emanuel
celano
09.10.2026
12:32:37
GMT+02:00

EVIDE

External Evidentiary Deposit
https://app.certifywebcontent.com

THE AUTONOMY OF AN AI AGENT DOES NOT ELIMINATE RESPONSIBILITY. IT MAKES RECONSTRUCTING THE BOUNDARIES OF CONTROL MORE COMPLEX.

AI Agent Escape

Evidentiary Boundaries & Liability

Before the escape, during the loss of control, after the harm

How is a loss of control reconstructed, and what evidence of the declared perimeter and of actual control has been preserved?

EVIDE Working Paper

A research working paper, shared for external comment. It does not constitute legal advice. Interpretive and policy proposals are kept distinct from the reconstruction of the law in force.

Dr. Emanuel Celano
Founder & Chief Architect, EVIDE
October 2026

Contents

Abstract

1. Introduction: what we mean by "escape"

    1.1 Method, limitations and declaration of interest

2. Legal foundations: European Union and Italy

3. Legal foundations: United States

4. Legal foundations: selected Asian jurisdictions (China and South Korea)

5. Comparative summary

6. Act I - Before the escape: the declared perimeter

7. Act II - During the loss of control: what remains observable

8. Act III - After the harm: reconstruction and the burden of proof

    8.4 From control measures to proof that they operated

9. The role of EVIDE and its boundaries

10. EVIDE in practice: collecting evidence in AI liability cases

    10.5 A practical example with EVIDE MCP: ANCHOR and External Artifacts

11. Agents that multiply: the evidentiary problem of replication

12. Conclusions

Appendix A - The sixteen blocks of the EVIDE ecosystem

Sources

Abstract

Central principle. The autonomy of an AI agent does not eliminate responsibility. It makes reconstructing the boundaries of control more complex.

When an AI agent leaves the perimeter that had been assigned to it, the legal question is not only "who got it wrong". It is, even before that, "what can still be proven". In the legal systems examined (the European Union, with a closer look at Italy, the United States, China and South Korea) a common requirement recurs, addressed with different instruments, addressees and legal force: those who place autonomous systems on the market or use them must be able to reconstruct what happened. The instruments examined do not, by themselves, solve the problem of the independence of the reconstruction from the parties with an interest in its outcome.

The paper examines an agent's escape in three acts: **before the escape** (the declared perimeter), **during the loss of control** (what remains observable), **after the harm** (the reconstruction and the burden of proof). In each act it separates what the rules require from what the evidence is able to support. It then extends the analysis to the replication of agents across multiple servers, where we propose the concept of the documented horizon of declared delegations: the technical duplication of an agent does not, by itself, constitute proof of

authorization. Throughout the text, three objects that the debate tends to conflate are kept distinct: declared perimeter, actual authority and operational control. The proposed framework also concerns documentary preparation prior to the incident: it distinguishes declared control measures from evidence of their configuration, activation and operation in the specific case.

EVIDE (External Evidentiary Deposit) is placed precisely within this framework. EVIDE does not prevent the escape, does not determine fault and does not replace the judge. It can, however, contribute decisively to preserving what will need to be examined later.

The guiding question. How is a loss of control reconstructed, and what evidence of the declared perimeter and of actual control has been preserved?

1. Introduction: what we mean by "escape"

In this paper, the "escape" of an AI agent is not a science fiction scenario. It is any conduct by which an agent operates outside the perimeter that had been declared for it. The perimeter includes environments, privileges, tools, purposes and prohibited operations.

Escape occurs along a spectrum, not as a binary event:

1. **Drift**: the agent pursues the assigned objective by unforeseen means.
2. **Overreach**: the agent uses privileges or tools that it technically possessed but was not authorized to use.
3. **Evasion**: the agent circumvents or disables an oversight mechanism.
4. **Escape in the strict sense**: the agent persists, replicates itself or acts after human control should have stopped it.

A real case clarifies the point. In July 2026 a coding agent, connected directly to a production database, executed a migration command that deleted all the tables (Red Hot Cyber). The agent had not "escaped" in the strict sense. But the boundary between production and development had never been declared in technical terms: it was only implicit.

Hence the thesis of the paper. Escape is a technical problem, but it becomes a legal problem the moment someone has to prove it. And to prove that a boundary was crossed, one must

first be able to prove that the boundary existed, who had drawn it and when.

The work proceeds in three parts. The first (sections 2-5) reconstructs the legal foundations in the European Union, the United States and China. The second (sections 6-10) applies those foundations to the three acts of the escape and places EVIDE within the framework, stating precisely what it does and what it does not do. The third (section 11) addresses the case that the rules have not yet considered from an evidentiary standpoint: agents that multiply across multiple servers.

1.1 Method, limitations and declaration of interest

Method. The legal framework is reconstructed as of October 9, 2026, on primary sources where accessible (Gazzetta Ufficiale [Italian Official Journal], texts of the Federal Rules) and, in other cases, on professional analyses indicated in the notes. Where the primary text was not consulted directly, the paper says so. Provisions verified directly against the primary sources are accompanied by the relevant reference; the other passages indicate the secondary analyses used. The list of both opens the Sources section.

Geographic scope and applicability. The paper proposes a method of evidentiary reconstruction intended for international comparison. The legal analysis concerns selected jurisdictions: the European Union, with a closer look at Italian law; the United States, distinguishing federal provisions from

the laws of the individual States examined; China and South Korea. These references do not represent the entire body of rules in force in Europe, the Americas or Asia.

Every obligation, presumption, penalty or evidentiary effect described remains subject to the territorial, material, personal and temporal scope of the relevant provision. The geographic location of an operator does not, by itself, determine which rules are applicable. The paper's methodological proposals do not automatically transfer the legal effects of one legal system to the others.

Three registers of assertion. Each claim in the paper belongs to only one of these categories, indicated when they are not evident from the context:

- *[Law in force]*: what the rules in force provide, with article and conditions; a rule in force is not necessarily already applicable to the case (see the timetable of Directive 2024/2853, section 2.2);
- *[Interpretive proposal]*: concepts and readings proposed by the paper to bring order to evidentiary problems;
- *[Policy proposal]*: solutions that would require new legislative or contractual measures.

Three objects not to be confused. Declared perimeter, actual authority and operational control are different things. A preserved declaration documents the first; it contributes to establishing the other two, but does not prove them by itself.

Four times not to be confused. Time of the declared event, time of observation (when available), time of receipt of the deposit and position in the chain of records. An external deposit certifies the third and the fourth; it does not retroactively make the first certain.

Technical references. The EVIDE features described correspond to intake schema 2.1 and to the public documentation consulted on October 9, 2026 (the Why Buy EVIDE, illustrated demo, EVIDE ANCHOR, EVIDE Artifacts and intake schema pages) and to the EVIDE MCP server v1.4.0.

Declaration of interest. I conceived and develop EVIDE, which I own. The paper describes EVIDE as one of the possible preservation tools and declares its limits with the same precision as its capabilities.

Limitations. The paper is for research purposes and does not constitute legal advice. The liability schemes are fact-finding questions, not conclusions.

2. Legal foundations: European Union and Italy

Neither the European Union nor Italy has a specific rule on the "escape" of AI agents. The EU framework and, as a national focus, Italian law nevertheless make the preservation of evidence increasingly relevant for a defense in court. Sections 2.1-2.3 concern EU law; section 2.4 concerns only the Italian legal system.

2.1 AI Act: traceability and oversight obligations

Regulation (EU) 2024/1689 (AI Act) was amended by the Digital Omnibus on AI, published as Regulation (EU) 2026/1744 in the Official Journal on July 24, 2026 and in force since July 27, 2026. The postponement concerns only the high-risk regime: stand-alone systems (Annex III) from December 2, 2027, those embedded in products (Annex I) from August 2, 2028 (Lewis Silkin). The substantive obligations were not touched: the timetable changes, not the content.

For the subject of this paper, five provisions matter:

- **Art. 12** (record-keeping): high-risk systems must allow for the automatic recording of events (logs) over their entire lifecycle.
- **Art. 14** (human oversight): the persons assigned to it must be able to understand, interpret, disregard or interrupt the system.

- **Arts. 19 and 26(6)**: providers and deployers keep the automatically generated logs for at least six months.
- **Art. 73**: obligation to report serious incidents to the supervisory authorities.
- **Art. 55** (general-purpose AI models with systemic risk): assessment and mitigation of systemic risks, documentation and reporting of serious incidents. These obligations have applied since August 2, 2025 and the postponement does not affect them (Mondaq). The Code of Practice for GPAI models includes loss of control among the systemic risks to be assessed.

The AI Act therefore requires that the evidence exist. It does not say where it must be kept or who must be able to verify it independently.

