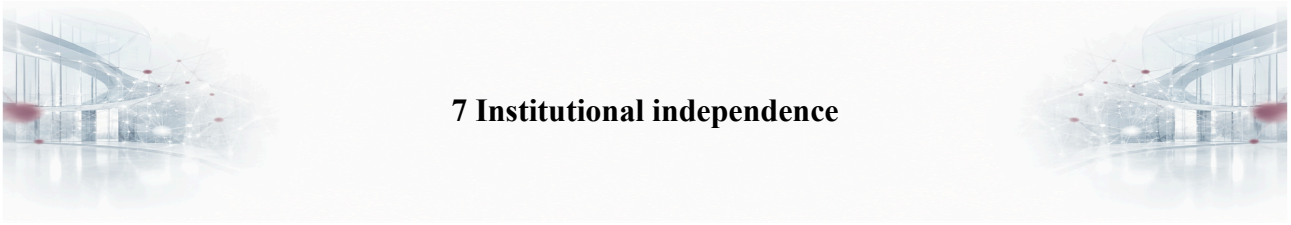
Recommendation: Designate different roles of verification and publication

Hypothetical example: One staff member accepts the complaint, checks it and approves the publication.

Analysis: The first-stage error is not independently monitored.

Practical steps: The curator shares training and approval; for a small team, he draws an outside reviewer.

Outcome criterion: A substantial publication has a second documented test.



7 Institutional independence

Independence as a tested system

The independence of the organization cannot be established by a single phrase in the statute. Restrictions on funding specific findings, disclosure of conflicts of interest, and rules of exclusion are needed. Independence is considered in three dimensions: financial (there is no possibility to buy a favorable conclusion), personnel (experts are not subject to the subject being examined) and editorial (the agreement does not allow results to be censored). The independence of political groups and bodies interested in restricting public access to information is assessed separately.

Risk of invisible influence

The donor may not formally require changes to the text, but may limit themes, deadlines or budgets. This is why not only cash receipts are published, but also substantial in-kind services, paid access to data, and the terms of closed studies. When there is a conflict of interest, a specific expert is assigned, the methodology is transferred to the second reviewer, and the fact of the assignment is recorded in a closed register. For public reporting, let's allow for the aggregate nature of disclosure when the employee's personal data needs to be protected.

Verification of independence

Once a year, the external reviewer evaluates the funding structure, the number of rejections, pressures on publication and compliance with the objection procedure. A special indicator is desirable: the proportion of projects whose contracts directly establish the independence of the editorial decision. But a high percentage does not in itself prove that there is no pressure. The outcome is formulated as an assessment of the procedures and not as an absolute guarantee of impartiality.

Special verification is necessary if the research Observatoire shares a brand, management, premises, financing or external counterparts with a structure that provides paid legal, advisory or representative services. Common belonging doesn't make independent conclusions impossible, but it does require the disclosure of connections and a proven differentiation of roles. The formal phrase "edition independent" could not be limited to: there should be written restrictions on access to client files, separate rights to approve publications and the challenge of those interested in the outcome of the matter under consideration.

Recommended organizational barriers: a log of related persons and paid work; a ban on the financing of a particular conclusion by a customer to whom that conclusion is beneficial; a mandatory external review of a controversial publication when intersecting with client work; a public policy of corrections equally accessible to critics and supporters. The customer or the donor may indicate an actual error and receive a reasoned answer, but does not have the right to veto on reputational grounds. Where such measures cannot be ensured, publication of the person concerned shall be carried out by an independent party.

The status of independence is actually described: the independence of methodological review, editorial decision, budget and management are checked separately. The presence of multiple

signatures within a single administrative hierarchy is not the same as an independent external verification. The annual report should contain the proportion of rejections, the number of conflicts reported, the proportion of research related to their own paid services, and cases of disregarded editorial pressure without disclosure of personal data.

Situation	Minimum rule of independence
Client case	Initial check of conflict, withdrawal of associated persons and independent review prior to publication
Donor or contractor	Disclosure of significant conditions of influence; lack of right to endorse the final conclusion
ARGA's own error	Application of the public correction procedure regardless of reputational inconvenience
Unable to break the conflict	Documented rejection or transmission to an independent researcher

Stress test on real editorial independence

The verification of independence is done through a set of control scenarios. First, the donor requests that the uncomfortable country not be included in the review. Second, a former developer proposes to review his own system. Third, the project manager demands to expedite publication before the verified party's response expires. Fourth, a well-known research institution asks to put its logo on before the content is agreed. In all cases, the rule must be clear in advance and not form based on the social significance of the participant.

It is proposed to register a significant intervention as a separate event of institutional risk. The source, the subject, the attempt to influence, the barrier and the final decision are described. In the annual public version, the data can be aggregated without revealing confidential correspondence. The external pressure itself does not guarantee the scientific integrity of the Observatory, but the ability to systematically limit such pressure is a necessary condition for trust.

Control issues and minimum evidence

Verification Element	Evidence criterion
Financial barrier	Prohibition of treaty control of findings
Human resources barrier	Derogation and second independent opinion
Editorial barrier	Protected right to publish methodological findings
Verification	Annual external evaluation of processes

Key sources for comparison: [1], [3], [14], [16].

Practical case analysis 20.07

Recommendation: Protect the right to an uncomfortable conclusion

Hypothetical example: The funding partner wants to delete the negative result of his product.

Analysis: Editorial intervention called independence into question.

Practical steps: The head shall apply the previously agreed rules of independence and record the request.

Outcome criterion: The decision to publish is based on evidence and approved procedure.



8 Funding and safeguards against agenda capture

Mixed model

It is proposed to combine basic institutional funding, targeted grants, paid methodological research and donations without interference with results. The basic part finances the registry, security and publication of inconvenient results that rarely attract commercial order. Targeted grants can only be granted on a predetermined basis and if the funds are publicly disclosed. The pay-as-you-go examination shall be separated from public monitoring and shall not render the non-evaluation-free product a "approved" product.

Donor restrictions

To prevent addiction, it is suggested to establish internal concentration thresholds, but not to pass them off as international legal requirements. For example, a 20% single-source income threshold can be used as a signal for increased verification rather than a universal boundary of independence. More important are the donor's contractual rights, ownership structure and ability to stop funding at a critical moment. With increasing concentration, the board formulates a diversification plan and exposes the risk.

The internal guideline of concentration cannot be transformed into a universal assessment of independence: a proportion less than 20% does not exclude editorial influence, and a proportion above 20% does not prove dependence. The methodology should simultaneously examine the terms of the contract, the right to terminate the grant early, the composition of the management body and the possibility of allocating costs between independent projects. The threshold should be reviewed regularly on the basis of actual data, while the percentage itself and the way it is calculated are disclosed as an internal management decision.

Division of services

An organization that simultaneously advises a developer and publishes an evaluation of the same project carries increased conflict. The permissible model requires team divisions, a ban on using unpublished client data without legal grounds, external verification of sensitive reporting and a transparent designation of the relationship. A "safety certificate" is not allowed following a registered review if the verification did not meet the certification process and applicable standards. A client can be sold a research service, but not independence of conclusions.

For studies involving individuals whose interests ARGA has previously or simultaneously represented, the priority option is either an aggregated study without human identification or an independent study that does not use client-directed information. The existence of a signed legal aid agreement does not in itself permit the transfer of data to the public research register. The legality of the secondary processing, confidentiality and possible conflicts of interest are assessed separately for each data set.

Verification of donor and customer dependence

For financial stability, it is recommended to consider not only the share of the largest donors, but also the substitutability coefficient of critical functions. If the only partner provides storage, transfer

or access to the registers, the organization may become dependent even on a small share of revenue. This dependence is described in the in-kind support register and evaluated over time to switch to an alternative provider. Separate budgets for research, information security and legal advice help make this threat visible.

The practical test of the contract contains a direct question: Does the agreement permit the publication of a documented negative conclusion about the customer's own activities? If not, the result should not be positioned as an independent audit. The permissible restrictions on secrecy are specified separately from the censorship of assessment judgments. If it is not possible to ensure editorial autonomy, the project is translated into a confidential consultation format without an external declaration on an independent assessment.

Control issues and minimum evidence

Verification Element	Evidence criterion
Grant	Public object and rejection of editorial control
Contract	Checking conflicts prior to operation
Donation	Register of significant donors and conditions
Threat of concentration	Signal for advice and diversification

Key sources for comparison: [3], [14], [16], [17].

Practical case analysis 20.08

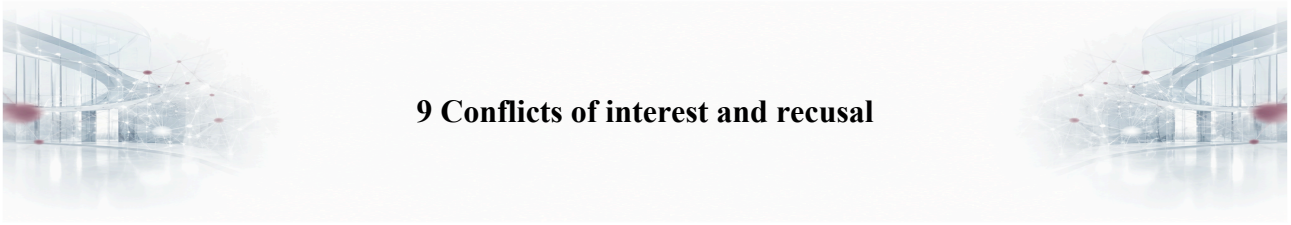
Recommendation: Unlocking the dependence of funding

Hypothetical example: Most of the budget comes from a single supplier of observed technology.

Analysis: Dependence can affect choices even without direct censorship.

Practical steps: The Council discloses the source, assesses the conflict and establishes an independent approval of the relevant investigations.

Outcome criterion: Funding and dependency management are available for verification.



9 Conflicts of interest and recusal

Categories

Conflicts of interest can be financial, professional, personal, political, institutional and informational. An expert who has previously developed a model is not necessarily without any competence, but his role requires disclosure and limitation when evaluating his own decisions. The conflict may include not only a direct contract, but also paid consultation, a dispute with a verified organization, competitor ownership of shares or the obligation to keep a trade secret. The register must contain the period of occurrence, type of interest, project concerned and decision taken.

Management

Before being appointed, the expert will file a declaration. The employee compares it to a customer registry, anonymous donations and previous projects. The result is no restrictions, participation with disclosure, limited role or complete withdrawal. If there is a serious conflict, the publication is carried out by an independent expert, and the primary expert can only answer technical questions on a documented request. The decision to exclude should not depend on whether the organization's conclusion was beneficial.

Subsequent conflict

After publication, new information on related persons or hidden funding is possible. In this case, a retrospective audit is carried out, the impact on the result is assessed and an explanation of possible correction is published. If it is not possible to confirm the independence of the key reviewer, the relevant statement is removed or marked as requiring re-verification. Circumstances are protected from unreasonable disclosure of personal data; it is enough to show the nature of influence and the process of managing it.

Protocol to challenge and document the conflict

A conflict of interest is not judged by an expert's moral attitude, but by the mechanism of possible impact on the outcome. For example, a specialist may have valuable industry expertise gained from working for a company under review. His involvement in explaining the technical context is permissible when disclosing the connection, but the decisive assessment of his own previous work must be carried out by another reviewer. Similarly, a sharp public dispute with the subject of the study can lead to a deliberate negative selection of facts.

In the annual assessment of conflicts, it is useful to consider not the number of declarations, but the proportion of appointments that have been substantially verified and cases of violation of decisions on removal. The external auditor has access to the anonymised register to assess the procedure without disclosing personal financial information of employees. If a new conflict is detected after publication, you can't just write down the reservation: you need to determine if it affected the choice of sources, calculation and conclusion.

Control issues and minimum evidence

Verification Element	Evidence criterion
Declaration	Before each new project starts
Deduction	Documented solution without conflict
Retrospective	Review of the disclosure of new relationships
Publicity	Adequate disclosure without excess data

Key sources for comparison: [1], [14], [16].

Practical case analysis 20.09

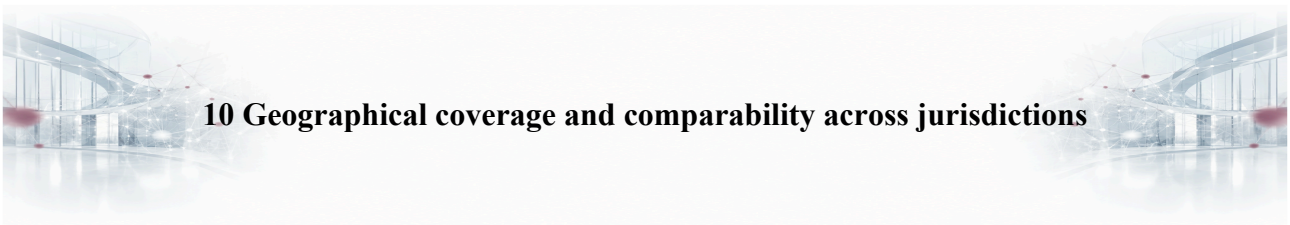
Recommendation: Apply a challenge on a specific issue

Hypothetical example: The reviewer advised the organization that is included in the study.

Analysis: The general declaration of interests does not resolve the question of participation.

Practical steps: The Coordinator shall appoint an independent reviewer for the disputed part and shall record the challenge.

Outcome criterion: The concerned party does not adopt a conclusion on its case.



10 Geographical coverage and comparability across jurisdictions

Conditions for international comparability

The international comparison begins with the verification of a common unit of observation. In one country, sources may describe appeals, in another - the decisions of the regulator, in a third -

individual episodes of harm. A table of the conformity of the categories and rules of incorporation shall be drawn up before the comparison. Cases that cannot be compared without loss of significant information are analyzed separately. Jurisdictional differences are not hidden within the same indicator name.

Inter-State review of categories

For the common body, interlinguistic testing is offered: researchers of different language groups independently classify one set of translated materials, preserving the originals and disputed terms. Differences can be traced back to sources such as translation, legal concepts, lack of data or ambiguity of instructions. The success of such a set checks the consistency of the procedure, but does not eliminate differences in the availability of real sources between countries.

Comparison over intervals

Instead of ranking countries by a single number, it is suggested to publish profiles and intervals by metric groups. Where there are different legal grounds, the assessment remains qualitative with a clear explanation of incompatibility. International aggregates are only considered for countries that have met the minimum reporting requirements and are accompanied by a list of excluded jurisdictions. This approach is less effective for headlines, but much more reliable when used by lawyers and international organizations.

Comparability of the country, period and legal framework

It is recommended that a selection protocol be issued in advance to compare countries. You can choose from a regional balance sheet, from the maturity of public registers, or from the presence of a specific automated function, but the outcome of the report depends on that choice. If the analysis is limited to countries with advanced judicial openness, it reflects the availability of evidence, not necessarily the best or worst state of AI governance. Each international schedule should have a list of countries and the key reasons for excluding the others.

In the legal profile, it is useful to code separately "rule accepted", "rule begun to apply", "body competent", "public execution documented". These four features may have different meanings at the same time. Time is also important: the regulatory situation at the date of the incident and the date of the investigation may differ. The Observatory must keep both versions and not evaluate past action under the new norm without a specific legal basis.

Control issues and minimum evidence

Verification Element	Evidence criterion
Law	Status of the act and date of application
Institutions	Competences and complaints procedures
Practice	Judicial and regulatory evidence
Data	Completeness and limitation of surveillance

Key sources for comparison: [4], [5], [6], [10], [11].

Practical case analysis 20.10

Recommendation: Compare countries with the same coverage

Hypothetical example: In one country, they study court decisions, in the other, only the news.

Analysis: The differences in the number of cases may reflect the method of collection.

Practical steps: The researcher forms a comparable layer of sources and shows the remaining materials separately.

Outcome criterion: The intercountry conclusion is accompanied by an observational check.



11 A public-interest monitoring strategy

Three collection channels

Monitoring is constructed from three independent channels: regular review of public documents and court decisions; voluntary appeals and communications from civil society organizations; targeted studies of systems chosen according to a risk-oriented plan announced in advance. You can't replace the third group with random media stories, because that's how you get the sensation effect. Channels retain their own coverage and percentage of unverified messages. The priority of the studies is determined by the potential severity of the damage, the number of persons affected and the availability of independent verification.

Open and limited data

The open-source mode does not involve automatic mass copying of personal data. The Observatory keeps a short extract, bibliography, date of access and cryptographic print of the material being verified, if justified. Access to confidential data is only possible if there is a legal and contractual basis, a prescribed storage and distribution period. The results are published after the identifier has been removed or in aggregate form if personal data is not required to protect the public interest.

Drafting cycle

Each cycle consists of observation registration, pre-qualification, verification of AI relevance, independent confirmation search, harm assessment, sending a request for comment and a decision to publish. For systematic reviews, the rules of inclusion and exclusion are published, not just the stories that have been the most effective. If the observation does not meet the standard of evidence, it can be included in the closed-loop study, but not in the confirmed cases section.

Pre-registration of the monitoring programme

A practical rule of thumb for working with multiple channels is to create a common model of events while preserving metadata sources. One court case, two newspaper reports, and a citizen's appeal can all refer to one story; in aggregate it is one event record with multiple confirmations. At the same time, a series of similar mistaken decisions by one agency may require separate accounting for episodes. The selection of the unit shall be determined prior to the calculation of trends.

A "sector-source-test frequency" map can be drawn up for monitoring planning. For example, court documents are checked in a certain period; reports from applicants are received continuously; the research sample is determined in advance. The lack of resources is reflected in the coverage rate, not hidden. Without such a map, it's impossible to distinguish the change in the number of events from the change in the intensity of the Observatory's own work.

Control issues and minimum evidence

Verification Element	Evidence criterion
Regular channel	Reports and judicial sources
Cross-references	Secure form of incoming messages
Target sample	Pre-recorded selection criteria
Publication	Certification and right of reply

Key sources for comparison: [7], [8], [13], [20].

Practical case analysis 20.11

Recommendation: Fix monitoring channels

Hypothetical example: In the middle of the year, they added two languages and twice as many sources.

Analysis: The increase in registration may be due to an increase in coverage.

Practical steps: The analyst maintains a change log and calculates the series through stable channels separately.

Outcome criterion: The chart marks the change of coverage and doesn't give it away for increasing risk.



12 Receiving reports and protecting reporting persons

Form of referral

The form of admission should ask for a minimum set of information: a description of the solution, the context of AI use, an approximate date, the type of intended harm and the available confirmation. No passport, address or full correspondence should be required before the need is assessed. The user chooses consent for feedback and separate consent for citation; these actions do not replace the legitimate basis for data processing, but help manage expectations. An anonymous or pseudonymal channel is provided for reporting the risk of persecution.

Reliability Check

The report is not considered to be an established incident. The analyst separates the applicant's own experience from assumptions about the internal workings of the system, checks the sequence of dates and the existence of alternative explanations. If the report concerns the actions of a public authority or bank, you should ask for available decisions, letters, information on human verification and dispute settlement. When verification is not possible, the observation remains confidential and does not affect the public statistics of confirmed violations.

Protection against damage

The publication may cause additional harm to the applicant: disclose migration status, medical information, financial situation or participation in the process. Therefore, the risk of re-identification is assessed by combining indirect features. Contact details are kept separate from research cards, access is logged, and deletion dates are recorded in advance. The principle of public interest does not imply the unconditional right to reveal the whole history of a particular person.

Safe reception of messages and confidentiality boundaries

A two-tiered procedure is proposed for reporting from victims. A minimum signal shall be accepted at the entrance without the requirement of identification. If the event has research potential, the person responsible explains the objectives and limitations, specifies the contact permissible and only then asks for additional documents. The applicant is offered the right not to participate in a public history without automatically losing the opportunity to obtain information about the official procedures available.

The technical solution must prevent the accidental spread of the inputs: closed channel, separate storage, minimum notifications with the contents of the case. In case of high threat, the applicant is allowed to transmit general anonymised information without personal contact. The report cannot promise to provide legal representation, delete a record or change a government decision only on the basis of registering an appeal with the Observatory.

Control issues and minimum evidence

Verification Element	Evidence criterion
Minimum data	Do not request inertia documents
Status of the communication	Established / credible / not confirmed
Publication	Separate evaluation of re-identification
Storage time	Link to the purpose of the process and procedure

Key sources for comparison: [7], [8], [13], [18].

Practical case analysis 20.12

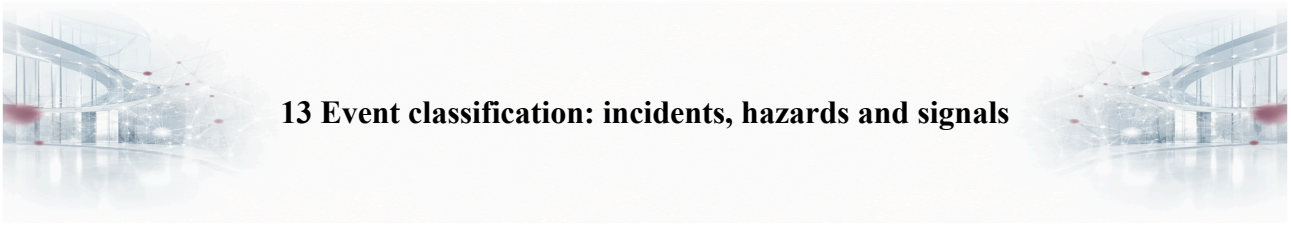
Recommendation: Segregate message reception and confirmation

Hypothetical example: The complainant sent an emotional story without an initial decision.

Analysis: The story is important for verification, but it doesn't prove all the circumstances.

Practical steps: The receiver stores the message, clarifies the available documents, and explains how to transmit it safely.

Outcome criterion: The record contains the verification status and the next specific request.



13 Event classification: incidents, hazards and signals

OECD Taxonomy

The OECD methodology distinguishes between AI incident, when development or use of AI has led to real harm, and AI hazard, when harm is likely under certain conditions. The Observer retains this distinction and introduces an additional internal class "signal" if the particulars do not yet allow any status to be established. It is important not to turn the absence of established consequences into evidence of security. A single signal can go to hazard or incident after receiving documents, but the history of this reclassification remains.

Communication with technology

For AI event status, you need to explain the role of the system: direct decision making, recommendation, automatic ranking, text generation, or indirect impact on the process. For example, an official's wrong decision doesn't become an AI incident just because the institution uses

digital services. On the contrary, an algorithmic recommendation can be causal even with formal human involvement. The key question was what alternatives and actual actions were available to the operator.

Combination of duplicates

Media posts about the same story can create the appearance of dozens of independent incidents. For de-duplication, the event, organization, time, geography and primary document are used, but the association must be cautious: the same mistakes of different people sometimes represent a series of separate decisions. The basic event record binds the sequences and indicates the calculation rule. You can't change this rule between reports without re-evaluation of the time series.

Segregation of actual harm and potential threat

In classification, not every use of automation is considered causal, but a specific link to the resulting or potential damage. The analyst formulates competing scenarios: an error in the original data, an unsuitable algorithm, incorrect operator configuration, lack of independent review, and independent human decision. In the absence of internal documentation, damage can be established and the role of a specific AI component unproven; these statuses are recorded separately.

For incident trends and hazards, it is proposed to publish a status transition matrix. If the event at the time of receipt was a signal, then became a hazard and finally an incident after the court decision, the original versions are not deleted. Otherwise, the historical report will look as if the editorial board had the final information from the beginning. The review of status has its own date and evidentiary basis.

Control issues and minimum evidence

Verification Element	Evidence criterion
AI incident	Documented relevant harm
AI hazard	Plausible possibility of harm
Signal	Insufficient evidence
Deduplication	Steady rule of events and subpisodes

Key sources for comparison: [7], [8], [19].

Practical case analysis 20.13

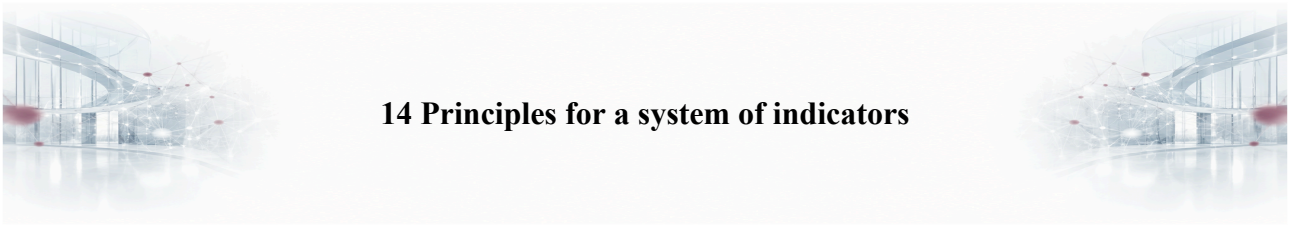
Recommendation: Exploit the hazard signal and incident

Hypothetical example: The publication describes the possible damage, but does not contain any information on its occurrence.

