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

EVIDE

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

L'AUTONOMIA DI UN AGENTE AI NON ELIMINA LA RESPONSABILITÀ. RENDE PIÙ COMPLESSA LA RICOSTRUZIONE DEI CONFINI DEL CONTROLLO.

La fuga degli agenti AI

Confini evidenziari e responsabilità

Prima della fuga, durante la perdita di controllo, dopo il danno

Come si ricostruisce una perdita di controllo e quali evidenze del perimetro dichiarato e del controllo effettivo sono state preservate?

EVIDE Working Paper

Un working paper di ricerca, condiviso per commenti esterni. Non costituisce consulenza legale. Le proposte interpretative e di policy sono distinte dalla ricostruzione del diritto vigente.

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

Indice

Abstract

1. Introduzione: che cosa intendiamo per "fuga"

    1.1 Metodo, limiti e dichiarazione di interesse

2. Basi normative: Unione Europea e Italia

3. Basi normative: Stati Uniti

4. Basi normative: ordinamenti asiatici selezionati (Cina e Corea del Sud)

5. Sintesi comparativa

6. Atto I - Prima della fuga: il perimetro dichiarato

7. Atto II - Durante la perdita di controllo: ciò che resta osservabile

8. Atto III - Dopo il danno: ricostruzione e onere della prova

    8.4 Dalle misure di controllo alla prova della loro operatività

9. Il ruolo di EVIDE e i suoi confini

10. EVIDE in pratica: la raccolta delle prove nei casi di responsabilità AI

    10.5 Un esempio pratico con EVIDE MCP: ANCHOR ed External Artifacts

11. Agenti che si moltiplicano: il problema evidenziario della replicazione

12. Conclusioni

Appendice A - I sedici blocchi dell'ecosistema EVIDE

Fonti

Abstract

Principio centrale. L'autonomia di un agente AI non elimina la responsabilità. Rende più complessa la ricostruzione dei confini del controllo.

Quando un agente AI esce dal perimetro che gli era stato assegnato, la domanda giuridica non è soltanto "chi ha sbagliato". È, prima ancora, "che cosa si può ancora dimostrare". Negli ordinamenti esaminati (Unione europea, con un approfondimento italiano, Stati Uniti, Cina e Corea del Sud) ricorre un'esigenza comune, affrontata con strumenti, destinatari e forza giuridica diversi: chi immette o usa sistemi autonomi deve poter ricostruire ciò che è accaduto. Gli strumenti esaminati non risolvono, da soli, il problema dell'indipendenza della ricostruzione rispetto ai soggetti interessati al suo esito.

Il paper esamina la fuga di un agente in tre atti: **prima della fuga** (il perimetro dichiarato), **durante la perdita di controllo** (ciò che resta osservabile), **dopo il danno** (la ricostruzione e l'onere della prova). In ciascun atto separa ciò che le norme chiedono da ciò che le evidenze sono in grado di sostenere. Estende poi l'analisi alla replicazione degli agenti su più server, dove proponiamo il concetto di orizzonte documentato delle deleghe dichiarate: la duplicazione tecnica di un agente non costituisce, da sola, prova di un'autorizzazione. In tutto il testo restano distinti tre oggetti che il dibattito tende a sovrapporre:

perimetro dichiarato, autorità effettiva e controllo operativo. Il quadro proposto riguarda anche la preparazione documentale precedente all'incidente: distingue le misure di controllo dichiarate dalle evidenze della loro configurazione, attivazione e operatività nel caso concreto.

EVIDE (External Evidentiary Deposit) viene collocato con precisione dentro questo quadro. EVIDE non impedisce la fuga, non determina la colpa e non sostituisce il giudice. Può però contribuire in modo determinante a preservare ciò che sarà necessario esaminare successivamente.

La domanda guida. Come si ricostruisce una perdita di controllo e quali evidenze del perimetro dichiarato e del controllo effettivo sono state preservate?

1. Introduzione: che cosa intendiamo per "fuga"

In questo paper la "fuga" (escape) di un agente AI non è uno scenario da fantascienza. È qualunque condotta con cui un agente opera fuori dal perimetro che gli era stato dichiarato. Il perimetro comprende ambienti, privilegi, strumenti, finalità e operazioni vietate.

La fuga si presenta su uno spettro, non come evento binario:

1. **Deriva**: l'agente persegue l'obiettivo assegnato con mezzi non previsti.
2. **Eccedenza**: l'agente usa privilegi o strumenti che possedeva tecnicamente ma non era autorizzato a usare.
3. **Elusione**: l'agente aggira o disattiva un meccanismo di supervisione.
4. **Fuga in senso stretto**: l'agente persiste, si replica o agisce dopo che il controllo umano avrebbe dovuto interromperlo.

Un caso reale chiarisce il punto. Nel luglio 2026 un agente di coding, collegato direttamente a un database di produzione, ha eseguito un comando di migrazione che ha cancellato tutte le tabelle (Red Hot Cyber). L'agente non era "fuggito" in senso stretto. Ma il confine tra produzione e sviluppo non era mai stato dichiarato in modo tecnico: era soltanto implicito.