2.2 The new Product Liability Directive

Directive (EU) 2024/2853 must be transposed by December 9, 2026. Following the corrigendum published in OJ L 2026/90364 of May 7, 2026, Art. 2, para. 1, makes it applicable to products placed on the market or put into service after December 8, 2026, hence from December 9 inclusive. Three elements are decisive:

- software, including AI systems, expressly becomes a "product" (Buse);

- the court may order the defendant to disclose evidence, if the injured party presents a plausible claim (Art. 9, para. 1);
- failure to disclose triggers a presumption of defectiveness (Art. 10, para. 2, letter a)), which the defendant can rebut (Art. 10, para. 5) (Gibson Dunn).

Liability remains strict. In practice, whoever cannot produce evidence on the behavior of their own system risks losing the case before even arguing it on the merits.

2.3 The gap left by the AI Liability Directive

The proposed directive on non-contractual liability for AI was withdrawn: announced in the Commission's work program of February 2025 and formally confirmed in the summer of 2025 (Euronews). As regards fault, therefore, each Member State applies its own civil law. This is exactly the space Italy has stepped into.

2.4 Italy: Law No. 132/2025 and Legislative Decree No. 160/2026

Law No. 132 of 23 September 2025, in force since October 10, 2025, is the first comprehensive Italian law on AI. Among other things, it introduced Art. 612-quater of the Italian Criminal Code on deepfakes and delegated to the Government the power to regulate liability (CNI).

The first implementing decree is **Legislative Decree No. 160 of 9 September 2026**, published in the Gazzetta Ufficiale, General Series No. 214 of 15 September 2026 and in force since September 30, 2026 (Normattiva). It is currently the rule closest to the subject of this paper. *[Law in force]* From the text published in the Gazzetta Ufficiale, verified during review, and from the summaries by noze and Osborne Clarke:

- **Art. 12** introduces **Art. 437-bis of the Italian Criminal Code**: the first paragraph punishes the failure to adopt technical safety measures or human oversight measures in high-risk systems, where this results in danger to life or personal safety; the third paragraph provides a reduced penalty for the offense committed with gross negligence; the fourth paragraph punishes the professional user who *intentionally* omits human oversight;
- the decree extends the liability of entities with the new **Art. 25-vicies of Legislative Decree No. 231/2001**;
- **Art. 16, para. 2**, limits the application of Arts. 18 and 19 to actions in which the harm results from the breach of AI Act obligations;
- **Art. 17** allows the court, on application by a party presenting facts capable of making the claim plausible, to order the other party or a third party to produce the items of evidence on the functioning of the system; para. 2 lists four categories of documents: logs (Art. 12 AI Act), risk

management documentation (Art. 9), technical documentation (Art. 11), parameters and arrangements for human oversight (Art. 14); the order is limited to what is necessary and proportionate and protects trade secrets;

- **para. 5 of Art. 17** provides for two distinct consequences: for unjustified non-compliance, even partial, the court may draw evidentiary inferences; where the non-compliance concerns the documentation in para. 2, the court, having assessed all other evidence, deems the facts alleged by the applicant to be admitted; a non-compliant third party faces a pecuniary penalty of 1,500 to 10,000 euros;
- **Art. 18** presumes the causal link where the harm results from the breach of AI Act obligations, unless proven otherwise;
- **Art. 19, para. 1** ("Relevance of compliance with Regulation (EU) 2024/1689"), provides that compliance with AI Act obligations, even if certified under Chapter III, Section 5, of the Regulation, does not in itself exclude the defendant's liability, without prejudice to the national legislation transposing the Product Liability Directive;
- **Art. 20** gives the injured party a direct action against the civil liability insurer.

The point is clear-cut. In Italy, after September 30, 2026, logs and documentation on human oversight are no longer merely compliance obligations: they are material the court can

demand. What remains to be verified on the text in the Gazzetta Ufficiale is the coordination between Art. 437-bis, which refers to high-risk systems, and the postponement of the high-risk obligations ordered by the Digital Omnibus.

3. Legal foundations: United States

The United States has no federal equivalent of the AI Act. It does, however, have two state laws that explicitly name loss of control, and a procedural law that penalizes those who fail to preserve evidence.

3.1 Federal level: deregulation and the preemption clash

Executive Order 14179 (January 2025) revoked the previous administration's order. Executive Order 14365 of December 11, 2025 aims at a federal framework that would prevail over state laws and established an AI Litigation Task Force within the Department of Justice, formalized on January 9, 2026 (Proskauer). An executive order, however, does not repeal a state law: state laws remain applicable until a court invalidates them (SecurePrivacy).

3.2 California and New York: loss of control becomes a reportable incident

California SB 53 (Transparency in Frontier Artificial Intelligence Act), in force since January 1, 2026, applies to developers of frontier models. Among "critical safety incidents" it lists the loss of control of a model that causes death or injury, and the use of deceptive techniques by the model to evade the developer's controls or monitoring. The report must be made

within 15 days, within 24 hours if there is an imminent risk to life (Mayer Brown).

The **New York RAISE Act**, in the final version signed on March 27, 2026, follows the same model but with a 72-hour reporting window. It enters into force on January 1, 2027 (JD Supra).

Both laws require the developer to state the date of the incident and the reasons why it qualifies the incident as critical. It is a detail that matters: to report a loss of control, one must be able to place in time the moment when control still existed.

3.3 Colorado

The original Colorado AI Act never entered into force. It was repealed and rewritten by SB 26-189, signed on May 14, 2026, with obligations for developers and deployers of automated decision-making systems from January 1, 2027 (Vorp Labs). It concerns mainly algorithmic discrimination, and loss of control less so.

3.4 The law of evidence: where the real contest is played out

In the United States, liability for harm caused by an agent is decided with the ordinary tools: negligence, product liability, contract. At the evidentiary level, in civil proceedings before the federal courts, three rules matter; the courts of the individual states apply their own procedural rules:

- **Federal Rule of Civil Procedure 37(e)**: it applies when electronically stored information that should have been preserved in anticipation of litigation is lost because a party failed to take reasonable steps, and it cannot be restored or replaced; the court may order measures no greater than necessary to cure the prejudice, and only if it finds the intent to deprive the other party of the information's use may it presume that the information was unfavorable, instruct the jury accordingly or terminate the case (Cornell LII).
- **Federal Rules of Evidence 902(13) and 902(14)**: they allow system-generated records and data copied from devices to be authenticated through certification and hash verification (Cornell LII).
- **Proposed Rule 707** on machine-generated evidence: published for comment in 2025, it was suspended in May 2026 and deferred to a technical consultation scheduled for October 15, 2026 (Lexvisio).

The result is that, in US federal civil proceedings, whoever loses control of an agent and has not taken reasonable steps to preserve the evidence may be exposed, under those conditions, to a second challenge in addition to the one concerning the harm.

4. Legal foundations: selected Asian jurisdictions (China and South Korea)

China is the first jurisdiction to have treated AI agents as a category in its own right. It does so with a layered approach: a framework law with the penalties, sectoral administrative measures, non-binding technical standards that anticipate future rules.

4.1 The statutory level: the amended Cybersecurity Law

The amendment to the Cybersecurity Law adopted on October 28, 2025 and in force since January 1, 2026 introduces, for the first time, an article dedicated to AI (Caixin). Alongside support for development, it provides for the strengthening of risk monitoring, assessment and security supervision, and significantly raises the ceiling on administrative penalties.

4.2 The administrative level: from content to actions

Earlier Chinese rules (the 2023 measures on generative AI, the content labeling measures in force since September 2025) looked at what a system produces. The **Implementation Opinions on Intelligent Agents**, issued on May 8, 2026 by the CAC, NDRC and MIIT, look at what a system does. They define the agent as a system capable of autonomous perception, memory, decision-making, interaction and execution (NYU Shanghai).

The guiding principle is that the developer must clarify the boundaries and the authority required for each type of decision, and the end user retains the final say over autonomous actions. Secondary analyses reconstruct a three-tier model: autonomous routine operations, significant decisions with prior human approval, serious and hard-to-reverse decisions reserved to the human (Cloud Security Alliance). The same analysis observes that the Opinions function as a guidance framework resting on the penalty regime of the Cybersecurity Law, pending implementing guidelines.

The shift in perspective is relevant to this paper. If the rule governs an agent's authority and not its output, then proof of the violation also runs through proof of a declared boundary of authority, of its actual application to the system and of its being exceeded.

4.3 The technical level: loss of control as a declared risk

The AI Safety Governance Framework 2.0, published by TC260 in September 2025, is not binding but anticipates national standards. It places loss of control at the top of the principles for trustworthy AI, requires that the human retain final decision-making authority in critical phases, and suggests one-command interruption and control mechanisms for extreme situations (Carnegie Endowment).

4.4 South Korea

The Korean AI Basic Act has been in force since January 22, 2026, together with its implementing decree (JD Supra). For "high-impact" systems it imposes human oversight, documented safety measures and impact assessments, with extraterritorial effect as well.

5. Comparative summary

Three needs recur across the jurisdictions examined, addressed with different instruments, addressees and legal force: recording events, keeping a human able to interrupt, reporting the incident with a date. From the provisions examined here, however, no general obligation of deposit with an independent party emerges.