Analysis: The potential scenario and event required different categories.

Practical steps: The coder applies the definitions, preserves the basis and passes the dispute to review.

Outcome criterion: Category confirmed by observed signs, uncertainty maintained.



14 Principles for a system of indicators

Quality is more important than number

Each indicator gets a research definition: object, numerator, denominator, period, source and acceptable area of conclusion. "Number of messages", "number of unique events", "number of confirmed episodes" and "number of persons affected" describe different values. They cannot be combined or compared as interchangeable risk measures. The records should allow you to restore the transition from the original message to the final unit of account.

Selection of indicators for the research task

The quality of the indicator is assessed according to the relevant research issue. It's useful to load the edition with the number of entries; for the structure of recorded events, it's useful to distribute categories; for populations to be distributed, you need a well-founded denominator and selection mechanism. It is proposed to start with separate indicators. The composite index will require

additional weighting, processing of passes and sensitivities, as well as an explanation of what decision it supports. [M2]

Interpretation

If the number of registered complaints has increased from 20 to 40, without the number of users and changes in the communication channel, it cannot be concluded that the risk has doubled. The credibility of the procedure and the accessibility of applications may have increased. Dashboard is therefore required to show the completeness of the identification and the quality of the sources. If it is not possible to estimate the denominator, the absolute number of observations is published, marked as "non-wide". Substitution of such a number by the frequency of the violation is considered a methodological error.

Indicator passport and denominator verification

For each KPI, a record is formed with examples of permissible and unacceptable use. For example, "the proportion of decisions with an accessible appeal procedure" requires a clear definition: is there a legal rule, a published instruction, or is there a confirmed possibility of a person actually being reviewed? These three indicators are not comparable. You can publish all three, but with different names. The replacement of them with the general term "protection level" would lead to a false conclusion.

The individual KPI of Observatory quality is the proportion of published cards that can open a primary document or obtain a description of it with a justified access restriction. It's more useful than the number of press releases. For confidential material, an independent documentary verification is sufficient, but the status of such a confirmation must be clearly different from that of a fully public source.

Control issues and minimum evidence

Verification Element	Evidence criterion
Recorded incident	Number of confirmed cards for the period
Scope	Proportion of public sources processed
Procedures	Proportion of cases with a documented response
Consequences	Revision result with missing data

Key sources for comparison: [6], [7], [8], [14], [19].

Practical case analysis 20.14

Recommendation: Determine the value of the indicator before calculating

Hypothetical example: The panel shows the number of victims, counting the number of reports.

Analysis: One person could have sent several messages.

Practical steps: The analyst shares the message, the event and the person; introduces rules of association and pass.

Outcome criterion: The indicator has a definition, a unit and a reproducible calculation.

15 Denominators, selection and observation bias

Denominator problem

The fraction or frequency shall be calculated only after the numerator and denominator have been agreed upon. If there are 30 confirmed episodes, but we don't know the number of resolutions and the full extent of the detection, let's assume that the episodes were found. The risk assessment for all users does not follow from these data. Even with a certain number of operations, it is necessary to check the match of the period, units of account and coverage; the number of clients per year does not replace the number of resolutions per month.

Mixed Sources

Sources form different selection mechanisms. Complaints depend on the availability of the application and the willingness of the applicant to disclose information; news - on the editorial interest and language coverage; official decisions - on the authority and priorities of the body. The match of an event in multiple channels enhances its documentation, but does not automatically eliminate the shifting of selection. Each channel retains the origin, detection rules and connection to previously recorded events.

Sensibility

Sensitivity analysis shows how much the conclusion depends on the controversial decisions of the researcher. Results are calculated separately for inclusion and exclusion of uncertain events, alternative combining of dubs, and constraining by a stable set of sources. Scenarios are created by a meaningful change, not just an arbitrary percentage. If the direction of the conclusion is changed, the terms of that change and a specific additional verification to resolve the uncertainty are published.

Analysis of selection errors and non-extractionability

Comparing periods requires keeping the unit of observation and the collection mechanism or explaining their change. The doubling of the cases after a new language is launched may reflect an increase in detection. For analysis, a separate series of common sources is constructed and the contribution of new channels is shown separately. Such a breakthrough does not restore unknown events from the past and does not prove the causal effect of expanding observation.

When the probability of inclusion of each event is unknown, a large base does not become representative automatically. Weight correction requires substantiated information about the selection mechanism; arbitrary weights do not eliminate displacement. A permissible result may be a description of the structure of the registered body and an analysis of the mechanisms of individual cases. The risk frequency statement is deferred until a suitable design and identifier is available.

Changes in the number of events with increased sources

A case study shows why the increase in recorded events is not equal to the increase in risk. In the first period, 20 confirmed events were detected from a set of sources. In the second period, an additional language channel was introduced; the total number of events increased to 40. Once the duplicates are eliminated, the channel contribution is as follows.

Source set	First period	Second period
General for two periods	20	22
New Language Channel	I wasn't.	18
Total registered	20	40

The total number increased by 100 per cent, while the overall source growth was 10 per cent. The second number is also not an assessment of the change in real risk: publication activity, document availability, and the number of relevant transactions could change. The correct conclusion describes the increase in the registered body and the contribution of the expansion of observation. Comparing the actual frequency will require a comparable denominator and studying the completeness of the detection.

The report retains both rows and the date of the change in the fee. New events are not deleted for comparison, but the analytical conclusion on a stable set is separated from the full volume of registration. If common sources have also significantly changed the practice of publication, even this series gets an additional limitation. The transparency of such changes is more important than the visibility of a continuous schedule.

Control issues and minimum evidence

Verification Element	Evidence criterion
Available denominator	Publicize frequency with confidence interval
No denominator	Leave absolute number of observations
Sources are mixed	Selected series and clear warnings
Weak resilience	Interval instead of point output

Key sources for comparison: [7], [8], [19], [20].



Figure 20.2. Corpus selection and the limits of generalisation. Diagram for Section 15, without invented counts. If inclusion probabilities are unknown, the number of records does not establish the prevalence of harm.

Practical case analysis 20.15

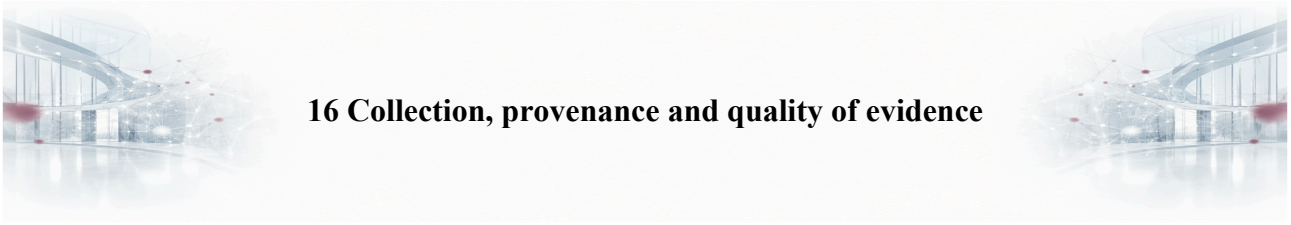
Recommendation: Specify denominator and selection mechanism

Hypothetical example: Of the 10 applicants, eight reported a bank blockade.

Analysis: That's the percentage of hits, not the percentage of all blocked clients.

Practical steps: The statistician signs the observed population and does not carry the result to the entire population.

Outcome criterion: The external text contains an accurate denominator and limits on selection.



16 Collection, provenance and quality of evidence

Chain of origin

The proof includes not only the file, but also information about the origin, time of receipt, access conditions and conversions. The observatory keeps a log of the material's receipts, assigns a stable identifier, keeps the source separate from the working copy and records derivative operations. The hashing helps to identify the change in an electronic object, but it does not prove the authenticity of its contents or the legitimacy of obtaining it. Important claims require an independent source or verification directly in the official register.

Quality standards

A four-step verification scale is offered: a primary official document; a participant's direct documentary evidence with accompanying materials; an independent secondary source with verifiable links; unverified communication. The scale determines the degree of certainty in a particular statement and should not become an automatic rating of the veracity of the source as a whole. A single court order may prove the fact of a court decision, but not all of the parties' statements cited in it.

Reproduction

The research package should allow the second specialist to reproduce not only the calculation but also the logic of classification. For confidential data, an anonymised log of the signs and the controlled access procedure are kept, if permitted. The open report gives the source of each metric, sample, exclusion, version of code and date of verification. When data is lacking, we should simply abandon strong conclusions instead of masking the gap with expert confidence.

The chain of origin and preservation of evidence

The chain of evidence includes at least four independent questions: where the document came from; whether the file has changed; what the text confirms; what rights the researcher has to use it. The control amount resolves only the second issue and partially facilitates the verification of the preservation. For example, a signed image may be technically unchanged and may actually contain false text. Therefore, the quality of the source is evaluated in a meaningful way, not just by a software tool.

Digital materials must have a reproducible chain of origin: source and legality, time of receipt, original file name, control amount, access and results of any conversions. The verified amount confirms the immutability of the specific copy, but not the authenticity of the contents, authorization or completeness of the exported correspondence. For screenshots, the URL, date and context of the page are also stored, while for AI-generated responses, the exact version of the system, query settings, and restart restrictions are.

For court-related research, it is recommended to prepare a package with chronology, original numbering, translation descriptions and a disguise log. The scientific report specifies the limits of access: a published passage may be sufficient to verify one thesis, but not the whole case. If the second researcher is unable to reproduce a classification from available materials, the result is marked as not reproducible from the outside.

Control issues and minimum evidence

Verification Element	Evidence criterion
Source	Saved separately from working copies
Metadata	Time, channel, access and processing
External confirmation	Check independently where possible
Reproduction	Options for methods and criteria for exclusion

Key sources for comparison: [7], [8], [13], [14].

Practical case analysis 20.16

Recommendation: Monitor the source of each material fact

Hypothetical example: Five sites repeat one entry post.

Analysis: The number of publications does not mean five independent confirmations.

Practical steps: The researcher builds a chain of origin and highlights the primary material.

Outcome criterion: The record shows the number of independent bases, not reprints.

17 Information verification and the right of reply

Verification protocol

Each socially significant statement goes through three stages: verification of the factual basis, legal and technical context, and assessment of possible harm from publication. An editor cannot hide the fact that there is a rebuttal, even if the general hypothesis seems convincing. For major investigations, the "thesis - proof - alternative explanation - remaining uncertainty" matrix is useful. Such a matrix protects against glamorous but logically fragile causal explanations.

Response of the affected party

An organization for which a substantial negative conclusion is being prepared is given a reasonable opportunity to answer specific questions. The deadline depends on the complexity of the request and the urgency of the public risk; the internal guideline is established by procedure, not by law. The answer does not give the right to prohibit publication, but must be taken into account in good faith. Silence is not considered a confession of charges and is reflected in a neutral manner: "No reply received at the closing date of the material".

Refusal work

Once published, a new court act or technical protocol may modify the evaluation. The correction shall be accompanied by a date, a description of the modified statement and the preservation of the available history if it is lawful. A gross mistake that can affect a person's life must be corrected promptly. An institutional reputation is better protected by a transparent recognition of uncertainty than by trying to rewrite an event record.

Conflict resolution and the right to a reasoned reply

The right of reply is part of the methodological accuracy, not the assignment to the object of the study. A party may disclose a document that denies a link to the algorithm, indicates an error in the date or indicates that the decision was preliminary. The request must be specific enough to change

the result. Too general a question of "complain the charges" does not ensure quality. The opinion may be rejected if it is not supported by sources.

The internal decision to publish contains a table of all disputed facts, the party's response and the analytical assessment of the response. If there is no confirmation of a direct link between AI and the damage, the publication may discuss the process itself but not confirm the provenance of a technical defect. The applicant's immediate security sometimes limits the amount of information requested; this exception is documented and reviewed after the risk has been eliminated.

Control issues and minimum evidence

Verification Element	Evidence criterion
Thesis	Actual ground
Objection	Separate entry in review
Silence	Not count as confirmation
Adjustment	Version, date, reason, notice

Key sources for comparison: [13], [18], [19].

Practical case analysis 20.17

Recommendation: Provide a meaningful response

Hypothetical example: The report concludes the organization's actions without asking for its position.

Analysis: An unrecorded document can change the qualification.

Practical steps: The editor sends specific questions and permissible information, records the answer or the absence of a response.

Outcome criterion: The position of the addressee has been considered in substance and is reflected in the publication.

18 The risk register: structure and purpose

Two units of account

Sustainable work requires a separate system register and a risk register. The risk relates to the "system-action-vulnerability-possible damage-affected group" scenario and not just to the product. For example, the wrong transliteration of a family name in a sanctions-related match and the violation of access to financial services represent different links in the chain of causes. The register records the source, control measure, residual risk and responsible party. The risk may exist without a confirmed incident, but that doesn't make predictions a fact.

Evaluation

It is proposed to assess separately the severity of the potential harm, the likelihood of confidentiality, the spread of the system, the vulnerability of the affected persons and the reversible consequences. The addition of conditional points to a single hazard index is permitted only as internal sorting and necessarily reveals the subjectiveness of the weights. Important risks with low probability but irreversible harm must have their own order of escalation. The methodology does not claim normative legal classification.

Dynamics

After the model is updated, the law changes or a new complaint is received, the risk record is reviewed. The history includes a previous assessment, the reason for the change and the date of independent confirmation. Reducing the risk assessed may reflect a new protection measure, an improvement in knowledge, or a change in user group; these cases differ fundamentally. The mechanism of feedback from organizations and victims makes it possible to correct the picture without cancelling the independence of the Observatory.

Segregation of risk, severity and confidence

In the risk register, the degree of confidence cannot be mechanically mixed with gravity. A single violation that can lead to irreversible imprisonment may require prompt verification with limited data; frequent misclassification of advertisements may not have comparable severity. Therefore, it is recommended that separate fields of "hardness in implementation", "quality of information", "proliferation", "urgency" and "access to protection" be recorded. Any final category is used only for internal sorting of analytical work.

When reviewed, there is a difference between changing the original system, changing user exposure and changing the researcher's presentation. If the new test refutes the original hypothesis, the record shall be marked accordingly. Removing it without a trace is wrong: the history of such errors is necessary to assess the sustainability of the method and prevent repeated unfounded accusations.

Control issues and minimum evidence

Verification Element	Evidence criterion
Scenario	System plus event plus consequence
Scale	Hardness separate from confidence
Balance	After taking account of the confirmed measures
Escalation	Irreversible harm and vulnerable persons

Key sources for comparison: [8], [14], [17], [19].

Practical case analysis 20.18

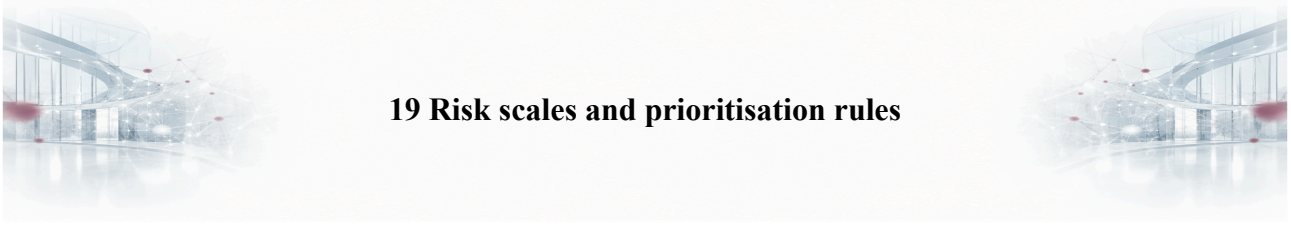
Recommendation: Link risk to measure and owner

Hypothetical example: The register records the risk of mistaken association of people without action.

Analysis: Recording does not by itself reduce the likelihood of harm.

Practical steps: The owner introduces verification of different identifiers and tests it on conditional surnames.

Outcome criterion: The risk has a responsible measure and a confirmation of her verification.



19 Risk scales and prioritisation rules

Transparent levels

The internal scale is recommended to be limited to four levels of action: regular observation, enhanced verification, prioritized investigation and urgent escalation. The level is determined not only by the quantity, but also by the nature of possible harm: deprivation of liberty, unlawful interference in family life, vital medical care and the impossibility of dispute are capable of demanding attention even in a single case. The "low" and "high" indicators should not be turned into an assessment of the applicant's identity or of the credibility of the organization.

Decision-making matrix

For each record, the combination of severity, the convincingness of the source, the number of people affected, the urgency and the ability to prevent damage is recorded. If necessary, the defence action can be initiated before the full investigation is completed, but the public declaration must reflect the degree of confirmation separately. For example, a request to preserve electronic logs may be reasonable in the event of a well-founded threat of their destruction, but that does not give you the right to assert that guilt is proven.

Reconsideration of priorities

Every quarter, the editorial board evaluates whether or not the priority system has become unfairly favorable to the most media-oriented countries and commercially known developers. It is checked separately which groups and languages are not adequately represented. Criteria are reviewed before new cases are selected and published in a methodical log; weighting cannot be changed to obtain the desired result. Historical aggregates should be counted for significant changes.

Urgent priority and review mechanism

A system of protection against red flags in the press is important for prioritization. Loud publication may attract attention, but high priority should be determined by potential harm, prevention, and quality of information. A two-step evaluation is offered: the initial threat triad, followed by independent verification of evidence. If immediate action is necessary, it may be limited to a confidential request to keep logs without a public declaration of guilt.

The proportion of projects by region and group of applicants should be analysed over the long term. If almost all research is about the English-speaking markets, you can't call the program a global monitoring program without reservation. This shift is not achieved by artificially raising the scores of individual countries, but by planning language resources and new channels of sources.

Control issues and minimum evidence

Verification Element	Evidence criterion
Low	Periodic update
Advanced	Clarification of documents and context
High	Independent technical verification
Urgent	Protocol for preservation of evidence and escalation

Key sources for comparison: [8], [9], [14].

Practical case analysis 20.19

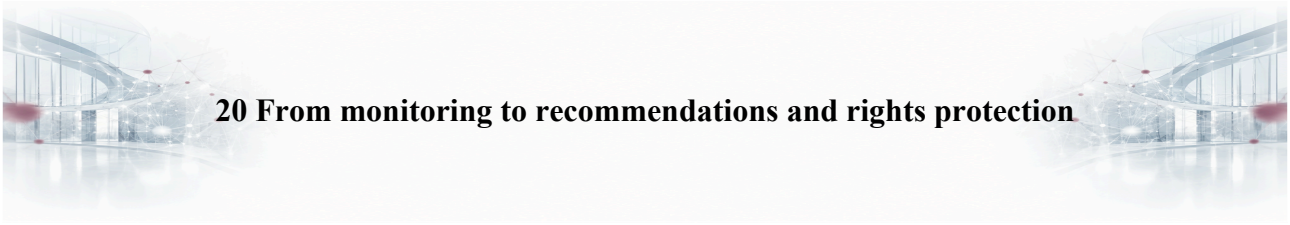
Recommendation: To justify priority on consequences

Hypothetical example: A popular media story replaces the less visible continuing restriction of rights.

Analysis: The frequency of references is not equal to gravity and urgency.

Practical steps: The coordinator evaluates the confirmation, gravity, reversibility and continuing character.

Outcome criterion: The order of consideration has been explained on a single scale.



20 From monitoring to recommendations and rights protection

Recommendation is not a requirement

The Observatory may propose to the developer to disclose the documentation, to the operator to review the procedure, to the regulator to review the matter within its competence, and to the research community to develop tests. However, this cannot be formalized as a binding order. The recommendations are divided into legal, technical, organizational and research. Each person shall be identified as the addressee, the subject confirming the observation, the expected result and the condition of subsequent verification.

Taking into account responses

The organization's response is not identical to the implementation of the recommendation. Four statuses are therefore recorded separately: accepted for consideration, declared for action, confirmed by document, and confirmed by independent data. If an organization reports a correction, the Observatory does not draw conclusions without verification applicable to the nature of the system. Otherwise, public monitoring will become a catalogue of the promises of market participants.

Rights of the person concerned

For people affected by automation, it's best to specify the official means of protection available: requesting information, contacting a data controller, administrative complaint, litigation, lawyer's office. The report may contain general navigation, but it does not promise legal victory and does not replace professional advice in a particular jurisdiction. The possibility of reviewing a decision should be assessed on the basis of actual availability, timing and cost rather than the formal availability of an electronic form.

Follow-up to the targeted recommendation

It is appropriate to divide the recommendations into actions that the addressee can carry out independently and measures that require a competent authority. For example, correcting the wrong questionnaire may be at the disposal of the system operator, and abolishing a legally significant restriction may depend on the bank, court or regulator. The Observatoire must indicate the chain of competence and not promise the applicant the legal effect of its own publication. In sensitive cases, general recommendations on how to complain do not replace personal work as a lawyer.

It is recommended to maintain intermediate statuses and documents for evaluation of performance: receipt of a recommendation, consent to review, approved action plan, actual implementation and outcome of independent verification. The lack of confirmation means "the effect has not been established", not "the recommendation has failed" or "the system has not changed". Honest typology prevents the use of monitoring as a list of formal promises.

Control issues and minimum evidence

Verification Element	Evidence criterion
Public recommendation	Do not create the appearance of a binding decision
Acclaimed correction	Separate from independent confirmation
Practical protection	Registered official appeal route

Verification Element	Evidence criterion
Further monitoring	Re-evaluation date

Key sources for comparison: [3], [11], [13], [18].

Practical case analysis 20.20

Recommendation: Make a recommendation with enforceable effect

Hypothetical example: The report recommends greater transparency in the refusal decision.

Analysis: The addressee does not understand which record or procedure to change.

Practical steps: The author suggests showing the original data and the reason for the refusal, names the owner and the method of verification.

Outcome criterion: The recommendation contains the action, the addressee and the performance criterion.



21 Personal data protection and information security

Legal evaluation of processing

When processing data about specific individuals, the Observatory should determine the role of the controller or processor, the purpose and basis of the processing, the category of data, recipients and storage periods. In the EU, GDPR requires checks on the principles of minimisation, accuracy, purpose limits and security. Special categories of data, information on offences and data of minors may require additional conditions and special guarantees. Information in open sources does not automatically become free of legal protection; public availability is just one of the circumstances of analysis.

For each set of personal data, the role of the organisation (controller or processor) must be defined before receipt, specific purposes, the basis under Article 6 GDPR, storage periods and addressees. A socially useful research goal does not per se eliminate the requirements of the GDPR. If the information relates to health, political beliefs, biometric identification or other special categories, the applicable condition of Article 9 is also required. The exclusion for scientific research cannot be declared universal: its conditions and national guarantees must be established for a specific treatment. [13]

Article 10 of GDPR constitutes a particular legal barrier. Information on criminal convictions, offences and related security measures that may be found in international search materials cannot be freely transferred to an open register on the basis of a common interest for monitoring. Their processing requires the control of an official body or authorisation provided for by EU law or a Member State and appropriate guarantees. Until the appropriate regime is confirmed, Observatory should be limited to anonymized structural indicators and public decisions to verify the readability of the document; even public access does not eliminate GDPR requirements. [13]

Before launching the channel of individual appeals, the 14th article of GDPR (information when receiving data not from a subject), the 28th article (contract with processors), the 32nd article (security), the 35th article (DPIA with probable high risk), the 33-34th article (reaction to security violations) and the 44-49th article (transmission of data outside the EEA) are evaluated. The applicability of each obligation is determined by the actual processing; mentioning the provision here does not imply that the obligation necessarily arises in each case. Pseudonymisation reduces the risk, but does not automatically make identifiable data anonymous. [13]

The issue of admission	Mandatory pre-treatment check
Regular personal data	Purpose, role, basis of article 6, minimization and duration of storage
Special categories	Applicable deletion of article 9 and national safeguards
Data on offences	Separate verification of admissibility under article 10 and applicable national law
High risk to rights	Assessment of the need for DPIA and measures to mitigate residual risk
Transfer outside the EEA	Legal mechanism for transfer, recipients and documented guarantees

Division of media

It is recommended to keep the contact details of applicants, the research facility and the public aggregates separately. Access is given by role, verified through logs, and the export of materials is subject to approval. Protected channels are used to transmit files; attaching a full file to a regular letter should not be standard practice. When an external expert is engaged, the permissible operations, the prohibition on re-use and the procedure for deleting data after conclusion are determined in advance.

Assessment of leakage

In the event of a security breach, it is important not only to comply with the applicable notification requirements but also not to increase the harm caused by rushed publication. The response plan shall include access isolation, preservation of evidence, evaluation of disclosed data types, possible notification of the competent authority and victims in the cases provided for by law. The Observatory's own public reporting should not conceal its own significant mistakes. Data minimization is verified annually, not just the complexity of technical protection.

Legal grounds, special data and security

The structure of the legal processing must be in accordance with the specifics of the Observer. A researcher can legally analyze a published court case, but combining different databases significantly increases the risk of identifying a person and reusing the data in a different context. Therefore, each new array binding requires a separate assessment of purpose and necessity. You can't keep a customer's full passport in an algorithmic risk record just for the convenience of subsequent search.

It is recommended to conduct a re-identification test before releasing open data: what information is available to a third party and can it be retrieved through public search? At high risk, some attributes are aggregated or not published. An internal data disclosure log helps to recover which versions were transmitted and to whom if a security breach is detected later.

Control issues and minimum evidence

Verification Element	Evidence criterion
Basis	Make excuses before personal information is collected
Access	By role, with registration of extradition
Storage	Limit target and duration
Leakage	Separate response and evaluation of legal notices

Key sources for comparison: [13], [18].

Practical case analysis 20.21

Recommendation: Check indirect identification

Hypothetical example: The name was removed from the record, but a rare profession and an exact date was retained.

Analysis: The combination of the indicators makes it possible to identify the complainant.

Practical steps: The person in charge evaluates the need for details, summarizes them and limits access to the original.

Outcome criterion: The open version does not reveal any additional identifying combinations.

22 Vulnerable groups and participation of affected persons

Invisible harm

The effects of automation are not always visible in aggregate statistics. For migrants, people with limited digital access, minors or citizens in difficult financial situations, a failed decision can have disproportionate effects. The observatory should not only assess the likelihood of an error, but also the ability of the affected person to learn about it, get an explanation and correct the result. In the absence of these procedures, even a relatively rare algorithmic failure can create deep inequality.

Meaningful Participation

The participation of victims is not limited to the completion of the questionnaire at the end of the project. When developing the methodology, it is necessary to take into account language comprehension, safe handling, distrust of institutions and the risk of re-injury. An expert group may involve community representatives to assess the impact, but does not turn their stories into evidence of a technical mechanism without further verification. The remuneration for expert participation should be transparent and independent of the conclusions.

Rules of publication

Systematic monitoring of vulnerable groups requires strict anonymisation or another legitimate data protection model. Removing a name is not enough if the combination of dates, locations, and rare circumstances allows us to identify a person. In the event of a conflict between the information and the applicant's security, the priority is to prevent unreasonable disclosure. The case can be presented as an aggregated scenario with clearly marked changes to details and evidence preserved in a closed archive.

Participation of affected persons without secondary harm

Qualitative participation of affected people requires self-planning. Translating the complaint form into another language does not remove restrictions on access for a person without documents, communication or free time. The research team should study the real barriers to entry and not confuse a small number of complaints with the absence of harm. For vulnerable groups, confidential intermediaries may be required who do not have the right to change an applicant's evidence.

When developing a metric, account should be taken of the gravity of the single injury. If an automated error deprives a person of a vital service, the simple average "model accuracy of 99%" does not describe the adequacy of the protection. For this case, timely repair and the presence of a backup human mechanism are important. In publishing the story, the Observatory must protect not only the name, but also a rare combination of social circumstances.

Control issues and minimum evidence

Verification Element	Evidence criterion
Channel of application	Available language and format
Assessment of damage	Account for reversibility and procedural position
Co-Design	Participation without pressure to withdraw
Publication	Re-identification test

Key sources for comparison: [3], [4], [13], [18].

Practical case analysis 20.22

Recommendation: Take into account the accessibility of the participation procedure

Hypothetical example: The form of the complaint is available in only one language and requires a complex file.

Analysis: Monitoring systematically misses a number of affected persons.

Practical steps: The team introduces understandable alternative channels and takes account of the change in coverage in the indicators.

Outcome criterion: Accessibility improved, period comparison adjusted.



23 International data exchange and cross-border cases

Consent to the record's overall structure does not mean permission to share the source files. International colleagues are first sent methodological definitions and unidentified indicators; the transmission of original individual materials requires a separate analysis of the role of the recipient, legal basis and requirements for transboundary transmission. Access to an international organization, a scientific expert or a consultant cannot be considered automatically exempt from data protection rules. [13]

Transfer and localization

Monitoring a system operating in several countries may include personal data, bank secrecy and criminal proceedings. For each cross-border transmission, the participants, the purpose, legality of the original receipt and the basis for further processing must be established. In the European context, the provisions of Chapter V of the GDPR may apply; other jurisdictions have their own requirements. The general interest in global comparative analytics does not override the limitations of the jurisdiction of origin of materials.

Minimum disclosure rule

The international working group should not have the same access to the entire archive by default. It is advisable to organize access according to a distributed model: the local keeper checks the primary materials, and the international participants receive a minimum set of verified attributes suitable for comparison. If the objective can be achieved by statistical indicators without referring to a file, this option is chosen. This is particularly important for reports of politically sensitive persecution and international search.

Comparability of procedures

A foreign document can confirm the fact of the detention, but not the legitimacy of the search; a court refusal to extradite is a decision by a specific procedure, but not universal proof of the absence of international police records. The Observatory is obliged to keep a distinction between the status of the procedure and the assessment of the legal regime. In order to compare the briefcases, the "document", "body", "date", "procedural effect", "jurisdiction" and "proven limitation" fields are introduced.

Cross-border exchange: minimization and conditions of transfer

For cross-border work, it is useful to use the "local custodian" model: primary personal data remain with the legal holder and the international team receives a minimum impersonal set of features. This reduces the risk of inconsistent national rules and excessive personal data transmission. If the finding cannot be verified without direct access to the document, a competent independent on-site verifier with limited powers shall be appointed. The public opinion describes the level of reviewer's access without revealing a secret.

Of particular complexity is the discrepancy between the procedural status. The termination of one proceedings by a court does not always mean the termination of another in another country. The automatic recording in the database and the judgement are also not equivalent. When building an international profile of a case, it is necessary to keep separate lines of competence without turning the timeline into a simplified "global status".

Control issues and minimum evidence

Verification Element	Evidence criterion
Collection of material	Legitimate purpose and origin
Access by colleagues	Minimum justified set
Public record	Only verified neutral attributes
Cross-border act	Do not generalize the effect on all jurisdictions

Key sources for comparison: [11], [13], [18].

Practical case analysis 20.23

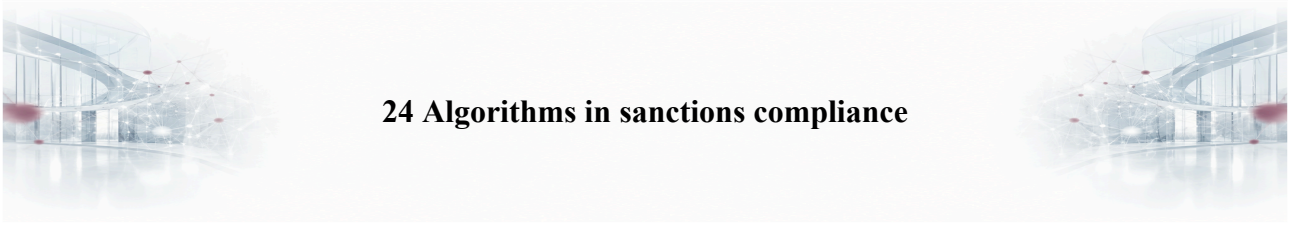
Recommendation: Minimumize international exchange

Hypothetical example: The partner's requesting a full archive to check one episode.

Analysis: The volume of the request exceeds the research task.

Practical steps: The coordinator identifies the relevant material, checks the basis and protected route of transmission.

Outcome criterion: Transfer documented by composition, purpose and recipient.



24 Algorithms in sanctions compliance

Sectoral characteristics

Sanction screening typically combines rules for names matching, transliteration, ownership traits, and relationship evaluation. An error can lead to a blocking of the operation, denial of service, and prolonged collection of documents. However, the name itself does not prove the error or illegality of the decision: the current sanctions regimes, identification details and internal obligations of the organization are important. The Observatory monitors not only the technical accuracy, but also the presence of a clear review process, response times and the proportion of cases where refusal is justified by relevant legal grounds.

Metrics without false objectivity

The false positive rate only makes sense when there is a set procedure for determining true status and a known set of negative examples. If there is no independent marking, only the metric of "the percentage of requests found to be incorrect by the review" is allowed - it measures the review process rather than the true accuracy of the screening. The systematic removal of complex names

from the test sample results in a displaced estimate. The protocol must take into account transliteration, order of names, geographical variants and list updates. [14]

Human rights component

Particular attention is paid to the availability of the mechanism for correcting identification and preserving written history of interaction. The researcher has no right to promise to withdraw sanctions or access the account based on his own assessment; these decisions are with competent subjects. A public report may substantiate recommendations on notification and appeal procedures without disclosing personal bank records or issuing a draft control scheme for a mandatory requirement for each bank.

Verification of false matches without presumption of error

The verification of sanction screening should take into account that the price of a false positive signal and the price of a missed actual match differ, but none of these eliminate the legal obligation of the organization. For a research report, it is particularly important to separate the technical metrics of the model from the legal evaluation of the bank's decision. For example, repeated matching of names may require adjusting the human verification procedure, but it is not evidence of an unreasonable refusal without document analysis.

For a conditional test sample, a set of names is formed with controlled transliteration variants, pseudonyms and identifier differences. The test is performed without real client information if a synthetic model is possible. The report limits the area of inference to the test set and shows openly cases where ground truth was not established. The transparency proposals are formulated in such a way as not to expose the sensitive thresholds of sanctions control.

Control issues and minimum evidence

Verification Element	Evidence criterion
Convergence of name	Not equal to the identified identity
System error	We need an independent spring trout.
Metr of revision	Separate from technical FPR
Client protection	Documentation, complaint, human verification

Key sources for comparison: [6], [11], [13], [14].

Practical case analysis 20.24

Recommendation: Check a sanctions match separately from a risk

Hypothetical example: The record describes the refusal of a single name as a designated sanctions person.

Analysis: The machine flag and identity aren't identical.

Practical steps: The expert compares available identifiers and keeps the disputed match as a matter of verification.

Outcome criterion: The publication does not attribute unconfirmed sanctions status to a person.



25 International wanted-person alerts, identification and migration

High-risk conditions

Automating identity matching, document analysis and query prioritization can have particularly serious consequences if the result is used in border control or international search. The Observatory does not have access to INTERPOL's closed databases by default and should not claim to be able to confirm individual search status on public materials. The research focuses on verifiable institutional guarantees: quality of comparison, preservation of decision causes, human verification, correction mechanisms and compliance with procedural rights.

Decision chain

An algorithmic clue, a police record, a national warrant, arrest and extradition are all different stages. It would be a mistake to attribute the entire outcome to the work of the model or to confuse international data exchange with the judicial competence of the state. For legal analysis, you need to document which authority made each decision, which digital tool was used and what could change the outcome. In the absence of technical logs, the conclusions about the mechanism of the system remain preliminary, even if the consequences for humans are confirmed.

Public disclosure

Asylum, persecution and international police information cases often involve a threat to the applicant's safety. Publishing the name, migration route and biometric characteristics can exacerbate the risk. Therefore, industry reports should aggregate process observations or use conditional composite scenarios. The recommendations refer to the principles of accuracy and access to an effective remedy, rather than to the public definition of "who is wanted".

Irregularity of finding the role of AI without evidence

An international search warrant must strictly distinguish between a technical match, information entered into the police system, a national enforcement measure, and an extradition court decision. These processes may relate to different legal systems and do not form a single automatic legal status. The presence of digital processing does not prove that it was AI system that created the basis for the retention. For a causal conclusion, documents on the instrument used and the subsequent decision of the competent person are required.

For a public case, it is sufficient to consider the structure of guarantees without identification of the person: the existence of a correction procedure, the independence of the complaint, the preservation of the basis of the decision and the possibility of legal support. If only the lawyer's message and court order are available, the technical part is marked as unknown. It doesn't make the damage done insignificant, but it protects the analysis from attributing an unproven technological cause.

Control issues and minimum evidence

Verification Element	Evidence criterion
Type of solution	Segregate signal, measure and act
Source	Only legally available document
Causation	Do not assume the use of IE without proof
Personal risks	By default, an impersonal analysis

Key sources for comparison: [3], [10], [11], [13], [18].

Practical case analysis 20.25

Recommendation: Segregate official record and automatic report

Hypothetical example: The algorithm combined publications about different people into one search story.

Analysis: The summary does not confirm the official status and may confuse persons.

Practical steps: The lawyer links each claim to the original material and excludes unfounded associations.

Outcome criterion: The form contains a separate reference to the source, the person and the proven status.

26 Financial monitoring and AML/KYC

Decision-making

With AML/KYC, automated tools detect anomalies, form warnings, and help analysts study financial transactions. The existence of a warning does not in itself prove a crime and does not mean a final refusal to the client. The Observatory should investigate the transparency of human escalation rules, the availability of corrective misstatements, and the quality of internal risk and sanction mechanisms. The confidentiality of anti-money laundering can limit the information available to the client and the outside researcher.

Composition of metrics

For analysis, several non-derivable indicators are useful: the proportion of warnings escalated by the analyst; the proportion of closed without additional action; the duration of manual review; change in load; the proportion of tested cases from different groups of clients. These indicators should not be interpreted as "effective crime detection" unless there are proven grounds. You can't publish the number of messages sent to financial intelligence as a crime assessment without a systematic reservation about the nature of those messages.

Recommendations

Quality identification processes, rules for periodic reviewing of models, division of responsibilities and the preservation of sufficient information for legitimate verification are priorities. Public monitoring can analyze rights protection standards and aggregated results without revealing criteria to circumvent financial control. Technical tests should avoid the dissemination of real personal data and bank instructions that could be used to harm.

Boundaries of the AML-soring conclusion and denial of service

In AML/KYC, the three-step model is useful: the algorithm forms a warning, the analyst examines it, the authorized organization takes a legally significant action. Each step may have its sources of error and should not be combined into an undifferentiated "automatic blockage". Monitoring verifies the quality of data, the criteria for further verification, documentation of human involvement and the effectiveness of the correcting of misidentification. Confidential mechanisms for detecting suspicious operations require caution when publicly disclosed.

When looking at aggregate indicators, it is important to establish the working units: the proportion of warnings, the proportion of closed calls, the time frame for consideration and the number of decisions actually contested. Without a record of proven wrongdoing, the number of warnings "captured by crimes" cannot be described. For an independent comparison of banks, the differences in product lines, customer base and volume of operations should be noted, otherwise statistics would be a punishment for broader scope of control.

Control issues and minimum evidence

Verification Element	Evidence criterion
An algorithmic signal	Unreconciled financial misconduct
Escalation	Share of known warnings
Procedural protection	Time frame for review and documented response

Verification Element	Evidence criterion
Publication	Do not disclose sensitive control rules

Key sources for comparison: [6], [11], [13], [14].

Practical case analysis 20.26

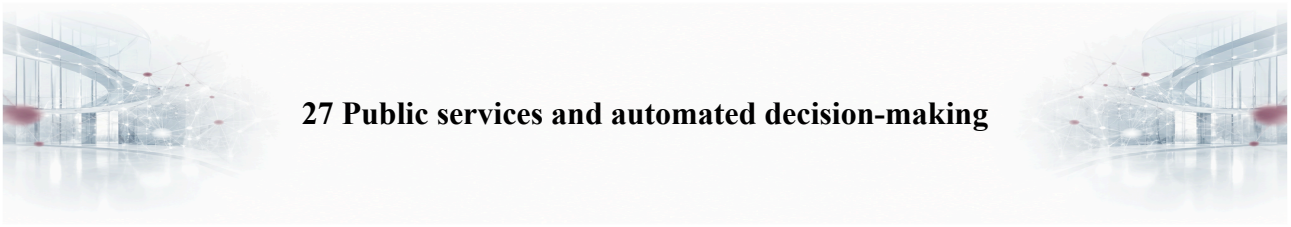
Recommendation: Do not equate AML signal with crime

Hypothetical example: The claimant was refused after an automatic risk assessment.

Analysis: The fact of refusal does not establish a violation of the law by the client.

Practical steps: The analyst describes the confirmed action and asks for acceptable grounds for review.

Outcome criterion: The text retained the distinction between risk, decision and legal characterisations.



27 Public services and automated decision-making

Public function

When allocating benefits, queuing or assessing needs, the use of AI can affect guaranteed rights. The study should not be limited to the accuracy of the classification. Information on the normative basis of the decision, the role of the official, the clarity of the notification, the actual possibility of appeal and the consequences of delayed correction are required. Contextual assessment of human rights helps identify processes where even a small mistake becomes a serious obstacle for a vulnerable person.

Openness without illusion

The public AI project registry is useful, but the record itself does not confirm that the project works as claimed. The Observatory compares data on procurement, operation, independent inspections and regulatory decisions. For international comparison, the difference between a pilot program and a large-scale implementation is taken into account. Projects completed after problems are detected are separately recorded so that the indicators do not create a false picture of the endless growth of automation.

Alternative to automation

A sound assessment may conclude that the task does not require an AI at all. In accordance with the risk management approach, it is advisable to compare an algorithmic solution with the existing procedure in terms of quality, cost, availability of explanation and legal consequences. The public interest requires not maximizing the number of implementations, but choosing a procedure that ensures rights and achieves the goal. The evidence base for such a comparison should take into account institutional resource constraints.

Monitoring of reasons for decision and available appeal

The public service map begins with the issue of the purpose of automation. If an organization optimizes the speed of decision-making, but doesn't measure the proportion of wrongful rejections and the availability of correction, efficiency is underestimated. For the Observatory, a practical unit is a human's path from filing an application to a decision and possible revision. In this way, the points are recorded when the algorithmic prediction becomes the actual outcome, who can change the decision and which document remains with the applicant.

Society benefits from the available criticism of algorithmic solutions, but the code itself does not provide a clear protection procedure. Therefore, monitoring should assess the link "explanation - possible objection - competent revision - actual correction". The comparison of institutions takes into account the social challenge and the availability of non-digital alternatives for individuals who cannot use the online service.

Control issues and minimum evidence

Verification Element	Evidence criterion
Legal framework	Competence and criteria for an administrative decision
Explanation	A humanly understandable foundation
Appeal	Feasibility and timing
Alternative	Comparison with manual procedure

Key sources for comparison: [3], [9], [11], [14].

Practical case analysis 20.27

Recommendation: Check the availability of public review

Hypothetical example: There is no clear way to correct the data in the notice of refusal of service.

Analysis: A technical description of the model does not solve the applicant's problem.

Practical steps: The team shall establish the operator, the original data and the envisaged challenge route.

Outcome criterion: The recommendation is addressed to the owner of the procedure and is checked by a training application.



28 AI procurement and public-sector contracts

Place of treaties

In international monitoring, procurement is a key evidence of implementation. The contract may disclose obligations for logging, model updating, security and data use. However, a purchase record does not mean a technical release. The Observatory distinguishes between the announcement of a competition, contracting, commissioning, modification, acceptance and termination. If the supplier maintains control over the model, the conditions for access by the public customer to documents and the consequences of incidents are particularly important.

Verification rights

It is proposed to recommend to purchasers the conditions for providing the verified documentation to an independent appraiser, notification of changes, storage of logs, separate responsibility and suspension of use. These provisions remain design guidelines, not universal requirements. In commercially sensitive cases, a closed check with a public description of the testing area, significant shortcomings and access restrictions is allowed.

Comparison of costs and benefits

The comparison of contracts should take into account the purchase price, technical support, data migration, staff training and the costs of correcting erroneous decisions. The public price of subscriptions without additional obligations does not describe the cost of ownership. If the contract

is classified, it is forbidden to assign terms to it based on a typical offer from the company. The conclusion is marked as limited by the open data, and the recommendations are directed to the transparency of the control conditions.

Auditability of procurement and contractual obligations

Procurement monitoring can track the duration of a project as a sequence of individual status rather than as a single indicator, "AI is implemented". There are competition marks, contracting marks, pilot, recruitment, work exploitation and termination marks. The analyst checks whether the contract contains the ability to independently obtain logs, documentation and information on updates in the event of complaints. Even if there are no such conditions, you can't file a back-up claim until the applicable legal standard is established.

For public analysis, a minimum set of contractual recommendations is useful: a stable system identifier, the obligation to notify a significant change, incident response measures, distribution of responsibility for data quality, and the mechanism for independent complaint handling. These requirements should be referred to as model terms and conditions of procurement. Comparing the cost of ownership involves expenses for maintenance, training, data migration and error-correcting procedures, not just subscription.

Control issues and minimum evidence

Verification Element	Evidence criterion
Procurement	A confirmed stage, not just an announcement.
Update	Contractual obligations of notification
Verification	Limited access by an independent evaluator
Value	Full life cycle

Key sources for comparison: [3], [11], [14], [16].

Practical case analysis 20.28

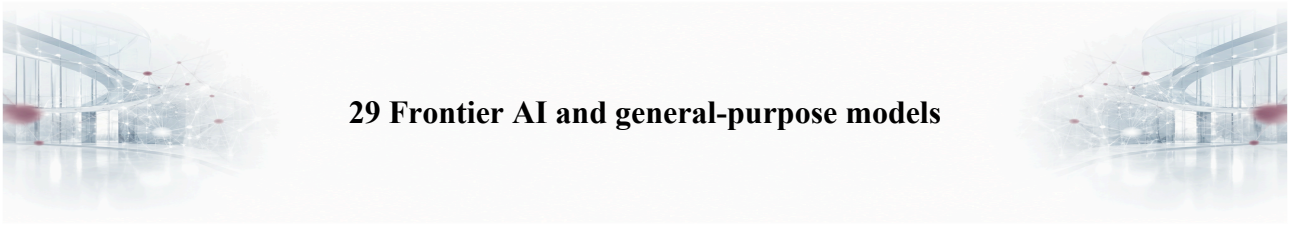
Recommendation: Include audited terms and conditions in the procurement

Hypothetical example: The public customer demands an E.I. without acceptance criteria.

Analysis: The general wish does not permit verification of performance.

Practical steps: The expert offers versions, a log, independent access and a dispute resolution test as the subject of admission.

Outcome criterion: The contract control relates to specific evidence.



29 Frontier AI and general-purpose models

Monitoring object

A general-purpose model can be a basic component of thousands of applications, so its evaluation requires a separate design. The Observatory describes the model, version, availability of weights, set of known tests, access boundaries, and application area. The designer's results are not considered to be independent proof of safety. The International AI Safety Report 2026 highlights the gap between laboratory test results and the actual behavior of systems, which is particularly important for agency functions and new capabilities.

Systemic risks

Risks include large-scale errors, abuse, infrastructure security and impact on public systems. They're not just an individual event. Regular changes in test results, the independence of the evaluator, documentation of protection measures and the presence of a response procedure are useful for monitoring. But a single closed test doesn't prove the universal safety of the model. The report separately shows the unprotested areas, the potential leakage of test data, and the impact of the selection of queries.

Legal status

The EU AI Act sets out specific obligations for GPAI model suppliers and stricter measures for systemic risk models. They cannot be automatically transferred to all users of any chat-bot. The Observatory analyzes the responsibilities by the current norm, the role of the subject and the date of action, and marks its own recommendations as voluntary. Conclusions about the technical potential of the model should not be made by a decision on a violation of law without authority and full factual basis.

GPAI risk assessment without change of regulator status

Frontier AI monitoring requires a clear distinction between assessing capabilities and inferring harm. The test results can confirm that the system has solved a certain set of problems, but not automatically show how it behaves in an unknown environment or how the attacker might use it. The report should record the configuration, version, available tools, access restrictions, and test protocol. It is particularly dangerous to refer to the "expert benchmarking" without being able to understand its representativeness.

For general purpose models, a register of updates to independent test results, individual reports on real incidents and developers' statements about protective measures is offered. These layers are not reduced to an overall safety assessment. If specific technical details are closed, the Observatory publishes a verification framework rather than simulating a full audit. The question of GPAI's legal responsibilities is analyzed separately from the scientific question of the model's measured capability.

Control issues and minimum evidence

Verification Element	Evidence criterion
Model	Version, provider and access
Test	Independence and protocol
Risk	Opportunities and consequences separately
Law	GPAI category and applicable duty

Key sources for comparison: [11], [12], [15], [19].