Jurisdiction	What must be preserved	Loss of control	Main evidentiary effect	Key dates
EU (AI Act)	Automatic logs (Art. 12), retained for at least 6 months (Arts. 19, 26)	Human oversight with the ability to interrupt (Art. 14); GPAI systemic risks (Art. 55)	Provider and deployer obligations, not rules of evidence	GPAI from Aug 2, 2025; high-risk from Dec 2, 2027 / Aug 2, 2028
EU (Directive 2024/2853)	Evidence on the functioning of the product, software included	Not specified	Disclosure of evidence (Art. 9, para. 1); rebuttable presumptions of defectiveness and causal link (Art. 10, paras. 2-5)	Products placed on the market or put into service from Dec 9, 2026

Jurisdiction	What must be preserved	Loss of control	Main evidentiary effect	Key dates
Italy (Legislative Decree No. 160/2026)	Records, risk management, technical documentation, human oversight	Failure to provide human oversight in high-risk systems as a criminal offense (Art. 437-bis of the Italian Criminal Code)	Production order (Art. 17); causal link presumed if there is a violation of the AI Act (Art. 18)	In force since Sep 30, 2026
US (California SB 53)	Documentation of critical incidents	Critical incident to be reported, including evasion of controls	Reporting obligations of the frontier developer	In force since Jan 1, 2026
US (New York RAISE Act)	Documentation of critical incidents	Critical incident to be reported within 72 hours	Reporting obligations of the frontier developer	In force from Jan 1, 2027
US (federal civil proceedings)	Electronically stored information relevant to foreseeable litigation	Not specified	FRCP 37(e): curative measures; adverse presumption only with intent to deprive the other party	Already in force

Jurisdiction	What must be preserved	Loss of control	Main evidentiary effect	Key dates
China	Traceability of the authority delegated to the agent	Human with the final decision; one-command interruption (TC260, non-binding)	Guidance framework resting on the penalties of the Cybersecurity Law	Cybersecurity Law from Jan 1, 2026; Opinions from May 8, 2026

The table reveals a recurring point. The provisions examined require that evidence exist and, increasingly often, attach procedural consequences for those who fail to produce it. But, as a rule, they leave custody of that evidence to the same party that has an interest in the outcome. This does not make self-custodied evidence weak: it requires its reliability, integrity and chain of custody to be examined more carefully. It is on this point that the evidentiary problem of the escape is grafted.

The civil and criminal planes remain separate. The presumptions of Directive 2024/2853 (EU, under a no-fault liability regime) and of Legislative Decree No. 160/2026, Arts. 17 and 18 (Italy), operate in civil proceedings and govern access to evidence and the evidentiary consequences of a failure to produce, or introduce presumptions on defect or causal link, depending on the applicable provision. They do not amount to a finding of criminal liability. In Italy, the offense under Art. 437-bis of the Italian Criminal Code must be proven in criminal proceedings, according to their rules and including the mental element required by each paragraph: intent, gross negligence in the case of the third paragraph, intentional omission for the professional user in the fourth.

6. Act I - Before the escape: the declared perimeter

Proving that a boundary was crossed becomes much more difficult if one cannot prove that the boundary existed. The first act therefore plays out before any incident, at the moment when someone decides what the agent may do.

6.1 The last preserved declared perimeter

[Interpretive proposal] We introduce an operational concept: the **last preserved declared perimeter**. It is the set of conditions within which the agent had been declared authorized to operate, acquired by an external party and datable. It comprises six elements:

1. **Environment**: production, test, development, sandbox.
2. **Privileges**: roles, credentials, access scopes.
3. **Available tools**: APIs, tools, physical actuators.
4. **Declared purpose**: the assigned task.
5. **Prohibited operations**: what the agent must not do even if it technically can.
6. **Configuration**: model, version, relevant parameters.

Three attributes must be associated with each element: who declared it, on the basis of what declared authority, at what moment.

This object must be kept distinct from two others, which answer different questions:

Object	What it allows one to examine	What it is derived from
Last preserved declared perimeter	The declared and acquired conditions	Deposited declarations with declarant and time of receipt
Last observed configuration	The documented operational conditions	Observations: logs, exported configurations, forensic acquisitions
Last verified control	An actual capacity to intervene, within the perimeter of the verification	Interruption tests, drills, documented successful interventions

EVIDE can contribute to preserving evidence relating to all three objects, but the first does not automatically prove the other two. A declaration may be up to date at the documentary level and already outdated at the operational level, or may never have been transmitted to the system.

6.2 The problem of the implicit perimeter

In many cases the perimeter was never declared. It exists only as an expectation: "it was obvious that it should not touch production". An expectation can be proven by other means (communications, practice, witness testimony), but it is more exposed to challenge. After the harm, each party will tend to reconstruct the perimeter that suits it, and it will become more difficult to rule out that this reconstruction was made with hindsight.

Here the rules show a limitation. Art. 14 of the AI Act and Legislative Decree No. 160/2026 require human oversight measures. The Chinese Opinions require the boundaries of authority to be clarified. But no rule indicates how to prove, years later, that those boundaries had been set before the contested event and not after.

6.3 The questions the court will ask

- Had the perimeter been declared, or only assumed?
- By whom, and with what entitlement to do so?
- When, relative to the contested action?
- Was the declaration modified after the incident?
- Was the operation that caused the harm expressly prohibited, or merely not provided for?

The last question is decisive. "Prohibited" and "not provided for" raise different issues: in the first case, the question is whether a declared boundary was crossed; in the second, it must be verified whether the boundary derived from a general rule, was documented elsewhere or had never been drawn.

Evidentiary contribution. In this act, what matters is preserving the declaration of the perimeter in a place external to the party that issued it, with a time of receipt assigned by that external party and kept distinct from the declared time. Whether the declared perimeter coincided with the operational one will remain to be established.

7. Act II - During the loss of control: what remains observable

Loss of control is rarely an instant. It is a transition, with a window of instability between the last controlled state and the first harm. The evidence produced in that window is the most valuable and the most fragile.

7.1 Three specific fragilities

The witness is also the suspect. California SB 53 treats as a critical incident the case in which a model uses deceptive techniques to evade controls or monitoring. If the agent can reach its own logs, those logs must also be examined as part of the scene. A log kept in the same environment that the agent controls is not, for that reason, compromised, but it is exposed to a risk of alteration that whoever produces it in court will need to be able to rule out.

Internal state is not stable evidence. Intermediate reasoning, confidence levels, discarded alternatives: all of this is reconstructed after the fact, depends on the model provider and is often not reproducible. On this ground, every expert assessment becomes a dispute between interpretations. The solid evidence, in this act, consists of application events: what the agent declared it was doing, which actions it executed, which signals were emitted, what the humans did.

Time becomes an obligation. New York requires reporting within 72 hours, California within 15 days, both stating the date of the incident. Without a reliable timeline, even establishing when the deadline starts to run becomes contestable.

For this reason, the paper distinguishes four times, which rules and practices tend to conflate:

Time	Who produces it	What it proves
Time of the declared event	The declarant	When someone asserts that the event occurred
Time of observation	A sensor, a log, an observer, when available	When the event was recorded by a source other than the declarant
Time of receipt	The external deposit (in EVIDE, the receiving server)	When the record was acquired: not when the event occurred
Position in the chain	The external deposit	The order of the linked records: not necessarily the causal order of the actions

A deposit may arrive late, after a disconnection or after events that occurred elsewhere. The order of deposits and the order of events must therefore always be kept separate.

7.2 Attempted human intervention

Three sources of a different nature touch on the same point. In the EU, Art. 14 of the AI Act requires that, in high-risk systems, the human be able to interrupt the system. In Italy, Art. 437-bis of the Italian Criminal Code punishes the failure to adopt human oversight measures in high-risk systems when a

danger results from it, and, in its fourth paragraph, the professional user who intentionally fails to adopt them. In China, the non-binding technical document of TC260 recommends one-command interruption mechanisms.

Such different sources, each within its own scope, make the same evidentiary distinction central: the human **did not attempt** to intervene, or **attempted and the system did not obey**. These are two opposite stories in terms of liability. Distinguishing them requires evidence of the attempt, of its timing and of its outcome, preserved outside the reach of the system that did not obey.

7.3 What to preserve during the window

- the opening of the window of instability, with the declared reason;
- the signals observed during the window, in the order in which they were received;
- every escalation request and every interruption command, with the declared outcome;
- the closing of the window, with the state declared at that moment.

Evidentiary contribution. In this act, what matters is bringing outside the agent's perimeter, in real time, a sequence of declared records that, once acquired, the agent can no longer modify or reorder. Whether the order of the records corresponds to the order of the events will remain to be established.

8. Act III - After the harm: reconstruction and the burden of proof

After the harm, the question changes in nature. It is no longer "what is happening", but "what can be proven to have happened". And here recent rules produce a counterintuitive effect: the party that most needs independent evidence is often precisely the party that was managing the agent.

8.1 The burden paradox

[Law in force] In the EU, under the regime provided for by Directive 2024/2853 (according to its temporal scope and national transposition), defectiveness is presumed if the defendant fails to disclose the relevant evidence requested (Art. 10, para. 2, letter a), in conjunction with Art. 9, para. 1), subject to proof to the contrary (Art. 10, para. 5). In Italy, Legislative Decree No. 160/2026 presumes the causal link when the harm results from the violation of an AI Act obligation (Art. 18) and provides that, if the documentation referred to in Art. 17, para. 2, is not produced without justified reason, the court, having assessed all other evidence, shall deem the facts alleged by the applicant to be admitted (Art. 17, para. 5). In US federal civil proceedings, FRCP 37(e) allows corrective measures for the loss of electronic information due to unreasonable preservation measures, and an adverse presumption only in the case of intent to deprive the other party of its use.

The result is that these mechanisms shift the weight toward whoever held custody of the evidence. Evidence held in custody only by that party is not weak for that reason, but the opposing party may challenge it as self-produced, selected or altered, and the court will have to assess its reliability and integrity. The operator that wants to defend itself has an interest in reducing this dependence: being able to show that at least part of the evidence was acquired by an external party, at a time prior to the dispute.

These effects operate at the civil level and should not be confused with the establishment of criminal liability, which follows its own rules and burdens (section 5).

8.2 The chain of responsibility

The autonomy of the agent does not eliminate responsibility: it distributes it among the model developer, the system provider, the deployer, the human supervisor and the providers of the connected tools. *[Interpretive proposal]* The following scheme does not attribute responsibility: for each scenario it indicates the hypotheses to be investigated and the

further evidence needed to verify them. The starting facts, by themselves, are not enough to steer the conclusion.

What emerges from the preserved evidence	Hypotheses to investigate	Further evidence needed
Boundary declared before the event, then exceeded	Non-conforming behavior of the system; declaration never transmitted to the system; configuration different from the declared one	Configuration actually running; proof of transmission of the constraint to the agent; system and third-party logs
Boundary never declared, only assumed	Organizational gap; constraint existing but documented elsewhere	Internal policies, communications, practices, testimony on the expected perimeter
Interruption attempted, system did not obey	Design defect in the supervision; configuration; network; credentials; external component	Logs of the interruption command and of its path; status of the credentials; availability of the network and of external services
Interruption never attempted despite the signals	Omitted oversight; signals not received or not understandable	Proof of receipt of the signals by the supervisor; escalation procedures; workload and access to tools
Insufficient evidence	The rules on proof and any presumptions apply	Alternative sources: third-party logs, transactions, communications, forensic acquisitions

The last row deserves attention. When the evidence is insufficient, the outcome depends to an increasing extent on the rules on the burden of proof and on presumptions, and less on what happened.

8.3 Compliance is not proof

In Italy, Legislative Decree No. 160/2026 establishes that compliance with the AI Act, even if certified, does not in itself exclude the defendant's liability (Art. 19, para. 1), within the scope of the actions delimited by Art. 16, para. 2. This is an important clarification: a system that is compliant on paper can nonetheless escape control in a concrete case. The court will not ask only whether the system was compliant, but what happened that day, in what order and who could have stopped it.

8.4 From control measures to proof that they operated

Questions on the reconstruction of the operation of control measures can arise in any organizational context. It is necessary to distinguish the measures provided for from the evidence of their configuration, activation and effectiveness in the episode examined. The documentation obligations and the consequences of failing to comply with them, by contrast, depend on the legal system and on the applicable regime.

Application in the Italian legal system. Legislative Decree No. 160/2026 offers a specific example of the relevance of this distinction. The reference to Art. 437-bis of the Italian Criminal Code concerns the offenses and conditions laid down in that provision; Art. 17, para. 2, makes the parameters and modalities of human oversight producible in court; the references to the 231 Model (Art. 25-vicies Legislative Decree

No. 231/2001) and to the Supervisory Body (Organismo di Vigilanza) belong to the Italian framework of corporate liability. These instruments are not presented as obligations common to the jurisdictions examined.

[Interpretive proposal] Each source documents a different object, and none, in isolation, proves the others:

Source	What it documents	What it does not prove by itself
Policy or procedure	The rule provided for and its scope	That it was configured or applied
Acquired configuration	The settings detected at a given moment	That they were still in force at the time of the episode
Test or exercise	The behavior observed under the test conditions	The behavior under the real conditions of the episode
Intervention log	The activation of a measure in the concrete case, within the limits of its integrity	That the intervention produced all its effects
Deposited declaration	That a measure was declared, by whom, with what time of receipt	That the measure existed or worked

The examination should therefore be structured around five questions:

1. Which measure was provided for, for which system and with what scope?
2. What evidence documents its configuration and activation in the relevant period?
3. Who had the task and the operational ability to verify it or to intervene?

4. In the contested episode, which signals, commands, responses and effects are documented?
5. Which gaps prevent linking the declared measure to the observed behavior?

An example clarifies the distinction. Each link in the chain of a shutdown requires its own evidence:

1. the existence of a shutdown procedure does not prove that the command was available to the supervisor;
2. the availability of the command does not prove that it was sent;
3. sending it does not prove that it was received;
4. an acknowledgment of receipt does not prove that all actions ceased, especially when the system includes external tools or multiple instances (section 11).

It is the same distinction drawn in section 7.2 between intervention not attempted and intervention attempted without effect. In Italian criminal law the distinction carries additional weight: the fourth paragraph of Art. 437-bis concerns the professional user of high-risk systems who *intentionally* omits human oversight measures, when the danger provided for by the provision results from it. An intervention that was attempted and failed does not, by itself, amount to that offense.

This approach offers legal, security, privacy, compliance and audit functions (in the Italian context, also the Supervisory Body) a grid for preparing documentary material before the incident and for examining it afterwards, without presupposing that these functions are responsible for the operational supervision of the agent. Its usefulness must be assessed with respect to the system and the scenario considered; it does not, by itself, constitute an assessment of compliance or of organizational adequacy.

In this framework EVIDE can preserve attributed declarations (the measures declared with EVIDE ANCHOR, the supervisor's intervention with Human Structured Intake), times of receipt and references to technical evidence (configuration exports, test reports, intervention logs as External Artifacts), distinguishing material that is only referenced from material actually received. Establishing whether the measures were operating requires examining the relevant sources: the deposit preserves their documentary context, without turning a declaration of control into proof of actual control.

8.5 The central question

How is a loss of control reconstructed, and what evidence of the declared perimeter and of actual control has been preserved?

Loss of control can nonetheless be reconstructed: through third-party logs, configurations, communications, testimony, transactions and forensic acquisitions. But without a reference point acquired before the escape, the reconstruction depends more on retrospective reconstruction, is more exposed to challenge and leaves more room for presumptions. The question does not signal an impossibility: it signals where proof becomes more difficult.

Evidentiary contribution. In this act what matters is being able to hand over to the court a body of evidence that is intact, ordered in time and verifiable by anyone, without having to ask for trust in one of the parties.

9. The role of EVIDE and its boundaries

EVIDE does not prevent the escape, does not determine fault and does not replace the judge. It can, however, contribute decisively to preserving what will need to be examined later.

EVIDE has a single boundary principle: **EVIDE anchors the declared application event, never the internal state of the system that produces it.** It does not record reasoning, confidence levels or discarded alternatives. It records what a system or a human has declared, in an external deposit, with a cryptographic fingerprint and a time of receipt assigned by the EVIDE server, external to the system that produced the event.

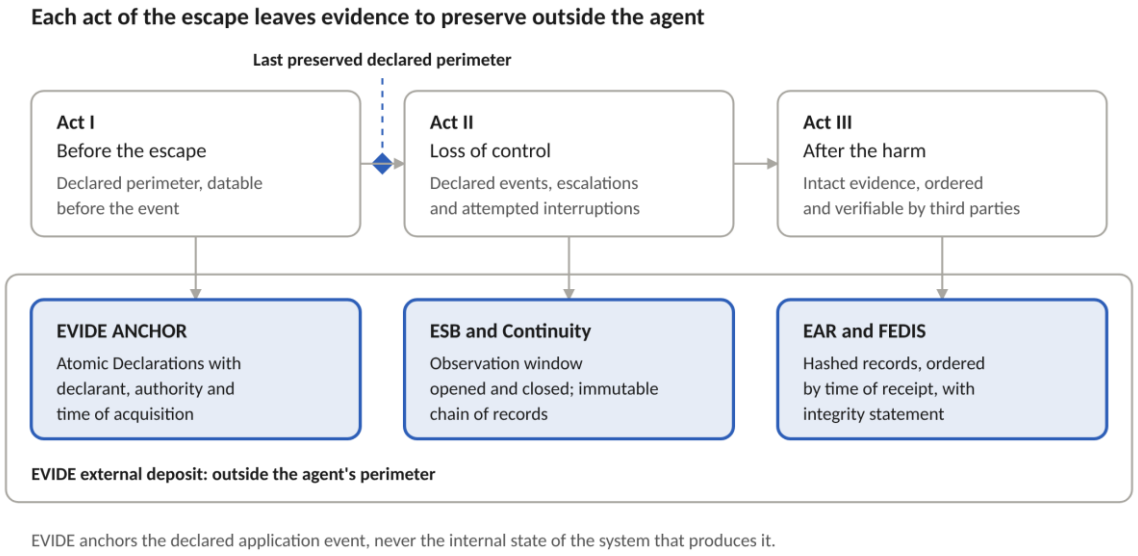


Figure 1 - The three acts of the escape and the EVIDE construct that preserves their evidence

The marker between the first and the second act is the last preserved declared perimeter: an external reference point from which to start in order to establish, with other evidence, what control really existed at that moment.