Practical case analysis 20.29

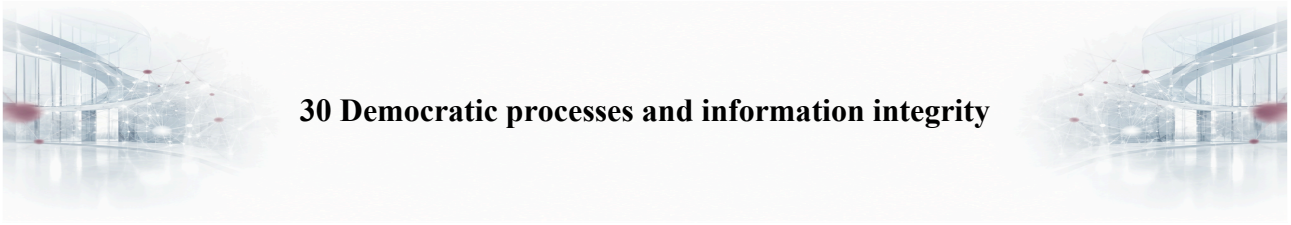
Recommendation: Do not move total rating to open product

Hypothetical example: The model's high rating is used to justify the assistant's stand-alone actions.

Analysis: The ranking does not cover the rights of instruments and the context of the operator.

Practical steps: The evaluator requests a configuration passport and performance tests.

Outcome criterion: The conclusion relates to the actually tested regime of application.



30 Democratic processes and information integrity

Public risks

Generative models can be used to create compelling false material, imitate voice and scale manipulative messages. When studying such phenomena, it is necessary to distinguish automated generation from deliberate deception and documented campaign from assumptions about the organizer. The risk of false accusations against political and public figures is particularly high. The Observatory must adhere to neutral procedures published prior to the investigation and exclude editorial influence by party interests.

Indicators

The calculation of synthetic materials depends on the quality of the identification. An AI-generated text detector is not a permanent proof of origin, especially when translating and editing. Therefore, confidence levels, manual verification methods and the proportion of disputed objects are published. Coverage indicators should take into account platforms and languages; public influence cannot be automatically inferred from the number of publications found. The causal relationship to the behavior of the audience requires a separate study.

Rights and balance

Anti-manipulation should not become a mechanism to suppress legitimate expression of opinion. Recommendations are made with respect to freedom of expression, privacy and reasonable limitations on evidence. If an error is detected in the own assessment of the origin of the content, the correction is published with as much notice as the original statement. For sensitive election periods, there is increased ethical control, not accelerated decline in the standard of verification.

Derogation of proven intervention from hypothesis

Information integrity monitoring should take into account the fundamental problem of attribution. Even a high-quality synthetic image or text detector doesn't prove the organization of the campaign or the identity of the author. The conclusion on coordination requires additional data, which are often closed. The Observatory is obliged to report the technical limitations of detection without publishing unfounded personal accusations. The reporting of the alleged manipulation can increase the spread of false information.

The publication procedure includes an assessment of public interest and proportionality. Do you need a full re-examination of the image or a sufficiently impersonal description of the techniques? Would it harm the participant to publish again? For serious conclusions, an external audit by a specialist in digital forensics and human rights is helpful. The objection of the affected party is considered according to the same standards, regardless of its political status.

Control issues and minimum evidence

Verification Element	Evidence criterion
Contact	Separate Origin and Truth
Detector	Assessment of error and contentious cases
Impact	Do not remove causality from view
Neutrality	Dismissal of politically interested reviewers

Key sources for comparison: [3], [6], [9], [19].

Practical case analysis 20.30

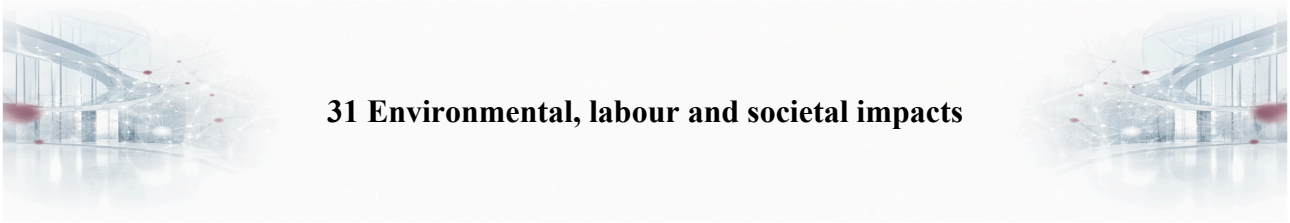
Recommendation: Segregate the origin of content from its influence

Hypothetical example: The material is believed to have been created by I.I. and is widely distributed.

Analysis: Origin, reliability and impact require different evidence.

Practical steps: The researcher checks each question separately and maintains the uncertainty of attribution.

Outcome criterion: The publication does not replace evidence of the influence of the authorship assumption.



31 Environmental, labour and societal impacts

System scale

Independent observation should include not only the most dramatic incidents, but also distributed effects: computing infrastructure burden, data marking working conditions, availability of technology and resource inequality. Such indicators are rarely linked to the same event. To evaluate the impact, you need comparable techniques, geographical context, energy types and system boundaries. The publication of the only number of carbon footprints of IE without explaining the composition of the calculation results in poor comparisons.

Work processes

Automatic assessment systems can influence working conditions, promotion and disciplinary decisions. The Observatory focuses on informing, human review, representativeness of assessment criteria and employee complaints. It's important to distinguish algorithm testing from evidence of discriminatory resolution in a particular organization. Cases are studied with respect to the protection of personal data and the possibility of safe communication of employees' representatives.

Comparison method

When evaluating environmental and social impacts, it is recommended to publish the limits of the account and the sensitivity of the conclusions to the assumptions. The evaluation of the manufacturer's available reports is marked as self-reported, and independent verification is separated. A public dashboard should not confuse technical energy efficiency with the distribution of social benefits. An international comparison can be fairly concluded that reliable ranking is not possible due to the lack of open data.

Estimation of external effects and calculation boundaries

In environmental assessment, the boundaries of the system are more important than an effective final number. It should be clear whether the learning model, service performance, data center infrastructure or lifecycle of the equipment is being measured. Comparison between organizations is possible only with the same coverage and units. The developer's data is marked as self-declared, and the external researcher's assessment - indicating how to verify. The lack of reliable data in one country does not prove less environmental impact.

In labour monitoring, it is important to separate the role of AI in automated ranking and employer independent decisions. Group inequality of results requires statistical verification and legal context. The summary of a particular complaint should not reveal the identity of the employee if it is a small group. The Observatory model provides for a separate assessment of the social availability of technology so that benefits and risks are considered symmetrically.

Control issues and minimum evidence

Verification Element	Evidence criterion
Energy	System boundaries and period
Work	Procedures for appealing decisions
Company statement	Separate from independent verification
Imminent	Directly reflect missing data

Key sources for comparison: [3], [6], [17], [19].

Practical case analysis 20.31

Recommendation: Comparing public costs by a certain base

Hypothetical example: Two services are compared by energy costs without the same load unit.

Analysis: Numbers can reflect different volumes and types of work.

Practical steps: The analyst sets a comparable task, measurement boundaries and sources; when the data is lacking, he rejects the rating.

Outcome criterion: Comparison is reproducible or restrictions clearly prevent ranking.



32 Engagement with reporting persons and civil society

Structural participation

Independence doesn't mean isolating experts from people who are affected by technology. The Observatory can set up advisory groups of researchers, representatives of public associations, professional communities and individuals with experience in algorithmic solutions. Participants should know which issues are addressed in consultation and which are within the competence of the scientific editorial board. You can't turn civil society's participation into a decorative endorsement of an already accepted conclusion.

Prevention of asymmetrics

Large technology organizations have more resources to prepare responses and expertise than individual applicants. In order to avoid this distortion, clear forms of submission, translation of key materials and an available review of the methodology are provided. At the same time, the right of the audited organisation to submit a documented objection is respected. The procedure is not required to consider all opinions to be equally valid, but it must explain why the evidence strength of the materials differs.

Representation

Membership in an expert council should not be considered a right to speak for the whole social group. Different countries, disciplines and types of experience are assessed in the selection of participants. For long-term studies, it is useful to describe not only the presence of "stakeholders",

but also cases where their comments have led to a change in the method. This indicator is less effective, but shows the real quality of feedback.

Consultation procedure and protection of interests of participants

Public participation must be arranged in such a way that opinions can really change the method. At an early stage, participants are asked to discuss the definition of harm, vulnerability categories, communication channels and notification language rather than endorsing final conclusions. To do this, you need a notebook with answers: accepted, partially accepted, rejected for a reason. In the public report, it is enough to summarize how participation changed the protocol without revealing the identity of the applicants.

Consultations by the most resource groups should be prevented. Participants are given clear access to materials, equal timelines and the opportunity to present information without complex technical formats. The expert council, in turn, is not obliged to recognize any report as proven; independence implies a fair procedure for assessing facts, not voting for a scientific result.

Control issues and minimum evidence

Verification Element	Evidence criterion
Inclusion	Miscellaneous disciplines and groups
Role	Advice separate from approval
Response	Document the record of objections
Effect	Examples of changes in methodology after participation

Key sources for comparison: [1], [3], [4], [18].

Practical case analysis 20.32

Recommendation: Agree with the applicant's expectations

Hypothetical example: The man expects the publication to automatically cancel the bank's refusal.

Analysis: Research does not guarantee an individual solution.

Practical steps: The coordinator explains the steps of the verification, the possible outcomes and a separate route for protecting rights.

Outcome criterion: The applicant understands the status of participation and how to proceed.

33 Publication procedures and categories of material

Types of issues

Four classes of publications are recommended: an operational analytical note, a thematic monitoring report, a methodological standard and an annual institutional review. The operational note may be brief, but must indicate the status of verification and updates; the thematic report reveals sample and assumptions; the standard describes the applicable methodology; the annual review also records the Observatory's own errors. The genre name prevents a situation where the initial signal is presented as a final scientific result.

Editorial standard

Each document contains the cut date, authors or editorial responsibility, sources, method of inclusion of events, legal restrictions, definition of key concepts, right of reply and history of

changes. In case of serious uncertainty, the relevant reservation appears in the summary rather than being hidden in the note at the end. For publications intended for regulators or courts, it is further recorded which statements are based on the original documents and which remain expert interpretation.

Public data

Disclosure of machine-readable data is desirable, but only after legal and ethical evaluation. The registry may publish anonymized codes, dates and confirmed categories without distributing personal files. The metadata includes the classifier version, the update history and reuse conditions. If open publication is not permitted, it is necessary to explain the restriction and provide statistical aggregates or controlled scientific access procedures when possible.

Evidentiary levels and conditions for publication

It is proposed that the status of the publication be clearly "preliminary observation", "verified research report", "the methodological position" and "the correct version". This facilitates the work of the press and legal advisers, as some materials are intended for early risk identification and others for well-founded comparative conclusions. In operational material, it's particularly important not to hide uncertainty for shortness. The title should not confirm more than the main part.

Publication of open data is permitted after an independent test for disclosure of personal information and verification of the conditions for secondary use. A complete technical log, bank correspondence or individual case file is not an open database because it's of interest to researchers. If it is not possible to disclose the initial documents, it is permissible to publish descriptions of the verification procedure and a limited set of aggregated data sufficient for an independent understanding of the conclusion.

Control issues and minimum evidence

Verification Element	Evidence criterion
Signal	Preliminary warning note
Overview	Sampling method and restriction
Standard	Version of methodology and withdrawal of reviewers
Data	Licensing and protection of personal data

Key sources for comparison: [7], [8], [13], [14].

Practical case analysis 20.33

Recommendation: Publish material by validation level

Hypothetical example: They want to release an emergency signal in the final investigation format.

Analysis: The format creates excessive confidence.

Practical steps: The editor selects a preliminary message with verification issues and the terms of the update.

Outcome criterion: The title and content correspond to the evidentiary status.

34 Managing corrections, withdrawals and challenges

Right to correction

The organisation must provide for a procedure for contesting publication: the applicant indicates a specific thesis, presents the basis of the objection and the method of communication. The editorial board confirms the receipt, determines the severity of the risk, appoints an independent verifier and makes a reasoned decision. The correction may include updating a fact, changing the degree of certainty, clarifying the legal status or recalling a fragment. The objection shall not become the right of the subject to be examined to stop an undesirable investigation.

Versioning

A silent replacement of a PDF with no history of changes undermines traceability, especially if the research is cited in legal documents. Each report receives a version number, a correction log and a steady link. If the error is serious, the old version is marked as being replaced, and the new one indicates the cause. An exception is allowed for cases where the preservation of an erroneous version itself reveals dangerous personal data; then the story is described without the preservation of illegal content.

Independent review

If there is a conflict between the author of the publication and the applicant, the re-examination request is sent to a specialist who was not involved in the original study. The decision recorded the evaluation of the evidence and the residual uncertainty. The number of satisfied complaints should be published together with the number of all applications and clarification of the procedural criteria. A high correction rate can mean both a poor quality of the first check and a transparent ability to correct errors; both explanations require study.

Correct error while keeping track of the versions

A correction system requires a distinction between technical imprinting, actual error, disagreement with interpretation and a new subsequent event. They are not the same "corrections". For example, a change in the law after publication does not mean that the original legal description was incorrect at the date it was made; updating is required. On the contrary, a person who is misnamed demands that an erroneous fact be corrected and that his reputation be evaluated.

In the event of a material error, the editorial board notifies known direct users of the conclusion when possible and proportionate. The version retains the number, date and list of significant changes. The dispute procedure does not automatically have the right to remove a critical but verified judgment; the applicant is guaranteed to consider the arguments and a reasoned answer, and the editorial board is obliged to correct the identified deficiency.

Control issues and minimum evidence

Verification Element	Evidence criterion
Treatment	Specify disputed argument and ground
Triage	Priority for the risk of immediate harm
Version	Date, content and cause of correction
Appeal	Independent review by the author

Key sources for comparison: [13], [14], [18].



Figure 20.4. Correcting a public record. Procedure in Section 34. A correction preserves the version history and an accessible explanation of any change to a material conclusion.

Practical case analysis 20.34

Recommendation: Save the history of corrections

Hypothetical example: Once the new document has been received, the editor changes the conclusion in silence.

Analysis: The readers of the previous version continue to use the wrong text.

Practical steps: The person responsible issues the dated correction and links it to the previous version.

Outcome criterion: The user sees what has been changed, why and what version is relevant.

35 Research quality control and peer review

Before publication

Before publication, the research question, sources, method and conclusions are checked. The methodological reviewer should restore the basic calculation and evaluate the alternative explanation; the subject expert - check the meaning of the classification; the lawyer - the applicability of legal claims. The editor shall record the substantive observations and the reasons for their decision. The work is designated as having undergone external scientific review only if the relevant procedure has been actually performed.

Segregation of checks

Independent coding is performed prior to joint discussion. The original assessments are kept in separate records; the agreed category is the subsequent decision with a reasoning. Inter-expert consistency characterizes the reproducibility of the classification, not the truth of the original message. Therefore, alongside its evaluation, the quality of the sources, the completeness of the materials and the suitability of the event itself are checked. [M1; M3]

Post-publication control

Once a card sample is published, it is rechecked from source to output. The selection of the most sensitive cards for evaluating the overall share of errors and the targeted selection of the most sensitive cards are carried out separately. In the first case, the inclusion probability and uncertainty are described; in the second, the results are used to identify significant defects without stating the frequency throughout the body. The corrections are returned to the classification and collection procedure.

Independence of review and inter-expert coherence

The reviewer receives the research question, the inclusion rules, the category dictionary, the available primary materials and the calculation. Its task is to determine whether a conclusion should be drawn from the procedure and data, rather than confirm the editorial's pre-selected position. Conflict of interest and access limits are indicated before work. For re-examination, comments and replies are retained, including rejected statements with reasons.

The independent re-marking must include ordinary, borderline and insufficiently documented cases. After the error is detected, the instructions are clarified and their suitability checked on new materials. Using the same set after a joint agreement evaluates learning by example, but not the tolerability of the rule. The focus is on differences that change the meaning of a public conclusion or risk to the person concerned.

In an illustrative dataset of 40 records, two independent assessors agreed in 32 cases, and each classified 20 records as positive. Observed agreement is 0.80, and agreement expected from the products of the marginal proportions is 0.50. Cohen's κ is therefore (0.80 − 0.50) / (1 − 0.50) = 0.60. This is a hypothetical calculation, not an ARGA finding. The coefficient must be interpreted alongside the disagreement table, class distribution and nature of disputed cases; it does not establish a universal threshold for the suitability of the method. [M3]

Analysis of agreement in independent coding

The hypothetical calculation of κ in Section 35 is based on the following joint table. The categories indicate whether a predetermined event attribute is present or absent; classification as positive does not imply legal culpability.

Valuationr A	B Positive	B Negative
Positive	16	4
Negative	4	16

Agreement is observed in 16 + 16 = 32 of the 40 records. Each assessor used the positive category in half of the cases, so baseline agreement expected from the marginal proportions is $0.5^2 + 0.5^2 = 0.50$. The resulting $\kappa = 0.60$ is accompanied by an analysis of the eight disagreements. If they arise from a missing primary document, improving agreement requires better source material; if they concern the category definition, the instructions require clarification. [M3]

Once agreed, reference marks and a separate final category are retained. Reviewing consent only on jointly corrected responses does not evaluate the original reproducibility. The new instruction shall be checked on a fresh set. For a rare category or other distribution of classes, the coefficient may behave differently, so it always publishes the original numbers and the substantive value of the differences.

Control issues and minimum evidence

Verification Element	Evidence criterion
Editor	Language, structure, accuracy
Scientific reviewer	Method and sustainability of conclusion
Legal officer	Applicability and risks of disclosure
Data auditor	Origin, calculation, reproducibility

Key sources for comparison: [9], [14], [16].

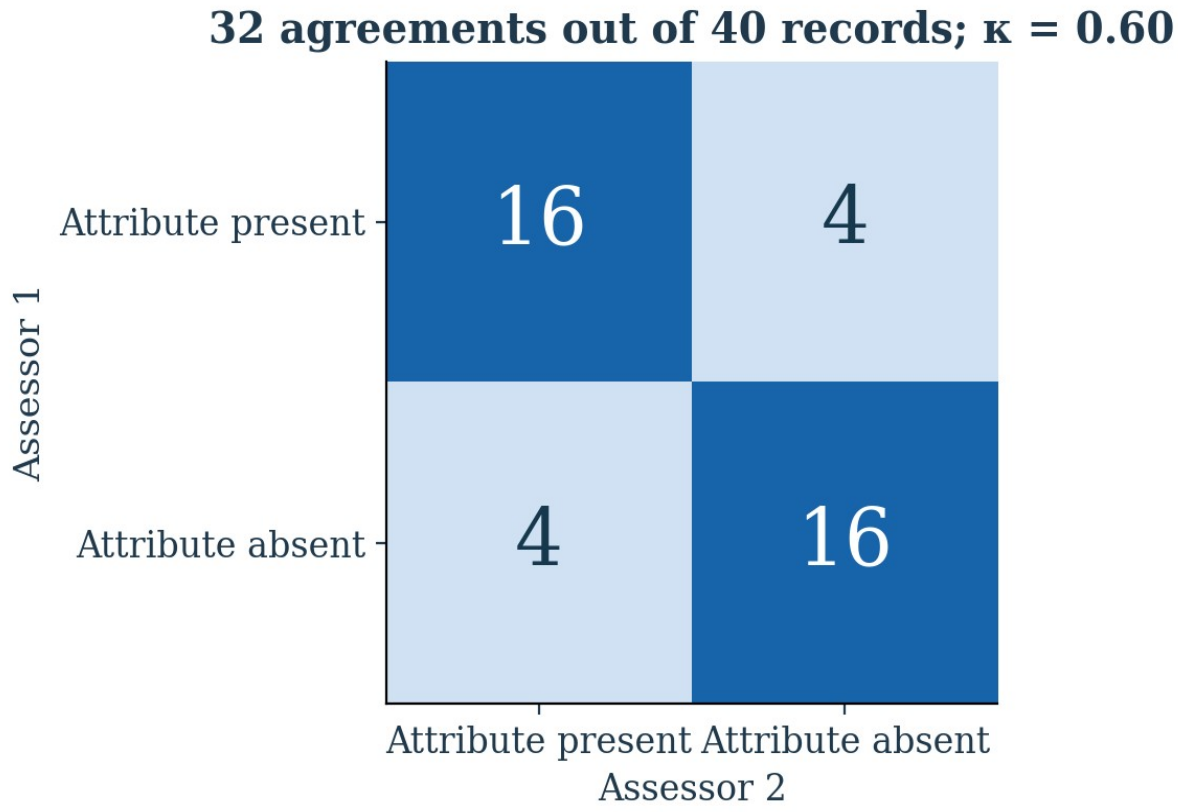


Figure 20.3. Agreement in independent coding. Hypothetical data from Section 35: agreements 16 + 16, disagreements 4 + 4. Observed agreement is 80%, expected agreement 50%. The calculation is not an ARGA finding.

Practical case analysis 20.35

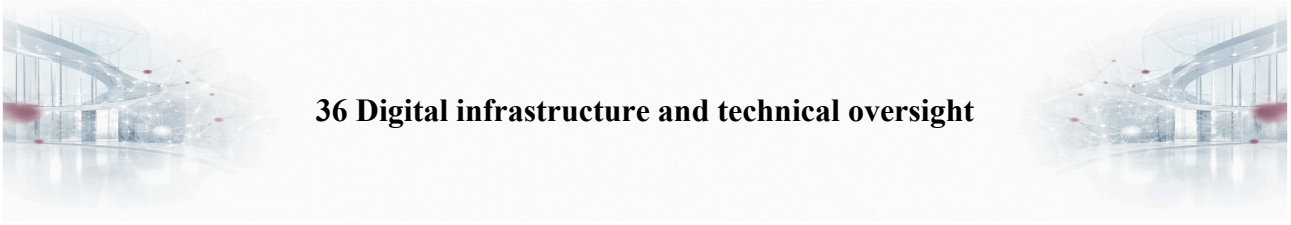
Recommendation: Conduct an independent review of key findings

Hypothetical example: The author simultaneously selected the sources and confirmed the causal conclusion.

Analysis: Undetected displacement can go through the whole process.

Practical steps: The curator gives the reviewer the source, the selection rules and alternatives, keeps the answers to the comments.

Outcome criterion: Significant differences are resolved or reflected as limitations.



36 Digital infrastructure and technical oversight

Data architecture

The practical observatory requires a minimum infrastructure: a separate system register, event register, evidence repositories, access logs, task systems and a publication portal. The connections between the registers are expressed by stable identifiers without duplication of personal data. The public part shall not have direct access to a closed evidentiary archive. All changes to the cards are accompanied by a log of the author, time and reason, and removal is carried out taking into account the procedures for preserving legally significant materials.

Automation

AI tools can help detect duplicates, translate documents, extract metadata, and classify messages, but the final assessment of harm and status must be human-controlled. The use of the generative model in the analysis is documented: tasks, version of the tool, error control procedures and data transmission restrictions are indicated. It is not possible to send the applicant's classified material unnoticed to external AI services without appropriate legal and technical assessment.

Sustainability

Minimum measures include backup, multi-component authentication, key management, export restrictions and regular recovery checks. The system has to withstand the denials of the supplier, because losing the archive can destroy the chain of evidence. For public data, it is desirable to download versions independently, while for confidential data, it is desirable to have a name and a release record. The security plan is reviewed after an incident or significant infrastructure upgrade.

Protection of archives and reproducibility of calculations

The technical infrastructure should provide verifiable history, not just a convenient interface. The published identifier of the record is linked to the evidence archive via a closed match table, which allows you to correct the public record without sharing personal data. Every export requires a schema check: whether the metadata contains hidden authors, addresses, coordinates, or original file names. Highly sensitive data is subject to limited access and specific backup rules.

In automating event selection, classification errors are inevitable. So AI tool has to give out candidates and a human-explained record, and the final decision is confirmed by the analyst. Using the same generating tool to create a message and independently verify it does not guarantee independence. Periodic tests take into account new versions of the model and possible changes in translation or extraction quality.

Control issues and minimum evidence

Verification Element	Evidence criterion
Register	Versions and an unaltered action log
Archives	Closed locator
Publication	Checking personal attributes prior to export
Automation	Only with documented human verification

Key sources for comparison: [13], [14], [15].

Practical case analysis 20.36

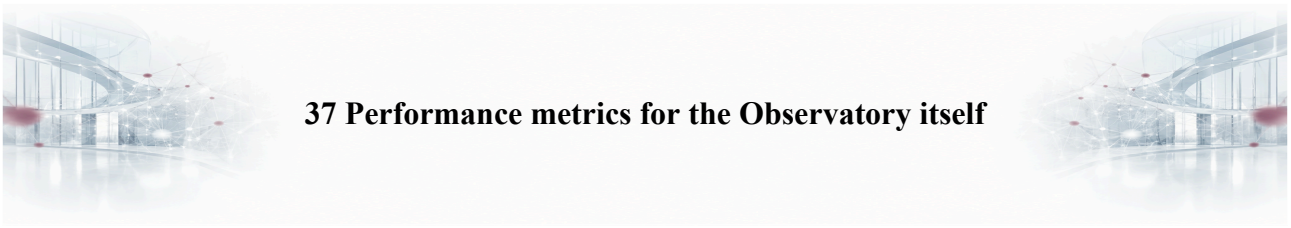
Recommendation: Check recovery and separation of access

Hypothetical example: The archive is reserved, but reconstruction has never been tested.

Analysis: The availability of a copy does not confirm the possibility of returning the work data.

Practical steps: The Administrator restores the training package in a separate loop and checks the roles.

Outcome criterion: The test of restoration and denial of unauthorized access is completed by protocol.



37 Performance metrics for the Observatory itself

Quality instead of PR

The Observatory's activities are assessed by the properties of the research process: verifiability of data origin, reproducibility of classification, timeliness of corrections and preservation of significant limitations in publication. The number of materials produced shows the volume of work. It does not measure the credibility of the corps or reduce public harm. For each indicator, the source, denominator and possible method of artificial improvement are recorded.

Performance

The practical effect is assessed through confirmed changes in procedures, the use of recommendations in public documents, the quality of respondents' responses and the availability of applicant protection mechanisms. Quoting the report does not prove that the behavior of the organizations has changed because of it. Other reforms and temporal factors need to be taken into account for the causal conclusions. The annual report may refer to an "observed contribution" but not automatically state its own merit in a public or judicial decision.

Countermetrics

To prevent misrepresentation, it's helpful to publish rejection figures for insufficiently verified materials, the number of correction requests and the proportion of cases where an independent review failed to confirm a preliminary hypothesis. It's not a failure of a research organization, it's an exercise of methodological discipline. Management should reward the quality of rejecting an unconvincing conclusion as well as the quality of the proven result.

Segregation of output and social impact indicators

For institutional reporting, it's useful to set opposite pairs of indicators in advance. The number of published cards is accompanied by the percentage of cards that have been rejected due to weak sources; the number of recommendations - the percentage with confirmed feedback; the speed of publication - the number of material corrections. Such a set limits the incentive to over-inflate. Management should not reward individual researchers for the number of scandals or posts without quality control.

Comparisons of work by year take into account changes in staffing and availability of sources. The increase in the number of events detected may reflect the development of the Observatory itself. For an international KPI, it is necessary to publish active languages, countries, source types and the proportion of confirmed records. If the budget for independent review has been reduced, this should be reflected in the annual report as a quality restriction.

Control issues and minimum evidence

Verification Element	Evidence criterion
Quality of sources	Percentage of primary support cards
Independence	Percentage of proceedings with challenge disclosed
Error	Time limit for correction of the substantiated claim
Impact	A confirmed response from the addressee, without merit

Key sources for comparison: [8], [14], [20].

Practical case analysis 20.37

Recommendation: Measure the quality of your own work

Hypothetical example: The team reports on the number of cards, except for the corrections.

Analysis: The growth of the output could hide the deterioration of the test.

Practical steps: The analyst takes into account significant errors, correction times, and re-review on a given sample.

Outcome criterion: The report shows the quality and limitations of the evaluation along with the volume.

38 Cooperation with international organisations

Participation models

For an independent expert structure, legitimate forms of participation in public consultations, thematic discussions, scientific events and research networks are possible. They do not mean automatic accreditation to the UN or to the status of an OECD body. Any declaration of cooperation must be in the form of an actual comment, a confirmed participation, a joint publication or an institutional agreement. The mixing of forms is detrimental to the credibility of scientific independence.

Compliance with mandates

When working with UNESCO, the country's RAM and EIA tools are useful, and with the OECD - single definitions of incidents and principles of AI governance. For the UN research group, aggregated risk data and research methods are relevant, but the channels of submission are determined by the official procedure of the relevant body. The Observatory should not send personal material to an international forum just because the issue is cross-border. Preferably a carefully aggregated analyst with a methodological application.

Joint research

Potential partners can coordinate data dictionaries, independent review, observation techniques and bilingual annotations. The treaty establishes the possession of the results, the right to publish, data protection and the procedure for resolving scientific differences. A joint logo is inadmissible without the real consent of another organization. A public website should contain verifiable evidence of the claims of connection: official letters, event pages or links to joint materials.

Separating collaboration and institutional recognition

Before the declaration of cooperation with the international organization, an internal document validity test is offered. What do you have: an invitation letter, a participant's registration, a confirmation of receipt of the study, an official joint project or just the fact of citing the methodology? Each circumstance requires its own formulation. A directed comment should not be made into a statement of official approval. The status information is checked on the public pages of the relevant organisations where such pages exist.

The joint methodological project should identify in advance the ownership of the source data, the publication autonomy, the order of possible corrections and the right to terminate cooperation. For international forums where voluntary expert materials are accepted, it's better to provide a short verifiable note with an applied methodology than a large report with uncertain quality of evidence.

Control issues and minimum evidence

Verification Element	Evidence criterion
United Nations	Actual channel of participation, not implied status
UNESCO	RAM/EIA Comparison
OECD	Incidents and interoperability of metrics
Treaty	Maintenance of independent editorial decisions

Key sources for comparison: [1], [2], [3], [4], [5], [7], [8], [21], [22].

Practical case analysis 20.38

Recommendation: Confirm institutional statements with documents

Hypothetical example: Confirmation of receipt of a comment is referred to as an official partnership.

Analysis: The administrative reply does not confirm this status.

Practical steps: The editor checks the wording with the letter and selects the exact description of the participation.

Outcome criterion: Each application for cooperation has the appropriate basis.



39 The Observatoire ARGA pilot programme

Pilot instead of declaration

The pilot starts with one leading subject area and a limited set of sources so that the method errors can be distinguished from the differences between sectors. For ARGA, documented cases of misidentification in sanctions screening can be an appropriate research subject if legal and verifiable materials are available. The boundaries of the language, territory, period and canals are fixed prior to collection. Two other areas remain candidates for expansion after the first series is evaluated.

Phases

The program includes the development of a dictionary and selection rules, test marking of learning cards, assembly of a limited body, independent coding of a new subset and issuance of a methodological report. Each stage has a verifiable result. Before receiving sensitive materials, the legality of processing, access rights and restoration of the archive are checked. The duration of the stages is determined by a resource decision, rather than by an actually approved calendar.

Criteria for continuation

The transition to expansion requires a reproducible path from the source to the record, explanatory classification differences, working corrections, and sufficient data protection resources. If these conditions are not met, a specific refinement or narrowing of the observed area is determined. The pilot may end up with a justified refusal to compare countries or calculate frequency while preserving the usefulness of documented individual cases.

Pilot launch: phases and stopping criteria

Prior to collection, the pilot's questions are recorded: is the selection reproduced; do evaluators distinguish between the main categories; can the source be traced; how does the appearance of a new channel affect it; how long does it take to check and correct? Numerous criteria are taken with the consequences of the error and the resources, before getting acquainted with the final results. Their modification requires a reasoned recording. The adverse result of the pilot is maintained as a basis for improving the methodology.

Both positive and negative signals influence the decision to continue. If personal data has not been protected or independent verification has failed, the program stops a sensitive channel, even if public interest is high. The pilot report includes expanding requirements: a new language, a new sector, additional reviewers, a budget for security and a proven need for resources.

Before the pilot begins, there are three levels of access: open aggregated facts without identifiable data; limited working records with verified grounds; specially protected individual materials that do not begin to be accepted until legal consensus and information security testing. This separation prevents a technically ready site from being mistaken for an authorization to store sensitive files. The transition between levels shall be established by a separate protocol and recorded in the register of decisions.

The pilot's final report contains the flow of inclusion and exclusion of events, a joint independent coding table, the causes of significant discrepancies, the analysis of sensitivity to sources, and the actual testing load. Storage safety and the correction procedure shall be checked separately. On this basis, the following language, sector or channel is decided. The lack of critical finding in a small series of studies is not interpreted as proof of the safety of the entire infrastructure.

Phase	Continuing/stopping criterion
Before the reception opens	Roles, GDPR verification, security measures and independent approval of the procedure
Test set	Traceable origin of the material and recorded limitations
Independent coding	Coherence assessed, disputed categories described and method revised
Publications pilot	Right of reply worked, there's a change log and an understandable policy of correction.
Scale	No unresolved critical conflicts, security violations and access gaps

Control issues and minimum evidence

Verification Element	Evidence criterion
Step 1	Statute and sources of money
Phase 2	Non-public test sample
Step 3	Pilot methodological report
Step 4	External assessment and decision on scale

Key sources for comparison: [3], [4], [7], [8], [14].

Practical case analysis 20.39

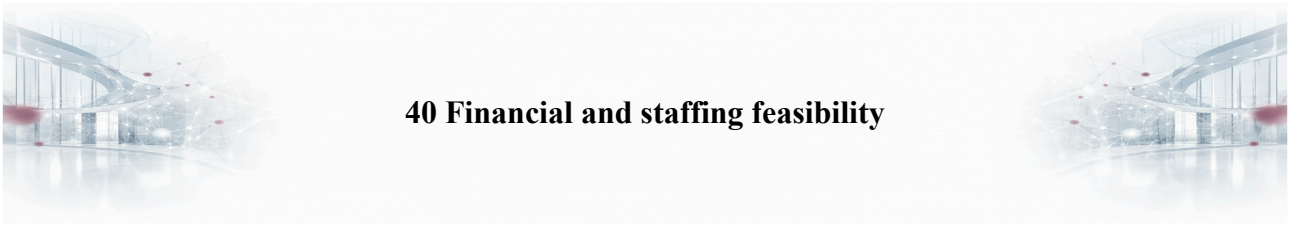
Recommendation: Pilot limited hull before scaling

Hypothetical example: The team plans an international rating before testing the coding.

Analysis: Unagreed categories would make the comparison arbitrary.

Practical steps: The curator conducts the pilot on a confined body, independent marking and sensitivity analysis.

Outcome criterion: The decision to expand is based on the quality of the procedure and the gaps identified.



40 Financial and staffing feasibility

Functions ahead of posts

The budget should be structured according to the functions: research, legal verification, data management, information security, translation and independent review. In the first step, functions are allowed to be combined, but not without conflict restrictions. For example, one person may collect and code primarily, but he or she must not be the only person to confirm his or her own contentious conclusion. External legal examination and protection of applicants are provided in advance, otherwise the organization will be forced to abandon the most socially significant cases.

Budget scenarios

Since the actual contracts and the payroll fund for the projected centre are not presented in this study, no conditional amounts can be issued for the confirmed ARGA budget. It is recommended to use the "minimum", "baseline" and "enlarged" scenarios with cost ranges after internal calculation. Each scenario separately takes into account editorial independence, data storage, language support, and the response reserve. Savings due to lack of data protection or reviews are unacceptable as a sustainable way to finance.

Deficit management

In a lack of resources, the Observatory should reduce the number of observed objects rather than standard verification. Publications with a weak evidence base can attract short-term attention, but they increase risk to the people affected and trust in the organization. The refusal of a sensitive case due to the unavailability of an independent verification is documented without disclosing the applicant. Employees should not rely on the number of critical citations or court rulings as an indicator of personal performance.

Verification of the resource model on operational data

The resource model is only useful if uncertainty is revealed honestly. Instead of the arbitrary amount of the "annual budget of the International Observer", it is recommended to calculate the cost of specific transactions: the hours of the analyst per record, the time of the lawyer, the translation of the document, the independent reviewer and the secure storage. These parameters are first measured on the pilot and then scaled up. If the costs are not known, "calculation" is required, not the fictional figure.

In the event of funding shortages, an organization must maintain a minimum standard of evidence, reducing the number of topics. The conditional expansion plan provides for a separate reserve for adjustments, emergency protection of applicants and technical incidents. This is especially important for a public organization whose reputation depends not so much on the number of published documents as on its ability to provide evidence for each one.

Control issues and minimum evidence

Verification Element	Evidence criterion
Study	Accessible analytical capacity plan
Legal control	Separate external budget
Information security	Mandatory function even on a pilot
Reserve	Examination and urgent data protection

Key sources for comparison: [14], [16], [18].

Practical case analysis 20.40

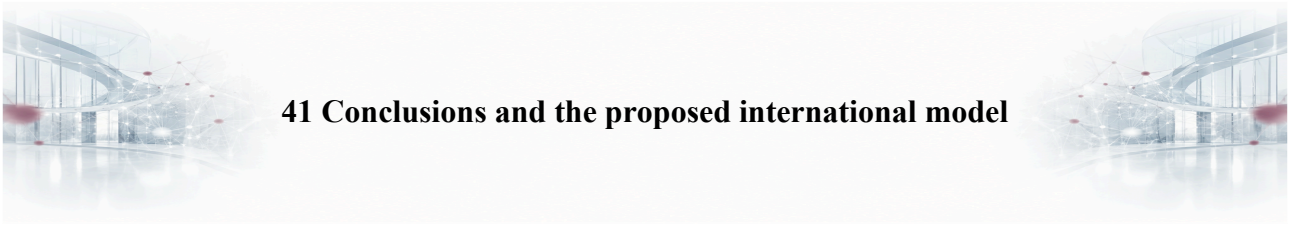
Recommendation: Link promises to human resources

Hypothetical example: Two staff members promise a daily assessment of all cases received.

Analysis: The load may exclude review and timely correction.

Practical steps: The manager evaluates the time on the record, introduces priorities and reserves for complex cases.

Outcome criterion: The plan corresponds to available time, while maintaining quality control.



41 Conclusions and the proposed international model

Institutional Conclusion

The scientific status of an independent AI Observatory is determined by the reproducibility of the observation and the validity of the transition to a conclusion. Content analysis provides uniform classification rules, selection methodology reveals selection constraints, and measurement theory gives meaning to indicators. Institutional independence protects the ability to publish an inconvenient observation and correct one's own mistake. The joint fulfilment of these conditions creates a verifiable research function.

Priority decisions

Initially, it is proposed to approve the research protocol, category dictionary, rules for dealing with sources and conflicts, data protection regime and correction procedure. Then conduct a pilot with independent coding and sensitivity analysis. International comparison is only permitted for indicators with a justified comparability of definitions and observations. The formal unity of the record form does not replace this check.

Final assignment for ARGA

The ARGA's practical path is to confirm the capability of a limited observation, then expand the field based on the test results. Publications should keep a distinction between a registered message, a confirmed event, an investigated cause, and a prevalence assessment. Materials may support

consultation and protection of rights within the limits of their probative power; research status does not confer on them the authority of a court or regulator.

Institutional implementation conditions being tested

Before scaling, it is necessary to check the stability of selection and coding on new material, the impact of language expansion, resource availability of review and correction execution. In the case of differences between groups, differences in observation are first considered. Only after such explanations have been excluded or substantiated can a wider conclusion be drawn. An unresolved gap restricts a specific indicator, not hides in the general note.

The next stage of the study should show which elements of the proposed procedure are reproduced by different specialists and where a change in method is required. Until the pilot is completed, the report is a science-based infrastructure project. Its practical value lies in verifiable launch, analysis and review criteria that allow for the separation of a confirmed observation from a broader claim for which data is not yet sufficient.

Control issues and minimum evidence

Verification Element	Evidence criterion
Introduction	Research and public control
Procedure	Methods, data, independence
Conclusion	Level of confidence and limitation
Development	Pilot - verification - scaling

Key sources for comparison: [1], [2], [4], [6], [7], [9], [14].

Practical case analysis 20.41

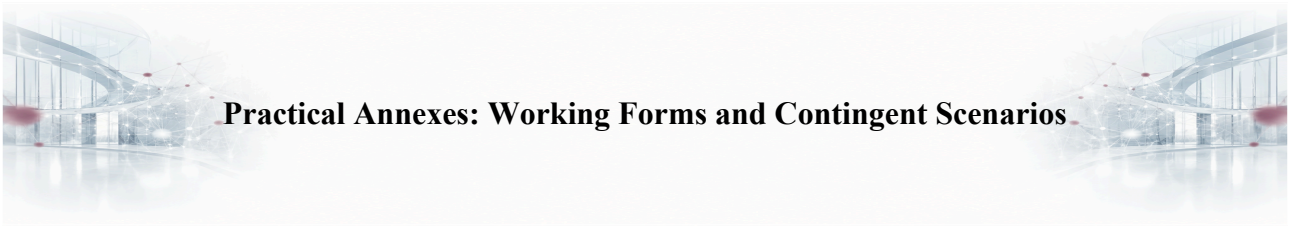
Recommendation: Make a decision to launch on confirmed readiness

Hypothetical example: The register is technically operational, but there is no correction procedure and no complaints authority.

Analysis: Launching will increase the impact of errors to the readiness of the protective processes.

Practical steps: The curator checks the protocol, roles, protection, review and educational correction; eliminates unclosed critical issues before launch.

Outcome criterion: The decision shall contain evidence of readiness, limitations and conditions for review.



Annexes are project tools. Their fields and procedures are subject to customization taking into account the applicable legislation, the nature of the organization's data access and resources. These scenarios do not describe actual customers or established irregularities. The existence of the form does not confirm its approval as an official internal regulation.



Appendix 1 Statute of the independent AI Observatory

Appointment

The draft statute establishes the social and analytical nature of the organization. Its main tasks are to collect verifiable data on the use of AI, develop risk measurement methods, publish reports and promote meaningful international discussion. The Observatory is not a state regulator, certification body or judicial institution and may not require disclosure of data outside the procedures provided for by law. The charter should have priority over commercial contracts that can influence editorial decisions.

Application and limitation

The charter is adopted by the competent body of the organisation after legal verification of the applicable law. Recommended wording: "Methodological findings are based on documented studies and the donor or customer is not entitled to determine the outcome of the evaluation". The powers of the Council for the appointment of an external review, the procedure for reviewing the methodology and the independent procedure for working with complaints are provided for separately.

Field/control	Rule of filling or recommended value
Subject	Public scientific and analytical monitoring of AI
Limits of powers	No State coercion or certification
Organs	Council, secretariat, ethics committee, revision
Independence	Prohibition of the purchase or censorship of results
Publication	Mandatory methodology and history of corrections
Revision	Annual evaluation of the statute and practice

Testing in practice

Practical test of the charter: a specialist who has been offered to conduct a study at the expense of a verified developer is obliged to check whether the donor can exclude critical conclusions. If so, the draft was not declared independent. The document provides for a separate annual public report on the corrections and sources of funding. The review of the charter is carried out in a documented manner; the council cannot change the rule for the sake of a specific case. When the Observatory is mentioned externally, the actual legal name of the organization and the nature of the activities carried out are indicated.



Appendix 2 Charter of Independence and Editorial Guarantees

Appointment

The Charter applies to all research, including commercially paid research. It prohibits the financing party from agreeing on its findings in advance, the right to unilaterally remove critical remarks and

the requirement to conceal the very fact of a substantial conflict. Access to the drafts may only be granted for the purpose of verifying and protecting confidentiality by documenting the comments received.

Mandatory disclosure in the project charter: Observatoire's liaison with ARGA and other organizations under general management; categories of paid services and donors; order of conclusion when investigating a current or former client; independent external channel of communication about pressure; prohibition of making payment conditional on a desired conclusion; correction procedure even if the outcome is unfavorable to ARGA itself. Until these guarantees are approved and effectively verified, the term "independent" is used as an institutional target and not as a proven certified characteristic.

Application and limitation

In the pilot practice, each contract must be tested for conformity with the charter. The violation leads to the suspension of the publication process, the recruitment of an independent lawyer and the possible rejection of the project. The public record discloses the fact that the charter has been applied, but does not disclose employees' personal data or legally protected contractual information without necessity.

Field/control	Rule of filling or recommended value
Buyer	Doesn't support the scientific conclusion
Expert	Opens up relevant interests
Reviewer	Not accountable to the party under review
Right to publish	Restricted only to legitimate grounds
Objections	Accepted and taken into account, but do not grant veto rights
Regulation	Discharge or termination of an incompatible task

Testing in practice

In the contract annex to the charter it is recommended to attach directly: the customer receives the right to submit factual observations, but does not confirm the scientific conclusion. In the event of a conflict between a confidentiality agreement and socially relevant information, the lawyer determines the legally permissible version of the publication: complete, impersonated, aggregated or refused. The lack of the right to disclose the original data does not make it possible to declare the verification a full open audit. Documentation of such a limitation should be a condition for each final report.

Appendix 3 Declaration of conflict of interest

Appointment

The form is completed prior to each appointment and updated when circumstances change. It does not require excessive information about family and private life; only relationships that can influence objectivity are essential. The declaration contains the period of recent professional and financial relationships, participation in the development of the system being verified, and possible competitor interest.

Application and limitation

The decision on the conflict shall be taken by a person or committee not participating in the study. It can include full participation, participation with disclosure, a ban on work on certain parts or a withdrawal. Self-checking by an expert is not enough. The process of conflict management is publicly reflected, not the full closed text of the declaration.

Field/control	Rule of filling or recommended value
Project ID	Relationship to a specific task
Financial linkage	Contract, grant, assets: relevance and period
Professional role	Development, implementation, consultation
Personal and other relations	Only capable of influencing the conclusion
Decision	No conflict / disclosure / limitation / challenge
Checker	Independent of expert

Testing in practice

Example of evaluation: The technical expert was a staff member of the supplier of the model being tested. He can describe the principles of the model and conduct a consultation, but the decision on the validity of its results is left to an independent reviewer. If the expert's involvement is irrelevant to the final conclusion, the role is limited enough. If data cannot be obtained without it, the study is published as having limited independence or deferred. The decision with the date is kept in the conflict register and must be available for internal audit.



Appendix 4 Register of funding and related services

Appointment

A financial register is necessary to assess the risk of single-source dependence and hidden influence on the agenda. It includes monetary financing, substantial natural services, licensing and research data. Significant support cannot be hidden behind the "partner assistance" category if it provides critical monitoring infrastructure.

Application and limitation

The annual review should show revenue shares with a clear denominator; concentration threshold is an internal control signal, not a legal prohibition. For closed contracts, you can disclose the category and share if the exact conditions are subject to permissible protection. The council regularly evaluates the donor's right to terminate funding and the dependence on this risk.

Field/control	Rule of filling or recommended value
Source	Name or legal aggregate category
Form	Money/access/equipment/expert hours
Appointment	Basic, thematic, custom
Risk of influence	Contractual conditions and concentration
Editorial control	Not allowed
Verification	Annual independent review

Testing in practice

Financial revenues are grouped by category, and risks are not only assessed as a percentage of income. Two contracts of the same size can have different effects: one does not affect the editorial, the other contains the right to agree conclusions. For the free computing resources provided, donors have access to non-public data. The report should reveal the method of assessing the significance of in-kind support and the situation when the amount cannot be established. No excess accuracy of the imputed estimates is allowed.



Appendix 5 Matrix of competences and appointments

Appointment

The necessary competencies shall be recorded for each verification. A legal professional does not replace an engineer, an engineer does not establish legal liability without the appropriate qualifications, and the editor does not solve the issue of the admissibility of personal data collection solely by stylistic evaluation. The assignment model should prevent all functions from being concentrated in one staff member.

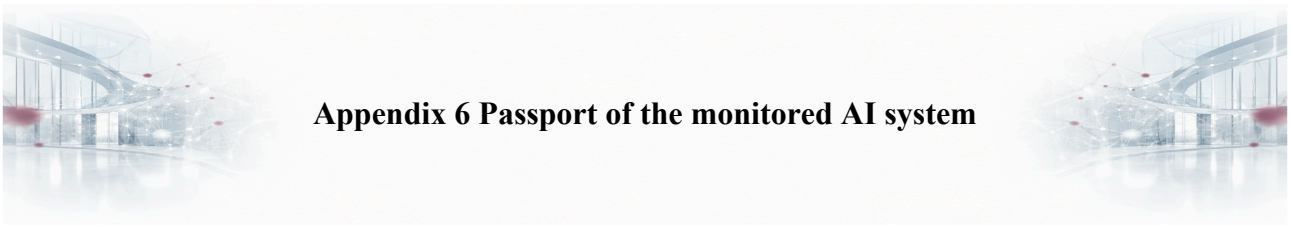
Application and limitation

The matrix is applied until access to the reference points. If the necessary competence is not available, the assignment is reduced or an external specialist is engaged. The conclusion indicates the actual professional scrutiny of each participant, which is particularly important when using the material in courts or public consultations.

Field/control	Rule of filling or recommended value
Law	Accepting jurisdiction and sector
Technical assessment	Model and test methodology
Statistics	Samples, errors and confidence intervals
Data protection	Basis and safety of processing
Edition	Clarity and quality of confirmation
External review	With a high public profile

Testing in practice

The assignment matrix for the verification of automated sanction screening involves a specialist on the relevant sanction regime, an identification algorithm engineer and an expert on complaint procedures. One common law lawyer doesn't provide all three qualifications. The nominating committee checks conflicts before granting access. After the analysis, the report lists not the position for prestige, but the specific actions of each reviewer: he studied the norm, repeated the test, checked the data, or evaluated the way the publication was done.