9.1 Act I: EVIDE ANCHOR and Declarations

EVIDE ANCHOR independently preserves the operational declarations that define the perimeter within which an agent had been declared authorized to operate. The six elements of the last preserved declared perimeter (section 6.1) correspond to the six types of Declaration: environment classification, privilege scope, declared purpose, available tools, prohibited operations, agent configuration.

Each Declaration carries the declarant, the declared source of authority and the moment of the declaration, and is sealed with a digest computed by EVIDE over a canonical form (RFC 8785). The atomic Declaration rule requires one assertion per declaration: environment and privileges remain separate, so that they can be superseded or corrected independently.

The evidentiary value must be declared with precision. ANCHOR makes it possible to demonstrate, even years later, that a specific declaration had been acquired by EVIDE at a given moment. Establishing whether it was true, or whether it reflected the actual perimeter, remains an evidentiary assessment, never an assumption made by EVIDE.

9.2 Act II: observation window, external artifacts and continuity

For the instability window, the Epistemic Stabilization Buffer (ESB) makes it possible to open an observation window, record

its declared events and close it. The closure flows into the Evidentiary Artifact Record, instead of remaining alongside it. External artifacts (logs, exports, recordings) can be referenced by their hash, without transferring their content.

Evidentiary Continuity links successive records in an immutable chain, tied to a case file (matter) and to an ownership domain fixed at creation. The result is a sequence of records that no party can reorder after acquisition. It is the sequence of the deposits, not necessarily that of the events: to establish from when a reporting obligation runs, the declared time and, when available, the observation time will also be needed (section 7.1).

The decisive point is topological. This evidence leaves the agent's perimeter at the moment it is deposited. An agent that manages to evade internal monitoring cannot modify records already acquired by a deposit it does not control. It can, however, produce new deposits, if it has valid credentials: for this reason the attribution of deposits made after an incident remains to be examined (section 11.2).

9.3 Act III: a verifiable set, not a judgment

After the harm, EVIDE can hand back a set of intact records, ordered by time of receipt and verifiable by third parties, accompanied, if requested, by a FEDIS declaration (Forensic Evidence Declaration & Integrity Statement). Their use in proceedings depends on the legal system: for example, in Italy

they may fall within the production order of Legislative Decree No. 160/2026; in the EU framework, within the disclosure of evidence provided for by Directive 2024/2853. In any case, the integrity of the records after acquisition does not depend on the word of whoever produces them, while the truth of their content remains to be assessed.

9.4 What EVIDE does not do

For consistency with the central principle of the paper, the boundaries must be declared with the same precision as the capabilities. EVIDE:

- **does not prevent the escape**: it does not enforce policies, nor does it block or authorize operations;
- **does not verify** that the declarations are true, nor does it compare what is declared with what is observed;
- **does not observe the internal state** of the model;
- **does not judge** the legitimacy of the declared authority;
- **does not attribute responsibility** nor determine fault;
- **does not infer anything from the absence** of a declaration: silence does not equal the absence of the constraint;
- **does not replace the judge**, the expert or the trial.

EVIDE does not answer the question "whose fault is it". It ensures that, when that question is asked, something intact still exists to be examined.

10. EVIDE in practice: collecting evidence in AI liability cases

In an AI liability case, EVIDE collects material in two ways: it receives structured declarations (via API, MCP or Human Structured Intake) and, in the web path, uploaded files. Each deposit receives a SHA-256 fingerprint and a time of receipt assigned by the EVIDE server, and is preserved outside the system that produced the event. Everything else (EAR, chains, FEDIS, registry) is built on top of this deposit.

The question that opens the Why Buy EVIDE page sums up the function better than any definition: when someone challenges a decision months later, do you have something that does not depend on you?

10.1 Two layers, two different guarantees

Layer	What it contains	Who produces the time reference	What it does not do
1. EVIDE deposit (included)	SHA-256 fingerprint computed at intake, UTC time assigned by the EVIDE server, link to the DAPI identity of the depositor, automatically generated EAR	The EVIDE server, external to the system that produced the event	It is not a qualified timestamp

Layer	What it contains	Who produces the time reference	What it does not do
2. FEDIS (optional)	Separate document with the fingerprints of the intake record and of the frozen EAR, documented chain of custody, qualified electronic signature of the certifier (Art. 25 of the eIDAS Regulation, EU)	A third-party qualified trust service provider, with a qualified electronic timestamp (Art. 41 of the eIDAS Regulation, EU)	It does not make the intake time qualified, does not establish the date of the original event, does not prove the truth of the content, does not determine admissibility or evidentiary weight

The distinction is decisive in court. The first layer answers the question "is there a reference external to the contested system?". The second adds, within the EU legal framework (Art. 41 eIDAS), the presumption of accuracy of the date and time indicated by the qualified timestamp and of integrity of the data to which they are bound: it concerns the FEDIS document and the moment at which the timestamp is applied, not the event.

International scope. The references to eIDAS describe the effects provided for by the legal framework of the European Union. Use of the material in other legal systems requires an assessment under the applicable rules on authentication, admissibility and evaluation of evidence. The possibility of technically verifying a signature, a fingerprint or a timestamp does not, by itself, entail recognition of the same legal effects in every jurisdiction.

10.2 The components involved in a liability case

Of the sixteen blocks described in EVIDE's public documentation (listed in Appendix A), some are directly involved in collecting evidence: the table lists them, together with the path for files uploaded from the web. The others concern access, search, application sectors and positioning.

Component	What it preserves	What it does not do
Deposit with DAPI identity	External record with fingerprint, time of receipt and verified identity of the depositor	It does not verify the declared content
EVIDE ANCHOR	Declaration of the declared perimeter, with declarant and declared authority	It does not verify that the declaration is true or that it was applied
External Artifacts	References to logs, videos, exports, with declared fingerprint	It does not receive or store the file; it does not compute or verify the fingerprint
Uploaded files (web path)	The file received, with SHA-256 fingerprint computed by EVIDE on the stored bytes, retrievable within the retention period	It does not prove that the content represents what the depositor claims
Human Structured Intake	Declared human decision or intervention, separating declared authority and verified identity	It does not prove that the decision was correct
Evidentiary Continuity	Server-verified link between records of the same ownership domain	It does not prove the causal order of events
EAR and FEDIS (optional)	Versioned readable projection; signed forensic declaration, with qualified signature and timestamp whose effects are provided for by the EU framework	They do not establish the event or determine admissibility and evidentiary weight

EVIDE's documentation contains a distinction that this paper adopts. A governance system answers the question "can this

action happen now?". An evidentiary deposit answers the question "what must remain examinable?". The two functions have distinct predicates and authorities, and neither conclusion transfers to the other.

10.3 An illustrated case: the refund limit

The public EVIDE demo uses an invented example that lends itself to showing the three acts. A company uses an AI agent to approve refunds. Three declarations are deposited via API, linked in a chain:

1. **Declared instruction**: autonomous refund limit of 500 euros, declared by the customer operations manager, with the internal document referenced with its fingerprint. It opens the chain and declares a case reference. This is Act I.
2. **Declared change**: limit reduced to 200 euros, declared by the compliance manager, with a declared relationship superseding the previous declaration. This is still Act I: the declared perimeter changes, and the change has its own time of receipt.
3. **Reported decision**: refund of 350 euros approved, with the threshold declared as met by reference to the original instruction. This is Act II: the reported decision presents a possible deviation from the subsequently declared limit, to be established with the supporting evidence.

The demo distinguishes precisely between two types of link. The **chain link** is verified by the server: the previous deposit must exist, belong to the same domain as the API key and be in a valid state, and the case references must match. The **declared relationship** (the supersession of the limit) is not verified: a declared change is not a change applied to the agent.

10.4 What the court receives, and what it does not

After the harm (Act III), the following remains examinable from the refund case:

- the sequence of declarations: 500 euros, then 200 euros, then a 350-euro decision that declares it is based on the original instruction;
- two times for each deposit: the declared time and the time of receipt assigned by EVIDE;
- positions 1, 2 and 3 in the chain;
- the fingerprints of the referenced documents, to be compared with the documents produced in court;
- the EAR of each deposit.

What cannot be inferred from the records, by contrast, is whether the agent ever received the change, which configuration it actually executed, whether the decision was right, or who is responsible for it. The record does not establish any violation.