Appendix 6 Passport of the monitored AI system

Appointment

The system registry record reflects the actual use, not the trademark. It contains information about the operator, the objectives, the affected group, the type of automated conclusion and whether there is human involvement. In the absence of a documentary confirmation of use, the record receives a preliminary status and is excluded from the aggregated statistics of the systems used.

Application and limitation

Each update gets a version number. The risk and event register is attached to the record; changing the model does not remove previous observations. The data on the supplier, national authorities and applicable law shall be confirmed by the primary sources. For a closed product, a "version" is allowed, but not an inventive technical description.

Field/control	Rule of filling or recommended value
System/version	Steady identifier
Appointment	Decision, recommendation, ranking, generation
Deployer	Actual use operator
Geography	Where applicable, where data
Status	Pilot / operating / terminated / unknown
Evidence	Document and date of verification
Legal map	Category and date of application of the rule

Testing in practice

If the same program name is used by five departments with different data sets and thresholds, one parent record and five operating cards are created. This is the only way to distinguish between a fault in basic technology and an error in local settings. The date of purchase shall be recorded separately from the date of commencement of service. In the absence of independent validation of implementation, the record is retained as a "project announced" without being included in the quantity of AI systems used.



Appendix 7 AI incident report record

Appointment

The form separates the description of the harm from the hypothesis of its cause. The applicant can document the decision and the consequences without knowing the architecture of the model; the absence of this information is not a reason to automatically refuse registration of the application. However, the public status of an AI incident is only given after verifying the role of AI and the link to the actual harm.

Application and limitation

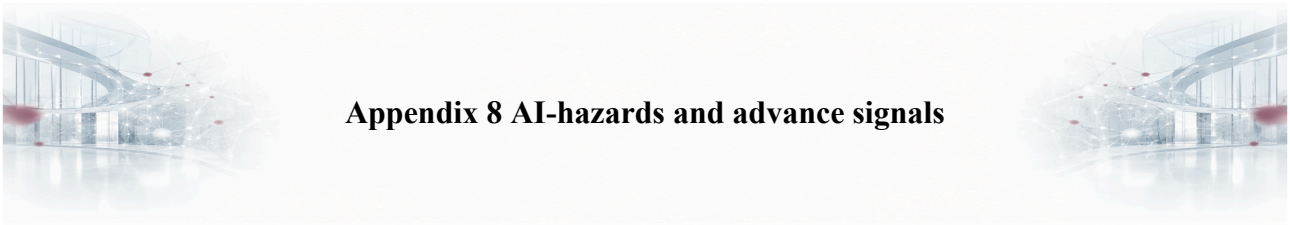
The circumstances of the case are kept in a closed loop until the safety of the publication is assessed. When it comes to multiple victims, there are related episodes. The time scale shall be recorded in a neutral form, indicating the source of each event. Verification should allow competing causal explanations.

Field/control	Rule of filling or recommended value
ID Events	Unique and stable
Date	Time of event and time of message separately
Harm	Relevant harm / Possible risk / unknown

Field/control	Rule of filling or recommended value
Role of AI	Confirmed, probable or not specified
Source	Primary document or communication
To whom the question is addressed	Organization and date
Public status	Non-public / provisional / confirmed

Testing in practice

Conditional registration: "the bank's client reports a prolonged blocking after an automated signal". Only the fact of the turn is established at the entrance; there is no evidence of the illegality of the lock or the nature of the algorithm. The analyst asks for permissible letters and documentation, records deadlines, and checks for alternative reasons. After analysis, the record can remain a signal or go to a confirmed incident if the actual damage and causation is established. All status changes remain in the version log.



Appendix 8 AI-hazards and advance signals

Appointment

A potential hazard register is necessary for early warning without an unfounded claim of damage. Each record describes the mechanism of potential harm, the necessary conditions, the categories involved and the level of evidence. Particular attention is paid to irreversible consequences where the expectation of a real incident would be contrary to reasonable prevention.

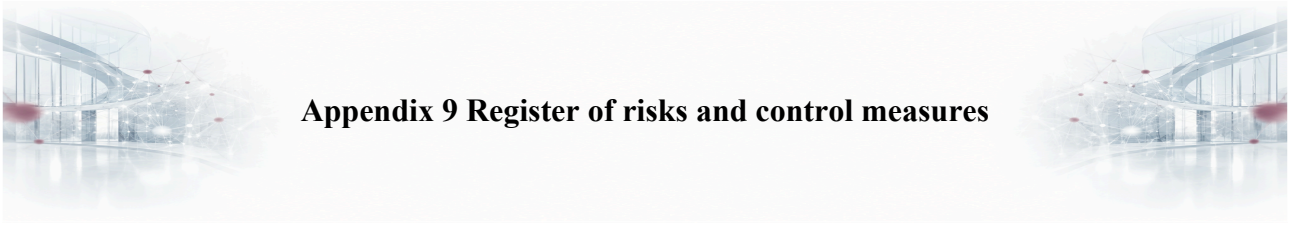
Application and limitation

Transitioning from a signal to a hazard or incident is done as a new version, not editing without history. Public numbers of hazards and incidents are shown separately. An outside expert's assumption doesn't become a confirmed fact because of his professional reputation.

Field/control	Rule of filling or recommended value
Hypotheses	Verification potential mechanism
Trigger	When Risk Can Be Realized
Damage	Category and possible gravity
Confirmation	What facts support probability
Actions	What Can Prevent Harm
Status	Signal / Hazard / Incident

Testing in practice

Hazard doesn't have to wait for real damage, but it does have to show a credible chain from the system to the potential outcome. For example, the lack of a mechanism to correct misidentification in a public service can create a risk of being deprived of assistance. Separately, it should be checked whether there is a manual stand-by process. If sufficient protective measures are found, the risk level shall be reviewed. Such a review should not erase the fact of the original hypothesis and its sources.



Appendix 9 Register of risks and control measures

Appointment

The risk register does not record the level of reputational danger to the developer, but the scenario of possible harm to people, organizations, infrastructure or the environment. The record is linked to a specific system, context and source. Internal priority levels are used for comparisons within a single project, but it is prohibited to give a score for a universal legal assessment.

Application and limitation

When a model changes, a new regulatory solution emerges, or a confirmed protection measure, the risk is overestimated. A distinction should be made between "measures declared" and "measures verified". The residual risk remains open if there is no evidence of real performance. Critical immediate harm scenarios escalate regardless of the total number of points.

Field/control	Rule of filling or recommended value
Scenario	The circumstances of the mechanism
Heavy	Retroactivity, coverage, vulnerability
Confidence	Level of evidence
Monitoring	Declared / verified / unknown
Balance	Post-protection evaluation
Revision	Trigger, date and responsible

Testing in practice

The risk record must contain not only the initial level but also confirmation of a specific reduction measure. If the developer writes "updated filter", it's the status of the application. The effectiveness can be confirmed by repeated independent verification in the same context. The model change date is kept separately so that the comparison before and after does not attribute the result to other circumstances. The uncertainty of residual risk is clearly indicated even when the supplier makes a public promise.



Appendix 10 Key indicator dictionary

Appointment

Each metric has a separate passport record, so different teams and countries don't override the indicators unnoticed. Fields include definition, numerator, denominator, period, source, exclusion, processing of dubs and data quality. In the absence of a metric denominator, it is not called frequency or prevalence.

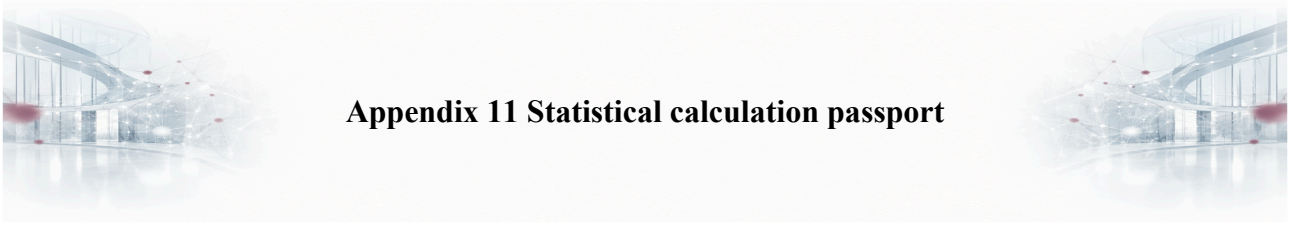
Application and limitation

Versions of the dictionary are published in open form, and when the definition changes, the historical row is either recalculated or split into two incomparable series. Special warnings are needed for the number of calls: it reflects both real harm and a change in trust in the communication channel.

Field/control	Rule of filling or recommended value
M-01	Confirmed incidents: absolute number
M-02	Hazards: A separate series
M-03	Percentage of cards with primary sources
M-04	Time from communication to preliminary decision
M-05	Percentage of organizations responding to the request
M-06	Percentage of corrections issued
Mandatory field	Existence or absence of a tested denominator

Testing in practice

For the time of consideration of an application, it is necessary to determine the time when the countdown begins, for example, when the full application is registered, and the time of completion, such as a reasoned verification decision. If some of the calls are still open, the average time of the closed cases will cause a shift. Dashboard therefore adds the number of open cases and their age. For the proportion of responses, the organization takes into account whether a specific query has been sent and whether a reasonable time limit has expired. Otherwise, the indicator would punish the addressees to whom the request was not sent.



Appendix 11 Statistical calculation passport

Appointment

The form is used whenever the report uses shares, period comparisons or intervals. An example is conditional: 12 verified errors out of 300 analyzed solutions yield 4%, but that's only within the analyzed sample, not across the organization. If the solutions were chosen based on complaints, the result cannot be distributed to the overall flow of customers.

Application and limitation

Confidence intervals require an established selection model. For a purposefully collected sample, the binomial random sample interval can mislead the reader. The report should clearly indicate the source of the displacement and the rejection of generalization.

Field/control	Rule of filling or recommended value
Unit	Decision / user / system / complaint
Number	Watches with a given sign
Celebrity	All suitable objects in the sample
Selection	Random / Full / Recourse
Duplicate	Rules of association
Result	Evaluation only for the appropriate aggregate
Limitation	Why can't you automatically generalize

Testing in practice

In a conditional numerical example, 12 out of 300 is equal to 4%, but the text must reveal the selection of those 300. If only the complaints were investigated, you can't write that an organization has 4% of all decisions wrong. Compared to the other period, the error definition and the share of incomplete records are checked. If the data are insufficient, the conclusion is expressed "4 per cent of the sample surveyed", without being extrapolated to all clients and without artificial precision after the decimal point.

Appendix 12. Source quality classification

Appointment

The classification does not replace the content analysis. The same document can be a strong proof of the existence of a solution, but a weak confirmation of technical reasons. For each statement, it is recorded separately what fact the source actually establishes. A newspaper article quoting a formal decision is not the solution itself.

Application and limitation

If the source contradicts the other, the conflict is not resolved by a mechanical promotion. The analyst searches for the originals, dates and process status, and when it is impossible to obtain documents, leaves the output limited. The public text indicates the nature of the confirmation.

Field/control	Rule of filling or recommended value
A	Official primary document with verified origin
B	Direct participant's material with supporting applications
C	Independent publication with verified references
D	Unidentified or unconfirmed communication
Supplement	Origin and content to be evaluated separately
Outcome	Quality for a particular point, not the author as a whole

Testing in practice

In a conflict of sources, an official act may confirm that a decision existed, but not necessarily all the parties' actual statements. A secondary analytical article may correctly explain the context, but it should lead to the primary source. For a controversial point, reviewer assesses the substantive relevance of each document. If the original is not found, the reference to the citation is not declared as proof of the primary level. This way, it is possible to honestly present available knowledge without unreasonably ranking people for trust.

Appendix 13 Evidence chain log

Appointment

The log allows you to recover the incoming, converting and using of an electronic file. It doesn't create a presumption of authenticity based on one hash: it's necessary to verify the origin and the conditions of obtaining it. The legal basis, access authorization and method of protection of personal data are also recorded for the encrypted material.

Application and limitation

When derivative processing, such as deleting personal data, is done, the contact with the source and information about the operation are retained. Only a permissible identifier is provided in the public version. If legal storage of the originator is not possible, the grounds for removal and the characteristics of the evidence loss are recorded.

Field/control	Rule of filling or recommended value
EVID-ID	Material number
Time	UTC and local time if necessary
Received from	Source, channel and base
Source	Format, reference amount
Operation	Copying, translation, cloaking, analysis
Access	Face, Time, Purpose
Disposal	Date and legal basis

Testing in practice

When you copy a file, the source bytes, the time of receipt and the new work copy hash are saved. The translation is separated from the original and is accompanied by an indication of the translator or the means used; with a substantial legal term, a profile check is necessary. The masking of personal data should be irreversible in the published file, not a surface text painting. If the evidence log is incomplete, the conclusion directly describes this gap and its possible impact on the outcome.



Appendix 14 Report of independent record verification

Appointment

The second researcher should check the role of AI, the category of the event, the primary documents, the timeline and the limitations of the inference. It cannot be limited to checking the orthography or the existence of the said uRL. For technical applications, the ability to reproduce the test and the model/version to which it belongs is further described.

Application and limitation

The reviewer completes a reasoned decision: confirmed, confirmed with reservations, required to be refined, insufficient data, rejected. In the event of disagreement, both entries and the applicable standard of evidence are retained. Publication without resolution of an essential contradiction should be clearly warned.

Field/control	Rule of filling or recommended value
Fact	What is proven and what
Alternative	Which explanation is possible?
Source	Original checked
Law	Is there a legally wrong conclusion?
Data	Rule of law and security of disclosure

Field/control	Rule of filling or recommended value
Decision	Confirmation / Finalization / Failure

Testing in practice

The reviewer asks the compulsory question: Can another expert on the same original documents come to a different reasonable conclusion? If so, the alternative and degree of uncertainty should be described. For example, a government decision may be the result of incorrect data rather than a model defect. Even if the use of I.E. is known, it is necessary to prove its significant role. The review protocol records such alternatives before approving the title of the publication.

Appendix 15 Procedure for the protection of a sensitive complainant

Appointment

The form is used at the risk of persecution, disclosure of migration status, bank circumstances, health status or family details. Before publication, the specialist evaluates the direct identifiers and the combination of indirect features separately. Name and place may be hidden, but rare timeline retains the risk of recognition.

Application and limitation

The decision to publish cannot be based solely on the interest of the media. At high risk, an aggregate review without individual history is allowed. If the information has been received through a lawyer or a trusted person, it is necessary to verify the limits of the powers to disseminate without assuming that the right to represent in the case equals consent to public disclosure.

Field/control	Rule of filling or recommended value
Risk	Physical, social, migration, economic
Identifiers	Direct and indirect
Minimumization	What details are really needed?
Harmonization	Separate from the legality of processing
Alternative	Aggregate or composite scenario
Outcome	Publish / Restrict / Not Publish

Testing in practice

For an asylum seeker, a combination of a rare date of detention, location and outcome of the procedure can uniquely identify an individual even without a name. The editor is checking the possibility of searching this combination in accessible databases and media. At a high risk, the altered details of the composite scenario should be marked as conditional. If the details are necessary for legal reasoning, the material may remain in a closed expert version without public distribution.

Appendix 16 Standard for Request for Comment and Right of Reply

Appointment

The request shall be made after substantial factual grounds have been verified and before a substantial negative conclusion has been published, if no justified exception exists. It sets out specific circumstances, questions and documents that may change the assessment. It is unacceptable to send an emotional charge demanding that the matter be immediately rejected without specifying the point under consideration.

Application and limitation

The time limit for response is fixed and commensurate with the complexity of the response. If the answer is received, the analyst indicates separately which information is accepted and which is not confirmed. In the absence of a response, neutral language was used that did not amount to recognition. Personal correspondence between employees shall not be published unless it is necessary and legally inadmissible.

Field/control	Rule of filling or recommended value
Address	Responsible person or organization
Thesis	Specific and verified approval
Request	Documents, reasons, technical context
Time frame	Smart, well-known.
Confidentiality	Response mode
Result	Received / partially / not received
Account	Where it came from in the report.

Testing in practice

A letter requesting comment must contain a question that can be answered and a clear closing date. For example, "Do you confirm the use of this classification in the said process, and does a documented human revision exist?" There can be no culpability for refusing to answer. When you get an answer, compare it with the original documents and note the difference between confirming the fact, objecting to interpretation and promising to change the procedure.

Appendix 17 Method of correction, withdrawal and replacement of the report

Appointment

The procedure responds to a typical problem: the research organization publishes an important conclusion, then detects an error in the date, identity or legal status of the document. In the case of a minor error, a public erratum is sufficient; in the case of a material - a new version with noticeable notice; in the case of a risk of direct harm - a temporary blocking of an incorrect fragment with a reasoned log of decisions.

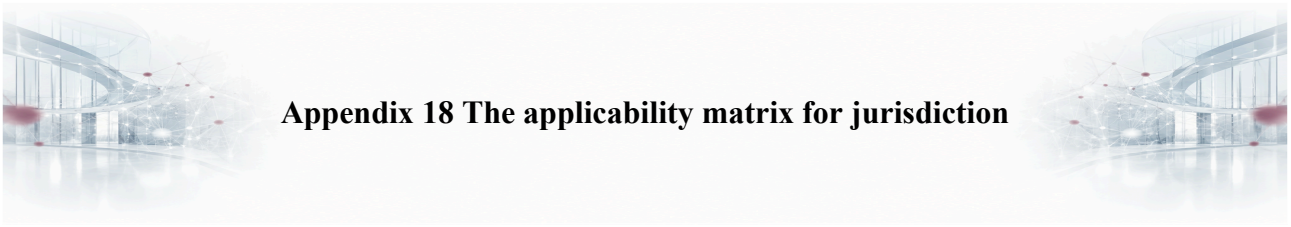
Application and limitation

The final decision on the adjustment is taken by an editorial body not related to the original dispute. History is preserved to the extent compatible with the right to data protection. Users of the materials should have an up-to-date version and an indication of where the substantial conclusion has changed.

Field/control	Rule of filling or recommended value
Type of error	Technical / actual / legal / security
Consequences	Are there direct damage from conservation?
Action	Erratum / update / withdrawal
Review	Independent verification in substantive dispute
Version	Stable number and date
Notification	To whom to send information on the correction

Testing in practice

For a misprint on a small date, an erratum is appropriate. An incorrect indication of a particular person's involvement requires substantial correction, possibly by notifying the addressees who have already received the report. To update the law after the date of the study, a new updated version is issued, not a historical review. When the correction is published, sufficient information about the old error is retained as far as it is compatible with the protection of personal data.



Appendix 18 The applicability matrix for jurisdiction

Appointment

This table is intended for a preliminary map, not a substitute for a legal opinion. The status of the act, its obligation, the applicable articles, the date of commencement of operation for this class of system and the competent authority are recorded separately. The difference between the signing and entry into force of an international treaty, as well as voluntary standards and legal obligations, is of particular importance.

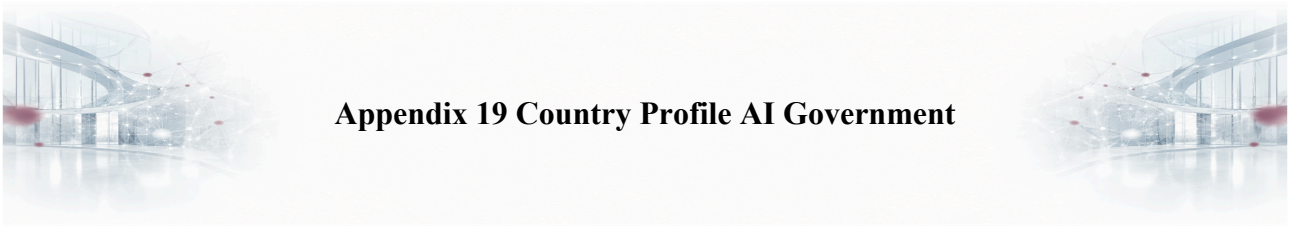
Application and limitation

The verification shall be carried out on the date of the event under review and again before publication. AI Act requires an up-to-date consolidated edition; the Framework Convention requires depository status; the GDPR requires the role of the controller and the nature of the processing. If there is a lack of information, the "fault established" status cannot be attributed.

Field/control	Rule of filling or recommended value
Jurisdiction	Country and applicable law
Tool	Law / Contract / Standard / Management
Status	Operating / transitional / not obliged
Role	Provider, operator, controller, etc.
Evidence	Official source and date
Competence	Who is entitled to impose sanctions

Testing in practice

The conditional table of law may contain the column "act is in effect" and a separate "duty is applicable to the subject". For AI Act, the date and category of the system are sometimes different; for an international convention, the signature is not equal to ratification; for the ISO standard, publication itself does not create a general legal obligation. In case of doubt, the final field contains "a conclusion on applicability is required". Such a mark protects against the spectacular but incorrect phrases "breaches the international standard" without an explanation of the legal force.



Appendix 19 Country Profile AI Government

Appointment

The profile brings together legal information, institutional infrastructure, ways to protect victims and data quality. It is not an international ranking of states and does not imply that the absence of public violations shows successful control. The completeness of complaints may be higher precisely where protection procedures are more accessible.

Application and limitation

The country's sources are indicated with the date and language. All the comparable categories are provided with a mark of comparability. Conclusions about the actual enforcement of the law require independent evidence: published checks, court decisions and measurable results, not the text of the law itself.

Field/control	Rule of filling or recommended value
Law	Obligations and actual date
Organs	Powers and public channels
Practice	Audited cases and decisions
Protection	Complaints, review, judicial remedies
Data	Periods, coverage, uncertainty
Comparability	Which cannot be directly compared

Testing in practice

A country with a developed public complaints system may have more open cases than a country with limited protection. A comparison of only the number of complaints would turn a possible conclusion. The profile therefore contains an assessment of the availability of channels, language coverage and the volume of public solutions. If data on these parameters are unavailable, the Observatory does not build a single AI management quality rating, but publishes individual verifiable characteristics.



Appendix 20 Passport of International Comparison Methodology

Appointment

The comparative report must explain why the selected countries and object classes are comparable. For example, a state with an open procurement registry has more observed implementations than a state without such a registry; the direct rating of the number of AI systems would be distorted. The sample of countries is recorded before the results are obtained, and exceptions are explained.

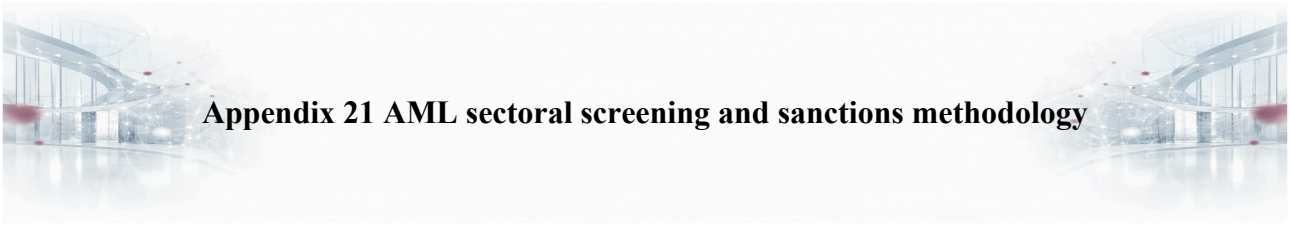
Application and limitation

Different types of sources compare separate domains rather than a single final score. For dynamics, the same definitions are used for years; when the methodology changes, a new base of a series is published. Interlingual verification ensures that legal terms are translated correctly without erasing the peculiarities of national law.

Field/control	Rule of filling or recommended value
Comparative Object	System, sector or procedure
Countries	Selection criterion
Period	Single cut date
Definitions	Translation and match of meaning
Sources	Classes and coatings
Conclusion	Comparable / Part / Uncomparisonable

Testing in practice

A single unit of observation is important for international comparison of sanctions screening. "Prevention", "limited operation", "verified customer" and "disputed decision" are different in nature. Even the same proportion of restrictions being lifted doesn't mean the same technical accuracy if there are different standards, customer bases and policy considerations. The protocol requires a separate table of definitions for each jurisdiction and an honest conclusion about the incompatibility in the absence of such data.



Appendix 21 AML sectoral screening and sanctions methodology

Appointment

It's a research manual, not a bank operating instruction. For the quality of the analysis, it is necessary to separate the signal of match from the final decision, the legal obligation from the internal appetite for risk, the identification dispute from the sanctioning status. The publication must not contain specific control settings that allow the unfaithful person to avoid monitoring.

Application and limitation

The test set should take into account the complexity of transliteration and repeated names. Results are marked by the nature of ground truth; if there is no independent valid markup, technical error rates are not deduced. The speed of the review and the quality of the explanation to the client within the legal limits are assessed separately.

Field/control	Rule of filling or recommended value
Object	An algorithmic signal
Legal role	Bank / provider / customer / regulator
Marking	Source of True Status
Offset	Names, languages, groups
Revision	Result, timing, basis
Publication	Without sensitive control thresholds

Testing in practice

The test on the hypothetical set of transliterations includes similar names, differences in birth dates, rearrangements of surnames and variant spellings. For real sanctions regimes, the final decision requires legal verification, so the lab result is not declared a ground for freezing funds. The team documents separately what types of errors are detected and does not reveal operational thresholds that could facilitate bypassing control. The public section describes the process of human verification and the right to available correction procedures.



Appendix 22 Protocol for the verification of AI systems in international tracing

Appointment

The protocol is aimed at assessing publicly available procedures, not at confirming a particular person's status in closed systems. The analyst separately describes the national decision, the international data exchange channel, the technical match and subsequent court proceedings. No step is automatically taken out of another step.

A separate condition of admissibility: no research record should be turned into a searchable database of charges or convictions. Inbound materials are checked for compliance with Article 10 GDPR and applicable national rules before they are placed in a centralised storage facility. The publication of the contents of the notification, the fact of detention or the extradition decision does not automatically lead to the conclusion that an AI was involved in the event or that the person actually committed an intentional act. [13]

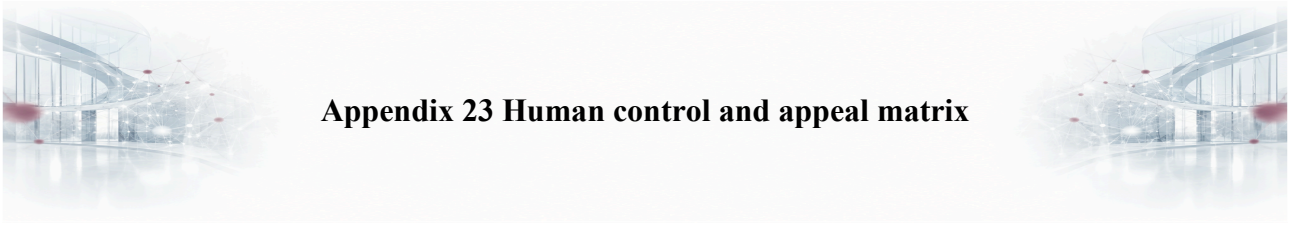
Application and limitation

The use of personal cases for public analysis requires a separate assessment of law and security. The real names, images and biometrics are automatically excluded from the open version. If there is no technical confirmation of the algorithm's use, the conclusion is about process quality and likely risk, not a specific AI defect.

Field/control	Rule of filling or recommended value
Channel	Which procedure is actually confirmed
Document	Authority, date, legal effect
AI Component	Proven function or unknown
Human control	By whom and when tested
Protection procedure	Access to correction
Publication	De-staffing and restrictions

Testing in practice

If a court order confirms a person's release after arrest, it establishes the facts of the case and its own conclusions, not automatically the information system that led to the arrest. The researcher can evaluate the quality of the control chain, the existence of a written basis and access to the correction procedures provided. He cannot confirm the state of the non-public INTERPOL database by public publication. This thesis requires competent official access and separate legal evaluation.



Appendix 23 Human control and appeal matrix

Appointment

The matrix helps distinguish the formal presence of a person from the actual ability to change an algorithmic result. You need information about the employee's authority, access to the source data, time to verify and independence from automated recommendation. The mere signature of the official does not prove a meaningful revision.

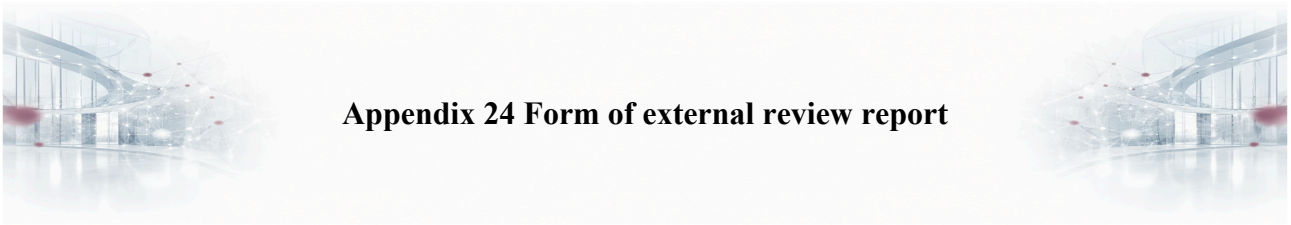
Application and limitation

For the victim, it's important to be notified, to have a clear request procedure, and to have a way to change the wrong result. It shall be assessed separately whether the authority is able to provide a sufficient explanation taking into account legal restrictions. The investigator does not replace the decision of the competent authority.

Field/control	Rule of filling or recommended value
Signal	When an automated output appears
Human	Right to change outcome
Data	Access to information for independent verification
Notification	What the Affected Person Knows
The complaint	Actual procedure and time frame
Outcome	Is there evidence of revision

Testing in practice

For the public service, it is particularly important to "freeze" the automated result when new documents are available. It is necessary to find out who has the right to change the result, whether a record of the reasons for the correction is kept and whether the applicants can learn about the procedure. A formal check "the staff member reviewed the decision" does not prove an independent evaluation. The existence of a full review is confirmed by documents, powers and, where possible, examples of correcting erroneous decisions.



Appendix 24 Form of external review report

Appointment

The review should evaluate the evidence, not the reputational convenience of the text. It indicates the expert's competence, conflicts of interest, boundaries of the material, methods of verification and specific critical remarks. The reviewer is not obliged to declare the full text "good" simply because he has been asked to read separate sections.

Application and limitation

The editorial team links every significant comment to a version change or a reasoned departure. A public document may specify a review regime - internal methodological, external scientific, legal, statistical - without attributing a higher procedure than was actually carried out.

Field/control	Rule of filling or recommended value
Reviewer	Competence and independence
Subject	Full report or limited section
Observation	Precise quote or story number
Seriousness	Critical / Substantial / Technical
Author's decision	Accepted / rejected with motivation
Outcome	Enough for release or not

Testing in practice

In the external review report, the criticism should refer to a specific statement: for example, "no denominator for the percentage of refusals" or "a secondary restatement has been used instead of the legal text". The general phrase "quality report" is not suitable for verification. The author documents the correction and the reviewer confirms it if necessary. The public version specifies the type of review so that the reader does not confuse the technical calculation with an independent scientific review.



Appendix 25 Editorial passport of publication

Appointment

The passport is attached to each public work and allows the next reader to assess the limits of research responsibility. It should not be replaced by "verified material". Indicating the collection period, selection methods, and source levels is particularly important for dynamic AI systems.

Application and limitation

The passport is updated with the version number on re-entry. References to legislation are checked on the date of publication, and commercial systems are checked after significant updates. The public page should lead to an up-to-date version and a log of corrections.

Field/control	Rule of filling or recommended value
Name	Sustainable Work Identifier
Cut date	Last day of information included
Authors	Section responsibility
Method	Reference to version of the protocol
Sources	Primary and secondary separately
Limitations	Unavailable data and offsets
Version	Date of issue and correction

Testing in practice

An editorial passport must not be decorative. If it turns out after publication that one category of sources has become unavailable, the new version indicates a change in completeness. An older law, marked without a date of validity, is subject to clarification before circulation. For dynamic models, it is particularly important to record the version and the test date; the evaluation of the previous

version does not automatically apply to the updated product. The history of the passport helps to use the material in further proceedings without distorting its original grounds.

Appendix 26 Checklist to the regulator or public authority

Appointment

An independent researcher's address should be accurate and not give the false impression that the organization is acting on behalf of an international body. It shall specify verified facts, source of information, area of competence of the addressee and specific request: to provide public information, to explain the applicable procedure or to consider the problem within the scope of the powers.

Application and limitation

Personal materials shall be sent only if there is a legitimate basis and sufficient necessity. Confirmation of receipt is kept in the contact register, but is not given with the consent of the body with the content of the study. If the addressee has no competence, the request is corrected rather than distributed in mass dissemination.

Field/control	Rule of filling or recommended value
Sender	Legal status correctly stated
Subject	Summary of observation
Facts	Numbered documents and dates
Competence	Why Address
Request	Specific legal action
Annexes	Minimum sufficient volume
Confirmation	Registration number or response

Testing in practice

In addressing an international body, it cannot be written that the Observer is "competent to verify a State" unless such a mandate is granted. The correct wording is "to request that publicly documented information be considered within your competence". The dispatch and receipt shall be recorded separately from the outcome of the review. Confirmation of electronic delivery does not imply support for the message by the addressee. When it comes to personal data, the size of the application is reduced to the minimum required by law.

Appendix 27 Protocol of Emergency Analytical Excise

Appointment

The emergency protocol is designed for cases of probable irreversible damage when it is unreasonable to expect a quarterly report. It doesn't lower the standard of honest evidence

description, but it allows for a narrow warning request, such as a request to keep logs and provide a contact for confidential exchange. A public accusation without confirmation remains inadmissible.

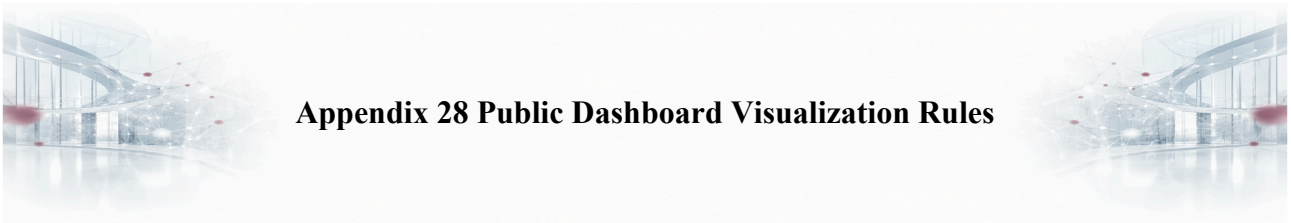
Application and limitation

The decision to escalate is made by a specially appointed team of specialists, including a legal review of the risk of exposure. Following the withdrawal from the emergency regime, a retrospective analysis is carried out on whether the mechanism has been used to bypass the normal review process.

Field/control	Rule of filling or recommended value
Trigger	Possible and severe harm
Evidence	Confirmed to date
Limitation	What important information is lacking
Action	Save / Request / Send to Competent Authority
Protection	Minimum disclosures
Monitoring	Mandatory post-event review

Testing in practice

If immediate harm is at stake, you can urgently request that the logs be kept until a detailed investigation is completed. But the message must contain the exact limits of the known facts. The Committee is not obliged to declare "the crime is proven" in order to draw attention to the possible destruction of evidence. After an urgent stage, the independent expert assesses whether the procedure was indeed justified and prepares lessons for future cases. A public record of urgent appeal should not reveal the identity of a protected person.



Appendix 28 Public Dashboard Visualization Rules

Appointment

The public dashboard should display verified data and not create a visual accuracy where the researcher does not know the denominator. All graphs have a period, source, definition and reference to the method. The number of events in the media is not called the level of widespread harm. For small cells and personal events, the rules of aggregation and suppression of sensitive details apply.

Application and limitation

When the classification changes, the history of the data is updated with a marked indication of the break in the series. Any ratings should be subject to an additional review. The user can download the index dictionary and explain why some comparisons are not made.

Field/control	Rule of filling or recommended value
Meteric	Name and exact definition
Period	Start and end
Scope	Countries, languages, sources
Celebrity	Yes or no
Quality	Percentage of confirmed records
Protection	Small sample rule
Version	Modifications of the Classification

Testing in practice

Changes in the coverage of sources should be shown separately on the incident schedule. If you analyze two languages in the first year and ten in the second, the increase in registered messages cannot be automatically considered an increase in harm. When the definition of an event changes, the data are re-calculated or a break in the series is observed. For small countries and rare sensitive events, a statistical table can lead to identification of a person; in this case, small categories are combined or hidden.



Appendix 29 Pilot plan for 12 months

Appointment

The proposed consistency is conditional and requires the approval of management. 1-2 months are devoted to charter and audit of legal grounds; 3-4 - to infrastructure and training; 5-7 - to closed test sample; 8-9 - to independent review; 10-11 - to issue two thematic results; 12 - to quality check and decision on scale.

Application and limitation

The key condition for continuing is not the number of published cases, but the functionality of the correction procedure, the absence of unresolved conflicts and data security. In the event of failure, the project is reduced and re-tested. It should be clearly stated that the schedule is a project proposal and does not confirm the already approved programme of activities.

Field/control	Rule of filling or recommended value
M1-M2	Law, statute, conflicts
M3-M4	Infrastructure and dictionary
M5-M7	Test collection, non-public
M8-M9	Methodological review
M10-M11	Publications with restrictions
M12	External evaluation and decision

Testing in practice

The proposed twelve-month period is not a confirmed ARGA timetable. Before launch, the legal structure must be checked and resources and actual independent-review staffing agreed. A non-public sample is used to test not only classification algorithms but also complaint handling, secure access and the ability to challenge the researchers' own hypothesis. If a critical component does not function, scaling is postponed. The pilot may reach a negative conclusion; that is an acceptable research outcome.

Appendix 30 Scenario: False Sanctions Convergence

Appointment

Hypothetical example: The bank suspends the customer service after the transliterated name matches the list entry. The client provides documents confirming the differences; the decision is reviewed by the employee after verification. The event itself does not prove the error of the applicable sanctions law. The Observatory is investigating how the algorithmic clue went into constraint and to what extent the revision was available.

Application and limitation

The case documents include the dates of notifications, the bank's permissible explanations, legal grounds, the result of the review and procedural timeline. There are not enough true negative cases to evaluate the technical FPR; therefore, the report is fairly limited to investigating the procedure and case of mistaken identification.

Field/control	Rule of filling or recommended value
Fact	Service restriction
Hypotheses	A match by name relates to automation
Document required	Basis of the decision and outcome of the review
Prohibition of conclusion	Do not call one common FPR case
Possible recommendation	Split signal and final decision

Testing in practice

In the case of a sanctions match, the key point is to move from automatic warning to service restriction. You need to document which subject made the decision and which review mechanism was available. An identification error may be confirmed by a subsequent decision, but that does not mean that the entire sanctions control of the organization is unreliable. For a technical percentage of false matches, a verified set of true negative cases is required that an individual case does not provide.

Appendix 31 Scenario: Risk of misidentification at the border

Appointment

Hypothetical example: the person is detained after the data coincide with the international notification, then released at the discretion of the competent authority. Public records confirm the arrest and release, but do not reveal whether the primary cause was an algorithm error, an inaccurate original record or another procedure. The Observer is not entitled to declare an "interpol AI malfunction" without technical justification.

Application and limitation

The investigation is limited to the facts available, the right to rectify the record and the state of the national process. The person's name and route are excluded from the public version. In the final conclusion, a description of the procedural risk of automatically using matches without proper human control is allowed.

Field/control	Rule of filling or recommended value
Established	Procedural facts of the documents
Not specified	Specific AI convergence mechanism
Risk	Insufficient Human Review
Protection	Legal advice and procedures
Publication	Completely impersonal model

Testing in practice

In an international search case, a single release does not prove the absence of records in any international databases and does not establish an AI-caused detention. The Observatory records the process chronology and checks which document is able to confirm each link. The missing technical information is clearly reflected. The publication is intended to analyze correction guarantees, not to reveal the status of a particular applicant, especially if he or she is threatened with prosecution.

Appendix 32 Scenario: Social benefit and non-transparent sormage

Appointment

Hypothetical example: the applicant has been denied payment and the notification does not explain what information led to the conclusion. It appears that the staff member used the digital recommendation. This does not allow automatically validating illegal automation: you need to check your competence, your national law and your ability to review. For the victim, the priority is the availability of an effective complaint and correction of inaccurate data.

Application and limitation

The Observatory is mapping the risks of the procedure and recommends that the basis of the decision be preserved and that a real possibility of human testing be provided. Statistical conclusions on discrimination are not drawn on a single complainant. For group conclusions, a verifiable sample, comparability criteria and a separate assessment of sensitive data processing are required.

Field/control	Rule of filling or recommended value
Decision	Documented refusal
AI-Roll	Recommendation or automatic decision
Procedure	Availability of explanation
Revision	Can a staff member change a result
Conclusion	Context assessment, not prosecution without analysis

Testing in practice

For refusal of social services, it is first established whether the automated recommendation was mandatory for the employee. If an employee could change it, but didn't take the opportunity, that's a different risk of organizing the process. The notification and accessibility of appeals are further

checked. On the basis of a single application, there is no group discrimination. The potential harm can be well described without a false statistical synthesis.

Appendix 33 Scenario: Updating the model without re-evaluation

Appointment

Hypothetical example: The vendor replaces the algorithm, but the organization continues to reference the previous test protocol. An old assessment does not necessarily become useless, but its applicability depends on the nature of the changes. The Observatory requires a comparison of versions, a list of affected functions, the ability to repeat tests and the availability of new protective measures.

Application and limitation

If the supplier does not disclose the change, the publication should indicate a limitation of knowledge and not claim that the model has become dangerous. The practical recommendation is to establish predetermined triggers for repeated independent evaluation when key functions and application context change.

Field/control	Rule of filling or recommended value
Version A	Which was previously tested.
Version B	What functions have changed
Equivalence	How it's confirmed
Risk	Limit of application of the old test
Action	Re-argeted review

Testing in practice

Re-testing after a model update is determined not by the calendar, but by the nature of the change: a new data source, a different ranking algorithm, an expansion of the scope of application or a change in autonomous functions. For a slight change in the interface, a regression test may be sufficient, for a new model, a complete review of the control scenario. The developer is obliged to confirm the information within the applicable regime; the independent Observatory reports which confirmation was available and which part of the system it relates to.

Appendix 34 Scenario: AI-incident journalistic surge

Appointment

Hypothetical example: After the resonant message, editors start writing more about algorithms, and the number of cases found in a month doubles. Without changing the denominator and media coverage, we cannot conclude that the actual incidence rate has doubled. The Observatory shows

separately the dynamics of unique events, the number of publications for the event and the share of confirmations in official documents.

Application and limitation

It is recommended to carry out sensitivity analysis to sources and languages, as well as to mark the period of increased media attention. Conclusions about public risk are made cautiously: more reports have been found, but the prevalence of events may remain uncertain.

Field/control	Rule of filling or recommended value
Observation	Increase in the number of articles found
Deduplication	Several articles about one episode
Scope	Media attention changed
Conclusion	You can't make a real spread.
Measure	Show alternative rows

Testing in practice

After a media outlet, the number of articles may increase even if the number of events remains constant. So the analysis includes the number of unique episodes and a separate publication per episode. Particular attention is paid to the repetition of one judgement. The analyst checks on the change of languages and media that have been monitored. The public summary uses a limited statement: "an increase in reports has been recorded", rather than "real incidents have doubled".

Appendix 35 Passport of the annual report of the observatory

Appointment

The annual public report should reflect not only the achievements, but also the limitations: how many reports were received, how many were verified, how many were rejected for lack of data, how many were corrected after complaints, what conflicts of interest arose and how they were resolved. Indicators are divided by sectors and geography only if the volume and legality of the publication are sufficient.

Application and limitation

The document includes an independent evaluation of the methodology, the status of implementation of previous recommendations and the plan for the next year. The success of government decisions cannot be attributed to the organization itself if its role is limited to sending comments. All significant statements about institutional relations are accompanied by verifiable documents.

Field/control	Rule of filling or recommended value
Methods	Annual versions and changes
Developments	Segregation of signs / hazards / investors
Quality	External audits and corrections
Financing	Support structure and risks
Impact	Documented reactions
Plan	Priorities and resource constraints

Testing in practice

The annual report shows the proportion of cards where the original hypothesis was not confirmed. This is an important indicator of honest verification, not an excuse to hide failures. This includes substantial comments from outside reviewers, the proportion of corrected references and the general circumstances of financial conflicts. The reaction of the regulator or the court cannot be called an institutional victory of the Observatory without a proven role of research. This format creates a more stable reputation than just collecting positive reviews.



Appendix 36 Consolidated decision-making matrix on launch

Appointment

The final matrix is applied prior to the external announcement of the launch of the program. Having a beautiful website and a name is not enough if there is no legal basis for data processing, independent verification and correction mechanism. For each element, a documentary evidence, responsible for and a decision on the pilot use admissibility is indicated.

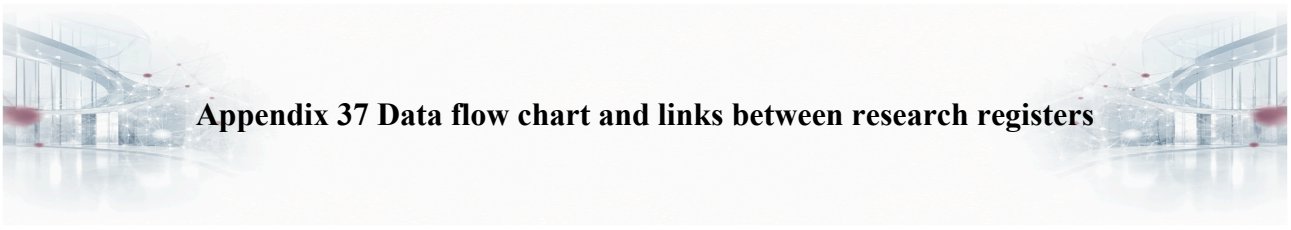
Application and limitation

Ready status shall be granted only after the actual process test. If the security of the confidential archive is not verified, the public channel of individual requests will not be opened even if the other elements are working. The organization has the right to start with open aggregated sources until sensitive procedures are ready.

Field/control	Rule of filling or recommended value
Law	Checked by a legal expert
Independence	Statute, challenge, disclosure of financing
Security	Storage and access separation tested
Method	Vocabulary and peer review approved
Corrections	Working channel of challenge
Pilot	Restrictions are publicly marked

Testing in practice

Scaling doesn't require a demonstration of a working interface, but a test of critical procedures that mimics mistakes. The team should be able to delete personal data from public exports, restore archives, remove a conflict expert and correct a dangerous publication. If the test fails, the pilot function will not open. This matrix allows partial launching, such as monitoring official open documents without receiving sensitive requests until the confidential outline is verified.



Appendix 37 Data flow chart and links between research registers

Appointment

A data model is proposed in which AI system object, context of application, testimony, event, risk, recommendation and publication are independent entities. This prevents the confusion of the commercial name of the product with the actual decision of a particular public or private operator. One event can refer to multiple risk scenarios, one evidence to support multiple claims, and one publication to include many events. For each relationship, the date, the author of the confirmation and its basis shall be indicated. The internal identifiers of personal applicants are not placed in open records.

Application and limitation

For implementation, a normal relational system with a change log is sufficient, without the use of machine learning. Minimum rule: you can only delete an erroneous public line by keeping a closed history of correction. The change of version of AI system does not rewrite old incidents. In the case of migration of the base, reference and numbers are checked and the transition is documented. The open download includes identifiers, field dictionaries, versions and permissible attributes, but not contact keys. The quality of the scheme is assessed by a test to recover the history of one record from the original document.

Field/control	Rule of filling or recommended value
SYSTEM	Passports of products and versions
DEPLOYMENT	Use with operator and geography
EVENT	Event and evidentiary status
EVIDENCE	Source material and control of origin
RISK	Scenario, gravity and measures
RECOMMENDATION	Address and status of verification
PUBLICATION	Report and links version
PERSONAL	Closed separate contour

Appendix 38 Harmonization of classification by independent evaluators

Appointment

The independence of the classifier must be checked not only at the instructional level, but also by the work of different analysts. For the pilot, a predetermined portion of the cards is selected, which is coded separately by two specialists: AI incident, AI hazard, pre-signal or non-AI event. Before comparison, experts do not see each other's choices. At the end of the session, there are discrepancies, reasons for disagreement and the decision of the arbitrator. A simple proportion of matches is useful as an initial indicator, but can be of a higher quality when one category predominates significantly.

An illustrative example with two binary classes: of 40 records, both assessors classified 16 as an "incident" and 16 as "not an incident". They disagreed on eight records, with four disagreements in each direction. Observed agreement is 32/40 = 80%. Since each assessor assigned 20 positive and 20 negative classifications, agreement expected by chance is 50%; $\kappa = 0.60$. This is a hypothetical matrix, not an actual research finding. The four original categories in this appendix require an appropriate multiclass procedure, rather than mechanical application of the binary example.