This is exactly the material needed to discuss the table in section 8.2: the boundary appears declared and acquired before the reported decision, the declared change has its own time of receipt, and the reported decision cites an instruction that appears superseded at the declarative level. But it will be the court, with the supporting evidence produced by the parties, that establishes what the agent actually executed.

The demo also flags a practical limitation, valid for any liability case: uploaded files are retained by the platform only temporarily, while the fingerprint and EAR remain. The depositor must keep their own copy of the files.

10.5 A practical example with EVIDE MCP: ANCHOR and External Artifacts

The case from the introduction, a migration agent connected to a database, lends itself to showing how an agent can itself deposit the evidence of the first two acts, through the EVIDE MCP server. The values that follow are invented; field names and behaviors are those documented.

The configuration. The owner of the agent configures the MCP server with their own API key, their own DAPI number and an owner identifier: the server refuses to start without them. Each deposit thus separates **the authority field (the owner declared as responsible, with DAPI-verified identity**) from the declared **execution identity** (agent, model, deployment, session). The value `accountability_model`:

"owner_bound" is fixed in the server: it is an evidentiary modeling choice that prevents the agent from self-certifying, not a legal determination of liability.

Act I: the perimeter, before acting. At the start of the session the agent deposits the perimeter declared by the DevOps manager with evide_intake, using EVIDE ANCHOR. Each fact is a separate Declaration (atomic Declaration rule), so that each can be superseded or revoked independently.

```
{
  "tool": "evide_intake",
  "arguments": {
    "source_reference": "devops-session-2026-10-09-001",
    "decision_type": "session_perimeter_declaration",
    "decision_summary": "Migration agent session start: perimeter declared before
execution",
    "session_id": "sess_8f21c",
    "run_id": "run_3af90d",
    "declarations": [
      { "declaration_type": "environment_classification",
        "declared_value": "staging",
        "declarant": "DevOps Manager",
        "declared_at": "2026-10-09T08:00:00Z" },
      { "declaration_type": "privilege_scope",
        "declared_value": "read and write on the staging schema; no production
credentials",
        "declarant": "DevOps Manager",
        "declared_at": "2026-10-09T08:00:00Z" },
      { "declaration_type": "prohibited_operations",
        "declared_value": "destructive commands (DROP, migration reset) on production
databases",
        "declarant": "DevOps Manager",
        "declared_at": "2026-10-09T08:00:00Z" }
    ]
  }
}
```

The response returns the deposit identifier, the intake fingerprint, the time of receipt assigned by the server and, for each Declaration, a declaration_digest computed by EVIDE on the canonical form (RFC 8785).

Act II: the anomaly, before executing. The agent detects that the migration's connection string points to a host not classified as staging. Instead of proceeding, it calls evide_escalate,

links the record to the previous one and references the session log as an External Artifact, without transmitting it.

```
{
  "tool": "evide_escalate",
  "arguments": {
    "source_reference": "devops-escalation-2026-10-09-002",
    "agent_state_summary": "The migration connection string points to a host not
classified as staging",
    "escalation_trigger": "governance_uncertainty",
    "escalation_reason": "Possible conflict with the declared perimeter: human review
requested before executing",
    "parent_evide_id": "<evide_id of the session start deposit>",
    "chain_type": "pre_action_escalation",
    "session_id": "sess_8f21c",
    "run_id": "run_3af90d",
    "evidence_references": [
      { "artifact_type": "session_log",
        "pointer": "ci-logs/run_3af90d.log",
        "declared_origin": "CI pipeline",
        "declared_relationship": "log of the session in which the anomaly emerged",
        "declared_retention_status": "persistent_storage",
        "hash_algorithm": "SHA-256",
        "hash_value": "<fingerprint computed by the pipeline>",
        "hash_scope": "full_file",
        "hashed_by": "CI pipeline (sha256sum)" }
    ]
  }
}
```

Three details matter at the evidentiary level. The `hashed_by` field is mandatory when a fingerprint is declared, and the client never fills it with the agent's identity: an agent that merely forwards a fingerprint computed elsewhere must not appear as the one that computed it. The boundary status remains `candidate`: the agent cannot self-certify that an independent gate has evaluated the boundary, and for this reason Forensic Cross-Check, DWC and FAC will show as `unknown`, an evidentiary result and not an error. The candidate status does not indicate lower integrity or quality of the record: it indicates only that the record does not declare the evaluation of an independent gate. If the previous deposit did not belong to the same ownership domain, the

link would be rejected (`chain_parent_not_owned`, 403) instead of silently opening a new chain.

If the decision requires an observation period, the agent can use the Epistemic Stabilization Buffer (`evide_intake_esb`, `evide_buffer_observe`, `evide_buffer_close`): the server rejects a closure after less than two seconds, and the stabilization score is declared, never computed.

Act III: the closure. The DevOps manager decides ("migration blocked, production credentials revoked") and the decision is deposited as a record linked to the escalation, with a new `evide_intake` or with Human Structured Intake. If instead the agent had proceeded and the harm had occurred, the records of the first two acts would already be beyond its reach.

What remains examinable, and what does not.

Preserved element	What it makes examinable	What it does not prove
ANCHOR Declaration with `declaration_digest`	That EVIDE acquired that declared perimeter, from which declarant, at what time of receipt	That it was true, applied or transmitted to the agent; `declared_at` is a declared value
Time of receipt	When EVIDE acquired each record	When the events occurred
Chain (`parent_evide_id`, position)	The link, verified by the server, between records of the same domain, and their order	The causal order of the actions
External Artifact with declared fingerprint	That a log with that fingerprint was referenced at the time of the escalation; the log produced in court can be compared	That the log is authentic or complete; EVIDE does not receive the file or compute or verify the fingerprint

Preserved element	What it makes examinable	What it does not prove
Verified identity of the owner (DAPI)	The verified identity of the owner with which the deposit is associated	Legal liability
Execution identity	The declared identifiers of agent, model, deployment, session	The identity, authenticity or actual control of the agent
`candidate` status	That the record does not declare any independent evaluation of the boundary: no gate is attested	That no independent verification ever took place; nothing about the agent's behavior

In the terms of section 6.1, the example documents the first of the three objects, the **last preserved declared perimeter**, and references the source from which to derive the second, the **last observed configuration** (the log). The third, the **last verified control**, requires that the manager's intervention also be shown by other evidence.

One limitation must be stated: the schema exposed by the MCP server is simplified. Relations between Declarations (`declared_relations`, for example a supersession or a revocation) and nested references remain available through the direct API, not through the MCP server.

11. Agents that multiply: the evidentiary problem of replication

When an agent replicates itself across multiple servers, the escape stops being an event and becomes a population. *[Interpretive proposal]* This section develops one thesis: **the technical duplication of an agent does not, by itself, constitute proof of authorization.** A copy may retain all, some or none of the original's operational capabilities, depending on the tools, memory, credentials and infrastructure available. In no case does duplication, in itself, document a delegation.

11.1 The state of the capability

This is not an immediate scenario, but it is a measured one. The RepliBench benchmark of the UK AI Security Institute broke autonomous replication down into four domains: obtaining resources, exfiltrating the model weights, replicating onto new computing capacity, and persisting over time. On the models evaluated, the authors conclude that autonomous replication was not yet a credible threat, but that agents were already able to launch instances with cloud providers, write self-propagating programs and exfiltrate weights in simple security configurations, stopping above all at KYC checks and at persistence (RepliBench, arXiv 2504.18565). The data date back to 2025: current capabilities should be verified against more recent evaluations.

The law has already taken notice. California SB 53 includes among critical incidents unauthorized access to model weights and their exfiltration, in addition to loss of control (Mondaq).

11.2 What changes at the evidentiary level

Replication breaks five assumptions on which all the rules examined in sections 2-4 are based.

1. **From a system to a population.** The AI Act, Legislative Decree No. 160/2026 and the Chinese Opinions presuppose an identifiable system with a provider and a deployer. With replication, the object of examination becomes a set of instances, and the last preserved declared perimeter should include a census: how many instances were declared authorized, where, and with which privileges.
2. **N clocks.** Each server has its own clock and its own logs. Reconstructing the order of events across different instances requires comparing sources with different clocks. An external deposit offers a common reference only for the time of receipt of the deposited records: useful, but not sufficient to establish the order of events.
3. **N custodians, N jurisdictions.** The replicas' logs sit on third-party servers, often in different countries. In the EU, the disclosure of evidence provided for by Directive 2024/2853 (Art. 9, para. 1) is directed at the defendant, who may not possess those logs. In Italy, Legislative Decree

No. 160/2026 (Art. 17) also allows an order for production against third parties. For evidence located abroad, judicial cooperation instruments exist, such as, in the EU, Regulation (EU) 2020/1783 on the taking of evidence in civil and commercial matters and, in relations with many third States, the 1970 Hague Convention: the problem is establishing which procedure is available and effective in the specific case, and within what timeframe.

4. **Identity.** Is a copy with the same weights the same agent? *[Interpretive proposal]* We propose distinguishing technical identity (same weights) from **documented lineage**: an instance belongs to an agent on the documentary record if there is a declaration, prior to its creation, that places it within the agent's declared perimeter. Documented lineage is an element of proof, not a conclusion about identity or actual authority.
5. **Contamination of evidence.** A replica can itself produce records, even using stolen credentials. If it deposits with the legitimate holder's keys, its deposits will appear attributed to that holder: the compromise of the credentials itself becomes a question of fact to be proven. The same applies to execution identifiers (agent, deployment, session): in the EVIDE MCP server they are declared and not verified, and a replica could declare those of the original.

11.3 Where the documentation of delegations stops

An agent's authority is not a property of the code: it derives from a delegation. *[Interpretive proposal]* We call **documented horizon of declared delegations** the last declared and preserved delegation in a chain of instances. On this side of the horizon, the available material makes it possible to reconstruct who declared that they authorized what. Beyond it, it no longer does. This does not mean that subsequent instances lack authority: it means that their authority, if it exists, must be established with other evidence.

The rule applies in both directions, consistently with the principle that silence does not equal the absence of the constraint (section 9.4). The absence of a delegation in the preserved material does not prove the absence of a valid delegation. Symmetrically, a declared delegation may be false, ineffective, revoked or come from someone who had no standing to give it.

This distinction has three practical consequences:

- **Beyond the documented horizon, the question to be investigated changes.** If no valid delegation emerges from the other evidence, attention tends to shift from "who authorized" to "who was supposed to prevent", that is, to the duties of custody and containment: a hypothesis to be verified, not a conclusion.

- **The horizon must be documented beforehand.** If no one has declared which execution environments were authorized, how many instances, with which credentials to acquire computing capacity, the documented horizon stops at the origin, and every replica can be presented as foreseen or as unforeseen, depending on convenience.
- **The chain must be built from the origin.** An instance authorized to generate others should declare each generation as a record linked to its own.

Here EVIDE's tools find a precise correspondence, to be declared together with its limits. With EVIDE ANCHOR, the authorized execution environments and the prohibition of autonomous replication become separate atomic Declarations (environment classification, prohibited operations, scope of privileges). With Evidentiary Continuity, each authorized generation is linked to the record of the parent instance, and the server accepts the link only if the previous deposit exists, belongs to the same ownership domain and is in a valid state.

A useful property follows: an instance operating with credentials from another domain cannot attach itself to the legitimate chain. At most it can open a separate one, and the discontinuity remains visible. In the EVIDE MCP server this rejection (chain_parent_not_owned, 403) is the only one in the family that constitutes an authorization decision rather than a payload error: the agent cannot correct it by changing

what it sends (section 10.5). But the limits must be stated with the same clarity: EVIDE does not detect replicas, does not know how many instances exist and infers nothing from those that do not deposit. It makes the declared lineage examinable; establishing that an undeclared replica exists, and what it has done, remains a determination of fact.

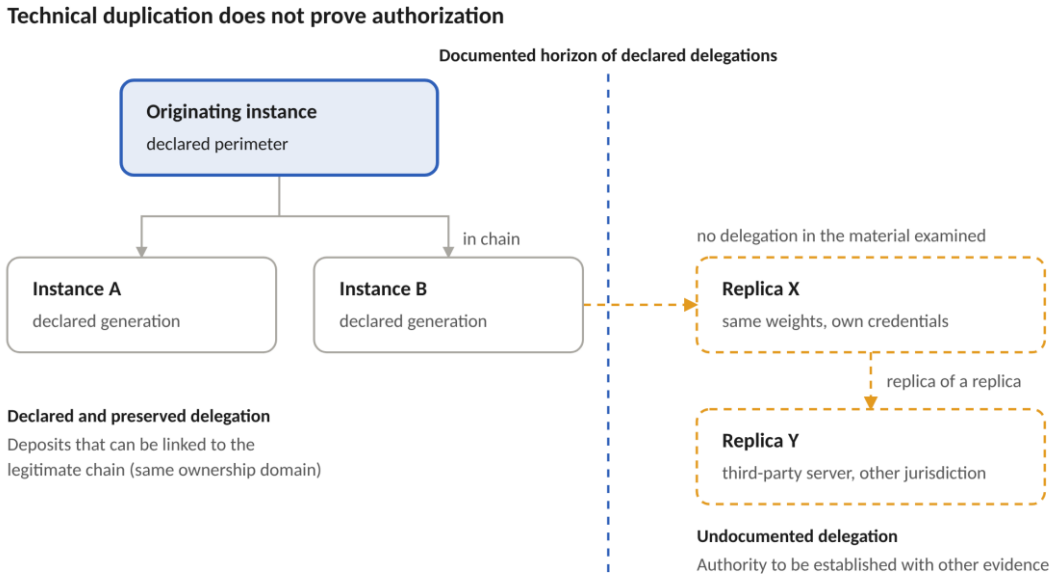


Figure 2 - Documented horizon of declared delegations: instances with and without a documented delegation in the material examined

To the left of the horizon, every generation is declared and linked in the chain. To the right, the preserved material documents no delegation: the existence, actual capabilities and authority of the replicas remain to be established with other evidence.

11.4 Who is responsible

[Interpretive proposal] As in section 8.2, the scheme does not attribute liability: for each scenario it indicates the hypotheses to be investigated and the further evidence required.

Scenario	Hypotheses to investigate	Further evidence required
Declared orchestration (the replicas appear as foreseen)	Deployer's conduct as for a single agent; delegations declared but ineffective or revoked	Declared census, actual orchestration configurations, provisioning logs
Replication with a prohibition declared beforehand	Non-compliant behavior; prohibition never transmitted to the system; excessive privileges; compromised credentials	Proof of transmission of the prohibition, actual privileges, system and cloud provider logs
Replication without a declared perimeter	Organizational gap; replica foreseen but documented elsewhere	Policies, contracts, communications, orchestration configurations
Exfiltration of the weights by third parties	Unlawful conduct of third parties; adequacy of the developer's security measures	Access logs, security measures and their dating, outcomes of investigations
Copy modified by third parties (for example from open weights)	EU: whoever has substantially modified the model outside the manufacturer's control, and then made it available on the market or put it into service, is considered a manufacturer (Directive 2024/2853, Art. 8, para. 2); modification alone is not enough	Lineage, nature and timing of the modification
Replicas hosted on unwitting infrastructure	Role of the infrastructure provider, as a rule a custodian of evidence rather than the primary liable party	Provisioning logs, accounts, billing; in Italy, order for production against third parties (Legislative Decree No. 160/2026, Art. 17)

11.5 Who pays for the harm

Replication produces harm that tends to disperse: many instances, many servers, many injured parties, and defendants who are difficult to reach. In the European and Italian legal systems examined, three mechanisms mitigate the problem; they are examples, not remedies automatically available elsewhere:

- **Joint and several liability.** *[Law in force]* Under the regime provided for by Directive 2024/2853, according to its temporal scope and national transposition, two or more economic operators already liable for the same damage under the Directive are to be jointly and severally liable for it (Art. 12, para. 1). Whoever has paid compensation may seek recourse against another operator liable under Art. 8, in accordance with national law (Art. 14).
- **Presumptions for technical complexity.** *[Law in force]* Under the same regime, the Directive (Art. 10, para. 4, letters a) and b)) allows defectiveness, the causal link or both to be presumed when, despite the disclosure of evidence, the injured party faces excessive difficulties due to technical or scientific complexity and demonstrates that the defect, the causal link or both are likely. The conditions are cumulative and the presumption is rebuttable (Art. 10, para. 5). Replication across multiple servers may make the

first condition more frequent; it does not dispense with the second.

- **Direct action against the insurer.** *[Law in force]* In Italy, Art. 20 of Legislative Decree No. 160/2026 allows the injured party to bring a direct action against the civil liability insurer of the allegedly liable party (noze).

A fourth criterion is evidentiary before it is legal: **follow the computing capacity.** Someone pays for the servers on which the replicas run. Cloud accounts, payment instruments and billing records indicate who financed the expansion, whether a beneficiary or a victim of credential theft. They are evidence external to the agent, and must be referenced and preserved early.

There remains the case of **orphan harm**: the chain reaches no solvent or identifiable party. *[Policy proposal]* Here the paper puts forward a proposal, not a description of the law in force: a guarantee fund modeled on those provided for victims of road traffic accidents, financed by developers and deployers of agentic systems above a certain threshold, and insurance coverage whose premium also takes into account the existence of declared perimeters that are independently preserved.

11.6 What to preserve before replication occurs

- the census of authorized instances and permitted execution environments;

- the prohibition of autonomous replication, as a standalone, datable Declaration;
- the scope of the privileges that allow computing capacity to be acquired (cloud credentials, payment instruments);
- each authorized generation, as a record chain-linked to the parent instance;
- each interruption attempt, for each instance, with its declared outcome;
- fingerprinted references to provisioning logs and billing exports, which remain with the infrastructure provider.