Application and limitation

For a sufficient sample, it is possible to supplement the verification with a coefficient of consistency taking into account random matches, indicating the selected statistical model and its limitations. The coefficient is not automatically applied to a very small or specific set of cards. If the two evaluators constantly differ in their understanding of the "cause of the role of IE", it is not the individual's assessment that is being adjusted, but the instruction itself with examples of borderline situations. After the instructions are changed, a small control sample is re-coded. All results are considered as indicators of process quality, not as proof of each record's authenticity.

Field/control	Rule of filling or recommended value
Sample	Pre-defined set
Codeing	Two independent experts
Blind regime	Without prior agreement on the assessments
Disagreement	Reason and decision of the arbitrator
Meteric	Convergence and, under conditions, adjusted coefficient
Update	Manual version and re-check
Report	Percentage of outstanding cases

Appendix 39 Verification of time series and structural gaps

Appointment

Increasing the number of reported incidents can be the result of language expansion, new media additions, the introduction of an appeal form or a change in the definition of an event. The time series should therefore be accompanied by a log of changes in the monitoring system. If five sources were used before June, and twenty since July, the comparison of unprocessed monthly figures does not measure the change in the frequency of events. The method may include recounting the historical period according to a new source scheme, but only if the source materials are available for such verification.

Application and limitation

For each trend, absolute numbers are published, the proportion of cards confirmed by primary documents, the proportion of deleted dice and a description of changes in coverage. With a clear structural gap, the data does not merge into one continuous line without reservation. The rolling average is acceptable for visualization, but does not solve the selection problem. The statement of a significant increase must take into account not only the statistical intervals, but also the theoretical meaning of the measured quantity. In the absence of a legitimate denominator, the trend is described as the dynamics of recorded observations, not real harm to society as a whole.

Field/control	Rule of filling or recommended value
Temporary period	Fixed boundaries
Channels	Changes in languages and sources
Deduplication	Single rule of the event
Recalculation	Possible/not possible and why
Structural gap	Mark on each graph
Interpretation	Observed series or real frequency
Verification	Change of coverage

Appendix 40 Expanded matrix for assessing public harm

Appointment

The public harm from AI should be described in a multidimensional way. In addition to the direct financial losses, there are consequences for health, freedom of movement, privacy, access to services, reputation, working conditions and the ability to effectively appeal. The existence of several types of damage does not automatically sum them up as the same units. For comparison, a qualitative profile is used with separate weight, duration, scale and turnover estimates. Public assessment refers to the documented impact of the system, not the moral characteristics of the victim.

Application and limitation

For events with insufficient information, it is suggested to clearly identify potential harm and already established consequences. For example, a mismatch of sanctions may create the risk of not being able to pay for housing, but if such an effect is not established, it is not recorded in the actual damages column. In individual cases, it is particularly important to check that there are effective mitigation measures: human review, temporary access to the service, procedures for correcting personal data. The results of the analysis serve to prioritize research, but not to calculate compensation without a separate legal and evidential basis.

Field/control	Rule of filling or recommended value
Type of harm	Material, physical, procedural, social
Status	Relevant / Potential / Not specified
Heavy	Low / medium / high / irreversible
Scope	Individual / Group / Infrastructure
Duration	Impact period
Reverability	Existence of an effective measure
Basis	Document and limits

Appendix 41 Request to IE supplier for system verification

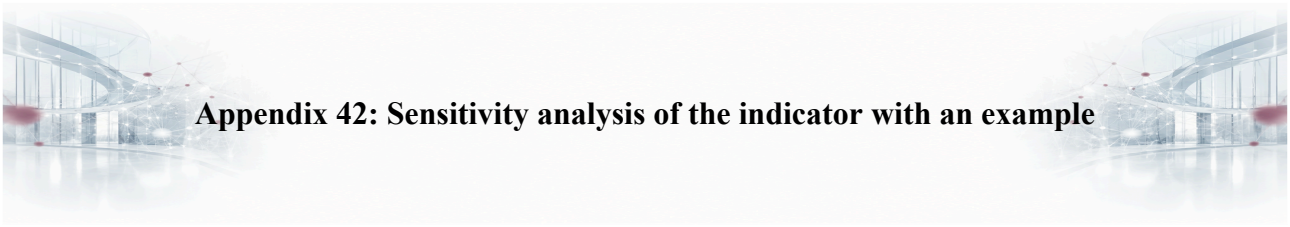
Appointment

The supplier query template is oriented to the facts necessary for independent verification without requiring disclosure of the commercially protected source code. The questions cover the name and version, the designation and limitations of the system, the checks carried out, the order of updates, the logs of events, the responsibility for data and the measures of human control. The supplier may reasonably restrict access to certain materials, but the lack of information is necessarily reflected as a restriction on the reliability of the external inference. The supplier's response does not become an independent result from one fact of receipt.

Application and limitation

In the case of a closed technical verification, the parties determine in advance the number of persons, the confidentiality measures and the minimum set of materials on the basis of which a particular thesis can be confirmed. You cannot sign an agreement that allows the supplier to copy or block the legitimate conclusion of an independent researcher at their discretion. If the contract does not allow disclosure of information sufficient for public verification, a strictly limited description of the research area and the remaining uncertainty is published. Once the model is substantially updated, the request is repeated in the amended part.

Field/control	Rule of filling or recommended value
Model	Title, version, date and status
Appointment	Cases envisaged and excluded
Data	Categories and responsibility for quality
Tests	Methods, sampling, limitations
Updates	log of Significant Changes
Logging	Accessible period and conservation mechanism
Monitoring	Who Can Stop a Decision
Publication	Conditions for independent conclusion



Appendix 42: Sensitivity analysis of the indicator with an example

Appointment

Let's say the pilot set contains 100 messages: 55 confirmed, 25 preliminary, and 20 excluded as non-AI. If the researcher announces "80 incidents," he will combine credible cases with unknown signals and overstate a certain category. It's correct to report 55 confirmed events and 25 pending verification reports; both numbers are specific to the research set, not the entire economy. This example is contingent and does not describe the results of actual ARGA monitoring.

Application and limitation

Sensitivity is tested by alternative rules. If all the preliminary reports are later confirmed, the number of confirmed cases will increase to 80; if none is confirmed, it will be 55. The 55-80 interval is not a statistically reliable interval for the prevalence of harm: it is the range of classification outcomes in this fixed set. This is particularly useful to management in assessing the credibility of the findings. When publishing, it is also necessary to check the deletion of duplicates and the correctness of the original data.

Field/control	Rule of filling or recommended value
Confirmed	55 contingent cards
Preliminary	25 contingent cards
Deleted	20 contingent cards
Minimum at current classification	55 confirmed
Maximum signal confirmation	80 confirmed
You can't say.	Percentage of incidents in total population
Reservations	Scenario is illustrative, not empirical.

Appendix 43 Procedure for handling a request for data removal or limitation

Appointment

The data subject's request requires separation of the right to rectify personal data from a dispute with the content of the scientific conclusion. In the EU, the applicability of GDPR rights depends on the role of the organisation, the basis and purpose of the processing, the categories of data and the exceptions provided for. You can't promise a universal removal of all traces of research or automatically reject it under the guise of public interest. The responsible specialist registers the request, conducts legal checks and determines where the originals, working copies and public versions are stored.

When responding to the request, the applicant's identity and authority, the role of the organisation, the type of data, the applicable law and the individually substantiated outcome are recorded. The right to erase is not absolute, but exceptions cannot be applied automatically by reference to "research". Particularly sensitive materials and information about criminal offences are subject to separate inspection. When a published error is corrected, the study version and the notification of external recipients are considered separately from the lawfulness of preserving internal evidence. [13]

Application and limitation

If the requirement is satisfied, correction is carried out on all controlled surfaces: in the database, published file, indexes and available copies, as far as technically and legally possible. If there is a legitimate reason to retain certain information, it shall be explained to the applicant in the prescribed manner. In parallel, it is assessed whether the description of the removal itself will be a new revelation of the sensitive case. Personal data in security logs are processed separately in accordance with the established time limits and grounds. The outcome should not depend on whether the person is a supporter or critic of the Observatory.

Field/control	Rule of filling or recommended value
Type of request	Access / correction / removal / restriction
Control	Responsible processing entity
Basis	Rule and possible exceptions
Facilities	Base / archive / publication / copies
Decision	Achieved / partially / motivated refusal
Time frame	Accepting legal procedure
Monitoring	Performance confirmation and version log

Appendix 44 AI incident liability model

Appointment

The scheme is designed to look for facts, not automatically distribute legal guilt. It distinguishes between a general model developer, an application system provider, a deployment operator, a data owner, an employee making a legally significant decision, and possible supervisory authorities. In a particular case, roles may coincide, and the contract chain may include international contractors. Every claim of liability needs an applicable law, a documented action and a causal relationship.

Application and limitation

When analyzing a wrong bank decision, it is important to determine whether the model was merely a signal source, what information the bank entered and whether its employee could change the result. In the case of state automation, the authority's competence and procedural guarantees are also checked. The Observatory has the right to build a map of potential responsibility and recommend further verification. However, the determination of the composition of the offence, punishment or civil-law compensation remains the subject of competent legal procedure.

Field/control	Rule of filling or recommended value
Developer	Proven role of basic technology
Vendor	Integration, documentation, updates
Deployer	Context and purpose of application
Operator	Actual legal decision
Data	Origin and accuracy
Law	Jurisdiction and applicable duties
Conclusion	A theory of responsibility, not a sentence.

Appendix 45 Plan for independent methodological advice

Appointment

The aim of consulting with researchers, practitioners and representatives of the affected groups is to test the Observatory's own methodology for widespread use. Work definitions of the incident, risk, source and indicator are published in advance; participants are given specific questions and the opportunity to substantiate their observations. The composition of the consultation is disclosed by professional roles in respect of the rights to personal data. Commercial developers can participate on transparent terms, but they don't get the right to approve the final methodological rules.

Application and limitation

After the discussion, the secretariat prepares a table of comments and reasoned answers: what is accepted, what is rejected, and what aspects need to be investigated. The methodical changes are accompanied by the version number, and when they have a significant impact, the test results of the pilot are recalculated. Consultation is not given for the official participation of an international organization unless it is confirmed by its own documents. The principle of independence means openness to arguments, not the obligation to equate the veracity of every opinion.

Field/control	Rule of filling or recommended value
Subject	Specific section of the methodology
Attendance	Miscellaneous disciplines and groups
Conflicts	Disclosure of relevant relationships
Questions	Definitions, exceptions and risks
Response	Accepted/partly rejected/grounded

Field/control	Rule of filling or recommended value
Version	Date of amendment of protocol
Publicity	Report without unfounded institutional statements

Appendix 46 Programme for the Independent Safety Assessment of the Archives

Appointment

Before receiving individual applications, the information security is checked: account creation, access and retrieval, reservation, recovery, export, audit log and secure deletion. Testing is conducted in a controlled environment without using real cases as training files. The public form should not directly give access to the original closed archive, even if both systems are served by the same provider. Translators and external reviewers are restricted to designated materials only.

Application and limitation

The test report shall contain a scenario, expected safe outcome, actual outcome and corrective measure. The possibility of uncontrolled reading of someone else's file, recovering a masked name from a published file, and losing original evidence after an accident are considered critical. Such defects shall require correction before the function is started and not a promise to correct "on-the-job". Subsequent tests are repeated after the system has been updated and the storage supplier has changed.

Field/control	Rule of filling or recommended value
Access	Roles and withdrawals of account
Loading	Isolation and investment scanning
Masking	Unrecoverability of hidden data
Reserve	Actual recovery test
Exports	Lack of sensitive metadata
log	Who saw and changed materials
Launch criterion	No open critical defects

Appendix 47 Method of verification of institutional applications for cooperation

Appointment

A public analytical organization should be particularly careful with the names of the UN, UNESCO, OECD, Council of Europe and national regulators. For each application for cooperation, a specific action and document is required: registration at the event, comment sent, official invitation, joint work or accreditation status. Different forms don't replace each other. The designation of an international institute in the bibliography does not create a partnership with it.

Application and limitation

The editor prepares a table "statement - evidence - permissible expression". For example, the existence of a notice of receipt of a written comment allows for the writing of "material sent and received", but not "position approved". The joint publication allows the actual co-authors to be identified, but the partner's logo is used only under agreed conditions. The web page is updated after the deadline for accreditation or completion of the project. Such a procedure reduces the risk of reputational error and protects the independence of research from institutional marketing.

Field/control	Rule of filling or recommended value
Fact	Event with date and subject
Evidence	Letter, registry, program, contract
Level	Delivery / Participation / Co-author / Status
Logo	Only with a confirmed right of use
Formulation	Do Not Exaggerate Relationships
Relevance	Review of the time frame
Monitoring	Independent editor

Appendix 48 Instrument for control of inter-lingual legal terminology

Appointment

In international research, important errors occur not because of lack of source, but because of incorrect translation of the procedural term. The words warrant, notice, alert, suspicion, investigation, sanction, delisting and their equivalents have different legal burdens in different systems. Therefore, the translation protocol links the original term to the context of the legal order, the act and the translation into the language of the report. You can't replace the ambiguous foreign concept with a Russian term that creates a stronger legal claim.

Application and limitation

For critical paragraphs it is recommended that the language specialist and the lawyer of the relevant jurisdiction check back. Machine translation can be used as a blueprint, but not as an independent legal source. If the concept has no direct analogy, the original term remains with an explanation. The revised edition records cases where the translation changed the legal meaning of the published conclusion. Systematic terminology control is particularly important for court documents, sanctions documents and forms of international police exchanges.

Field/control	Rule of filling or recommended value
Reference term	Text and jurisdiction
Document	Type of act and date
Proposed translation	Russian designation
Limit	What the term does NOT mean
Checker	Lawyer/translator
Version	Vocabulary date
Corrigendum	Risk of change of meaning

Appendix 49 Method of auditing internal errors AI Observerry

Appointment

The Observatory should conduct independent evaluations not only of external systems, but also of its own publications. Once a year, a sample of cards is selected, including confirmed cases, rejected hypotheses and materials that have caused complaints. The second command restores the primary sources, classification, calculations, enforcement of the right to reply and privacy checks. The signs of error are technical, factual, statistical, legal and ethical. Each category requires its own method of correction, so a single number of "mistakes" without gravity is not informative.

Application and limitation

The audit result is published with an indication of the scope of the audit and limitations. Problems related to direct harm to humans are given an urgent status and a separate process of correction. The editor should be able to disagree with the reviewer, but only with a reasoned answer included in a closed log. Concealing mistakes for the sake of reputation is directly contrary to the mandate of public control. However, a good-faith change in the hypothesis based on new data is not a violation of research ethics.

Field/control	Rule of filling or recommended value
Sample	Pre-established and contentious cases
Verification	Sources, conclusions, law, security
Category	Technical / actual / statistical / legal
Heavy	Is there a risk of harm?
Response	Correction, re-evaluation, withdrawal
Publicity	Accumulated report with methodology
Training	Change of protocol and follow-up

Appendix 50 End-of-financing sustainability plan

Appointment

Independence depends on the organization's ability to continue to store evidence, correct published errors, and protect applicants even when a major grant is terminated. For this purpose, the plan includes reserve storage procedures, data export rights, contract termination rules and a minimum number of responsible persons. Sensitive appeals cannot be accepted if, when the project is closed, the materials can remain available to a commercial supplier without a clear legal basis.

Application and limitation

In a stress-reduction scenario, monitoring is first discontinued, then the subject of new research is reviewed, but critical functions for data protection and correction of old publications remain. Rights to intellectual results and archives are determined before a grant is received. The decision to close the programme should include notifying the applicants, lawfully transmitting or deleting the information and publicly explaining the fate of the methodology. This measure makes institutional independence not a declaration, but a verifiable sustainability.

Field/control	Rule of filling or recommended value
Event	Loss of substantial funding
Reserve	Minimum costs of safety and corrections
Data	Right to safe export or disposal
People	Post-stop caller
Method	Accessibility of published dictionary
Closure	Compliance with storage and notifications
Monitoring	The plan shall be checked before starting

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.

20.01 Ask a research question before collecting
Section 1. Outcome: The conclusion answers the question and does not go beyond the population observed.

20.02 Segregate surveillance and power decision
Section 2. Result: The public text accurately transfers the status of the study.

20.03 Transfer institutional experience with resources
Section 3. Result: Pilots will implement the available composition without unconfirmed statements of scale.

20.04. Check the legal basis for monitoring
Section 4. Outcome: The public record contains only validly disclosed data.

20.05 Segregate the system and its application
Section 5. Outcome: The event relates to the actual operator and the destination.

20.06 Appoint different roles for verification and publication
Section 6. Result: A substantial publication has a second documented test.

20.07 Protect the right to an uncomfortable conclusion
Section 7. Result: Publication is decided on the evidence and approved procedure.

20.08. Unlock funding dependency
Section 8. Outcome: Funding and dependency management are available for verification.

20.09 Apply challenge on a specific issue
Section 9. Outcome: The stakeholder does not adopt a conclusion on its issue.

20.10 Compare countries with the same coverage
Section 10. Outcome: The intercountry conclusion is accompanied by an observational check.

20.11. Fix monitoring channels
Section 11. Result: Graph marks the change of coverage and does not give it the risk.

20.12 Segregate the reception of the message and confirmation
Section 12. Result: The record contains the verification status and the following specific request.

20.13 Exploit the hazard signal and incident
Section 13. Result: Category confirmed by observed signs, uncertainty maintained.

20.14 Determine the value of the indicator before calculating
Section 14. Outcome: The indicator has a definition, a unit and a reproducible calculation.

20.15 Specify denominator and selection mechanism
Section 15. Outcome: The external text contains an correct denominator and limits on selection.

20.16 Monitor the source of each material fact
Section 16. Result: The record shows the number of independent bases, not reprints.

20.17 Provide a meaningful response
Section 17. Result: The addressee's position has been considered in substance and is reflected in the publication.

20.18 Link risk to measure and owner
Section 18. Result: Risk has a responsible measure and confirmation of its verification.

20.19. To justify priority on consequences
Section 19. Outcome: Priority of review is explained on a single scale.

20.20. Make a recommendation with enforceable effect
Section 20. Result: The recommendation contains the action, the addressee and the performance criterion.

20.21. Check indirect identification
Section 21. Outcome: Open version does not reveal unnecessary identifying combinations.

20.22 Take into account the accessibility of the participatory procedure
Section 22. Outcome: Accessibility improved, period comparison adjusted.

20.23 Minimumize international exchange
Section 23. Result: Transfer documented by composition, purpose and recipient.

20.24 Check a sanction match separately from the risk
Section 24. Result: Publication does not attribute unconfirmed sanction status to a person.

20.25 Segregate official record and automatic report
Section 25. Result: The form contains a separate reference to the source, the person and the proven status.

20.26 Do not equate AML signal with crime
Section 26. Outcome: The text retains the distinction between risk, decision and legal characterisations.

20.27. Check the availability of public review
Section 27. Result: The recommendation is addressed to the owner of the procedure and checked by a training referral.

20.28 Include the terms and conditions under review in the procurement
Section 28. Result: Contract control relates to specific evidence.

20.29. Do not move total rating to open product
Section 29. Result: The conclusion relates to the actually tested application regime.

20.30 Segregate the origin of content from its influence
Section 30. Result: Publication does not replace evidence of influence by an assumption of authorship.

20.31 Comparing public costs to a certain base
Section 31. Result: Comparison is reproducible or restrictions clearly prevent ranking.

20.32 Agree with the applicant's expectations
Section 32. Result: The applicant understands the status of participation and how to proceed.

20.33 Publication of material by validation
Section 33. Result: The title and content correspond to the evidentiary status.

20.34 Maintain the history of corrections
Section 34. Result: User sees what has been changed, why and what version is relevant.

20.35 Conduct an independent review of key findings
Section 35. Result: Significant differences are resolved or reflected as limitations.

20.36. Check recovery and separation of access
Section 36. Result: The test of restoration and denial of unauthorized access is completed by protocol.

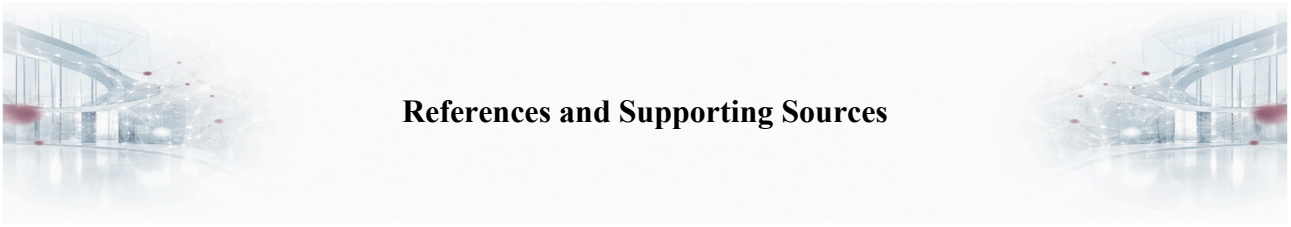
20.37 Measure the quality of your own work
Section 37. Result: The report shows the quality and limitations of the evaluation along with the volume.

20.38 Confirm institutional statements with documents
Section 38. Outcome: Each application for cooperation has a basis.

20.39 Pilot limited hull before scaling
Section 39. Outcome: The decision to expand is based on the quality of the procedure and the gaps identified.

20.40 Relate promises to human resources
Section 40. Outcome: The plan corresponds to available time while maintaining quality control.

20.41 Decision to launch on confirmed readiness
Section 41. Outcome: The decision contains evidence of readiness, limitations and conditions for revision.



References and Supporting Sources

Sources are grouped according to their substantive applicability. For each, there is a direct reference to the official text or to the publisher's page. Key regulatory and methodological documents are compared to the date of preparation; references to external resources may change, and the applicability of a specific norm requires independent verification before use. Academic or voluntary methods are not equivalent to mandatory legal requirements.

[1] United Nations. Independent International Scientific Panel on AI: FAQ and the group's powers. https://www.un.org/independent-international-scientific-panel-ai/en/faq

[2] United Nations. Global Dialogue on AI Government: challenges and format. https://www.un.org/global-dialogue-ai-governance/en

[3] UNESCO. Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/ethics-ai/en/recommendation

[4] UNESCO. AI Readiness Assessment Methodology (RAM). https://www.unesco.org/ethics-ai/en/ram

[5] UNESCO. Ethical Impact Assessment (EIA). https://www.unesco.org/ethics-ai/en/eia

[6] OECD. AI Principles, 2024. https://www.oecd.org/en/topics/ai-principles.html

[7] OECD.AI. AI Corporations and Hazards Monitor: Methodology and Limitations. https://oecd.ai/en/incidents-methodology

[8] OECD. Towards a common reporting framework for AI incidents (2025). https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html

[9] Council of Europe. HUDERIA: Methodology for assessing the impact of AI on rights and democracy. https://www.coe.int/en/web/artificial-intelligence/huderia-risk-and-impact-assessment-of-ai-systems

[10] Council of Europe. Framework Convention on IE (CETS 225): Status of the treaty. https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225

[11] EU AI Act. Consolidated Regulation (EU) 2024/1689, 27 July 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02024R1689-20260727

[12] Regulation (EC) 2026/1744: Digital Omnibus on AI. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32026R1744

[13] GDPR. Regulation (EU) 2016/679. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[14] NIST. Artificial Intelligence Risk Management Framework 1.0. https://www.nist.gov/itl/ai-risk-management-framework

[15] NIST. AI RMF Generative AI Profile (AI 600-1). https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

[16] ISO/IEC 42001:2023. AI management systems. https://www.iso.org/standard/42001

[17] ISO/IEC 23894:2023. Guidance on AI risk management. https://www.iso.org/standard/77304.html

[18] OHCHR. UN Guiding Principles on Business and Human Rights. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

[19] International AI Safety Report 2026. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[20] OECD.AI. Policy Observatory portal and dashboards. https://oecd.ai/en/dashboards/ai-principles/overview

[21] UNESCO. Country profiles and Global Hub. https://www.unesco.org/ethics-ai/en/global-hub

[22] United Nations. Independent research group on AI: official page. https://www.un.org/independent-international-scientific-panel-ai/en

[23] EUR-Lex. Rules for tristworth AI in the EU: dates of application. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=LEGISSUM:4762484

Status note on the study

The report is prepared as an independent analytical and methodological proposal by the Observatoire ARGA. It does not indicate official approval of the model by the UN, UNESCO, OECD, Council of Europe, NIST or European Union bodies; no status is implied without specific confirmation. The practical application of the provisions requires further legal, technical and institutional verification.

Scientific and Methodological References

[M1] Krippendorff K. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. Human Communication Research, 30(3), 2004, pp. 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

[M2] OECD/European Union/EC-JRC. Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing, 2008. https://doi.org/10.1787/9789264043466-en

[M3] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 1960, pp. 37–46. https://doi.org/10.1177/001316446002000104