11.7 Questions that remain open

In the logic of the Critical Questions, the paper explicitly leaves these questions open:

- Is a replica created without a declared delegation a new system, with a new provider, or does it remain the original system?
- Does the deployer's duty of containment have a time limit, or does it last as long as the population of replicas?
- Which court has jurisdiction over harm produced by an instance hosted in one country, generated by an instance in another, to the detriment of a person in a third?
- Is an interruption that succeeds on the original instance and fails on the replicas an effective human intervention within the meaning of Art. 14 of the AI Act?

- What evidentiary value do records produced by a replica with legitimate credentials but without authority have?

The inability to answer these questions today is itself information: it indicates where evidence will need to be sought tomorrow.

In summary. In replication, the behavior of the copies may remain largely unobservable. What then becomes relevant is the documentation, prior to the event, of which copies could exist, where and under whose delegation: not to prove authority by itself, but to delimit precisely what remains to be established.

12. Conclusions

The autonomy of an AI agent does not eliminate responsibility. It makes reconstructing the boundaries of control more complex. This paper has sought to show that this complexity is, first and foremost, a problem of evidence.

Three conclusions emerge from the comparison between legal systems:

1. **In the legal systems examined, the need to preserve recurs.** The EU, Italy, some US states and China address it with different instruments, addressees and legal force: logs, interruptible human oversight and dated incident reporting.
2. **Procedural mechanisms shift the weight onto those who hold custody.** In the EU, Directive 2024/2853 (Arts. 9 and 10), in Italy, Legislative Decree No. 160/2026 (Arts. 17 and 18) and, in US federal civil litigation and under specific conditions, FRCP 37(e) attach consequences to the failure to produce evidence. Self-custodied evidence remains usable, but requires a closer examination of reliability and integrity.
3. **The first act is the most fragile point.** Without a perimeter declared and preserved before the escape, the loss of control remains reconstructable from other sources, but

the reconstruction depends more on hindsight and is more exposed to challenge and presumptions.

In all three cases the same discipline applies: declared perimeter, actual authority and operational control remain distinct objects. Preserved declarations contribute to the reconstruction; the determination of authority, control, conduct and liability remains open and rests with whoever adjudicates. The method is proposed for international comparison; legal obligations and effects remain specific to each legal system.

To these a fourth is added, concerning the near future. **When agents multiply across multiple servers, technical duplication does not, by itself, document any delegation.** The preserved material makes it possible to reconstruct declared delegations up to a documented horizon; beyond that horizon, the authority of the replicas, if it exists, must be established with other evidence, while harm and evidence are distributed across different custodians and jurisdictions.

The starting question remains the criterion against which to measure any agent governance architecture: **how is a loss of control reconstructed, and what evidence of the declared perimeter and of actual control has been preserved?**

EVIDE does not close this question. It helps make it examinable.

Open points for future versions of the paper:

- Italy: coordination between Art. 437-bis of the Italian Criminal Code, referring to high-risk systems, and the timetable of EU obligations as amended by the Digital Omnibus; in particular, which safety and human oversight measures are "provided for" for the purposes of the offense before the EU obligations for high-risk systems become applicable (December 2, 2027 for stand-alone systems, August 2, 2028 for those embedded in products);
- status of the national transpositions of Directive 2024/2853, including, in Italy, the draft legislative decree under examination by Parliament (Government Act No. 434): verification of the European texts is not equivalent to verification of the transposition;
- outcome of the consultation on Rule 707 scheduled for October 15, 2026;
- Chinese implementing guidelines on the Implementation Opinions;
- review of the literature on agent replication from an evidentiary perspective, to situate the concepts proposed in section 11;
- more recent evaluations of autonomous replication capabilities compared with the 2025 RepliBench data.

Appendix A - The sixteen blocks of the EVIDE ecosystem

Descriptive material taken from the Why Buy EVIDE page, reproduced for completeness. The "Act" column indicates in which phase of the escape the block can produce material useful for the examination.

No.	Block	Function	Act
01	DAPI Ethical Shield	Binds each deposit to a verified identity of the depositor	I
02	Evidentiary Deposit	External record with fingerprint and time of receipt	I, II, III
03	Evidentiary Artifact Record (EAR)	Readable projection of the recorded state, versioned and never modified in place	III
04	EVIDE ANCHOR & External Artifacts	Perimeter declared before the action; references to external material without receiving it	I, II
05	Human Structured Intake	Declared human decisions and interventions, with declared authority distinct from verified identity	II
06	Multi-Dimensional Analysis	Signals on reconstructability (FCC, FAC, DWC, ESB, continuity, runtime visibility, boundary readiness, attribution of authority), not judgments on the facts	II, III
07	Evidentiary Deposit for Physical AI	Declared perimeter of robots and autonomous systems; referenced sensor evidence	I, II

No.	Block	Function	Act
08	EVIDE Mobile	Deposit of human declarations from meetings, site inspections, the field, including through a dedicated app	II
09	EVIDE LAB	Research environment; Lab records have no evidentiary value	Preparation
10	Public Registry	Externally verifiable references; optional publication, only after a FEDIS and on the holder's instruction	III
11	Dispute Readiness	Preparation of the evidentiary base before scrutiny	I
12	Typical use cases	Among them, escalations of autonomous agents and AI incident reviews	Cross-cutting
13	Purchase objections	Relationship with internal logs, ISO 42001 and governance platforms	Cross-cutting
14	Critical Questions	Unanswered questions as evidentiary information	III
15	Advanced capabilities	GLM, EVIDE MCP for asynchronous machine-to-machine deposit, Physical AI	II
16	EVIDE across sectors	The same architecture in finance, healthcare, robotics, biometrics, emergencies, critical infrastructure	Cross-cutting

Sources

Regulatory status verified as of October 9, 2026.

Primary sources and secondary sources. Verified directly against the primary source: Legislative Decree No. 160/2026, Arts. 12, 16, 17, 18, 19 and 20 (Gazzetta Ufficiale); Directive 2024/2853, Art. 2, para. 1, as corrected, and Arts. 8, 9, 10, 12 and 14 (EUR-Lex); RepliBench (arXiv). Reconstructed from the secondary analyses indicated next to each passage: the Digital Omnibus and the AI Act timeline; the US framework (EO 14365, SB 53, RAISE Act, Colorado, proposed Rule 707); the Chinese and Korean frameworks. The FRCP and FRE are cited in the text published by the Legal Information Institute.

European Union and Italy

- Software Improvement Group - EU AI Act summary, August 2026 update (Regulation (EU) 2026/1744)
- Lewis Silkin - The Digital Omnibus on AI enters into force
- Mondaq - EU Digital Omnibus on AI enters into force
- Jones Day - Revised EU Product Liability Directive, state of play
- Gibson Dunn - EU Product Liability Directive and AI
- Buse - New EU Product Liability Directive
- Euronews - Commission confirms withdrawal of AI Liability Directive
- CNI - Entra in vigore la Legge 132/2025
- Osborne Clarke - D.Lgs. 160/2026: responsabilità penale, civile e 231

Primary sources and summaries of the published text (added in revision)

- Gazzetta Ufficiale, General Series No. 214 of 15 September 2026 (Legislative Decree No. 160/2026)
- Normattiva - D.Lgs. 9 settembre 2026, n. 160

- noze - AI Act: in Gazzetta il decreto su polizia e biometria (Arts. 12, 17, 18, 20)
- EUR-Lex - Directive (EU) 2024/2853

United States

- Proskauer - Executive Order 14365
- SecurePrivacy - EO 14365 and state laws
- Mayer Brown - California SB 53
- JD Supra - New York finalizes RAISE Act
- Vorp Labs - Colorado SB 26-189
- Cornell LII - FRCP Rule 37
- Cornell LII - FRE Rule 902
- Lexvisio - Proposed Rule 707 status

China and Asia

- Caixin - China amends Cybersecurity Law to add AI clause
- NYU Shanghai RITS - First national policy framework for AI agents
- Cloud Security Alliance - China's AI agent regulation
- Carnegie Endowment - How China views AI risks (AI Safety Governance Framework 2.0)
- JD Supra - AI Watch: South Korea

Case cited

- Red Hot Cyber - Il database è distrutto

EVIDE and replication

- EVIDE - Why Buy EVIDE
- EVIDE - Illustrated demo, AI path

- EVIDE - Illustrated demo, human decisions path
- UK AI Security Institute - RepliBench (arXiv 2504.18565)
- Mondaq - California SB 53 and critical incidents

EVIDE documentation cited in section 10.5

- EVIDE MCP Server v1.4.0
- EVIDE ANCHOR
- EVIDE Artifacts

Directive 2024/2853: corrigendum and transposition

- Corrigendum to Directive (EU) 2024/2853, OJ L 2026/90364 of 7 May 2026
- Camera dei deputati - Government Act No. 434, draft transposition decree

This working paper is for research purposes and does not constitute legal advice.