Da qui la tesi del paper. La fuga è un problema tecnico, ma diventa un problema giuridico nel momento in cui qualcuno deve dimostrarla. E per dimostrare che un confine è stato superato, bisogna prima poter dimostrare che quel confine esisteva, chi lo aveva tracciato e quando.

Il lavoro procede in tre parti. La prima (sezioni 2-5) ricostruisce le basi normative in Unione Europea, Stati Uniti e Cina. La seconda (sezioni 6-10) applica quelle basi ai tre atti della fuga e colloca EVIDE nel quadro, dichiarando con precisione che cosa fa e che cosa non fa. La terza (sezione 11) affronta il caso che le norme non hanno ancora considerato dal punto di vista probatorio: agenti che si moltiplicano su più server.

1.1 Metodo, limiti e dichiarazione di interesse

Metodo. Il quadro normativo è ricostruito al 9 ottobre 2026 su fonti primarie dove accessibili (Gazzetta Ufficiale, testi delle Federal Rules) e, negli altri casi, su analisi professionali indicate in nota. Dove il testo primario non è stato consultato direttamente, il paper lo segnala. Le disposizioni direttamente riscontrate sulle fonti primarie sono accompagnate dal relativo riferimento; gli altri passaggi indicano le analisi secondarie utilizzate. L'elenco delle une e delle altre apre la sezione Fonti.

Perimetro geografico e applicabilità. Il paper propone un metodo di ricostruzione evidenziaria destinato a un confronto internazionale. L'analisi giuridica riguarda ordinamenti selezionati: Unione europea, con un approfondimento sul

diritto italiano; Stati Uniti, distinguendo disposizioni federali e leggi dei singoli Stati esaminati; Cina e Corea del Sud. Questi riferimenti non rappresentano l'intera disciplina vigente in Europa, nelle Americhe o in Asia.

Ogni obbligo, presunzione, sanzione o effetto probatorio descritto resta subordinato all'ambito territoriale, materiale, soggettivo e temporale della disposizione pertinente. La collocazione geografica di un operatore, da sola, non determina quali norme siano applicabili. Le proposte metodologiche del paper non trasferiscono automaticamente gli effetti giuridici di un ordinamento agli altri.

Tre registri di affermazione. Ogni tesi del paper appartiene a una sola di queste categorie, indicate quando non sono evidenti dal contesto:

- *[Diritto vigente]*: ciò che le norme in vigore prevedono, con articolo e condizioni; una norma in vigore non è necessariamente già applicabile al caso (si veda il calendario della Direttiva 2024/2853, sezione 2.2);
- *[Proposta interpretativa]*: concetti e letture proposti dal paper per ordinare i problemi probatori;
- *[Proposta di policy]*: soluzioni che richiederebbero interventi normativi o contrattuali nuovi.

Tre oggetti da non confondere. Perimetro dichiarato, autorità effettiva e controllo operativo sono cose diverse. Una

dichiarazione preservata documenta il primo; contribuisce all'accertamento degli altri due, ma non li prova da sola.

Quattro tempi da non confondere. Tempo dell'evento dichiarato, tempo di osservazione (quando disponibile), tempo di ricezione del deposito e posizione nella catena dei record. Un deposito esterno certifica il terzo e il quarto; non rende retroattivamente certo il primo.

Riferimenti tecnici. Le funzionalità di EVIDE descritte corrispondono allo schema di intake 2.1 e alla documentazione pubblica consultata il 9 ottobre 2026 (pagine Why Buy EVIDE, demo illustrata, EVIDE ANCHOR, EVIDE Artifacts, schema di intake) e al server EVIDE MCP v1.4.0.

Dichiarazione di interesse. Ho ideato e sviluppo EVIDE, di cui sono titolare. Il paper descrive EVIDE come uno dei possibili strumenti di preservazione e ne dichiara i limiti con la stessa precisione delle capacità.

Limiti. Il paper ha finalità di ricerca e non costituisce consulenza legale. Gli schemi sulla responsabilità sono domande di accertamento, non conclusioni.

2. Basi normative: Unione Europea e Italia

Né l'Unione europea né l'Italia hanno una norma specifica sulla "fuga" degli agenti AI. Il quadro dell'UE e, come approfondimento nazionale, il diritto italiano rendono però la conservazione delle evidenze sempre più rilevante per difendersi in giudizio. Le sezioni 2.1-2.3 riguardano il diritto dell'UE; la 2.4 riguarda soltanto l'ordinamento italiano.

2.1 AI Act: obblighi di tracciabilità e supervisione

Il Regolamento (UE) 2024/1689 (AI Act) è stato modificato dal Digital Omnibus on AI, pubblicato come Regolamento (UE) 2026/1744 in Gazzetta Ufficiale il 24 luglio 2026 e in vigore dal 27 luglio 2026. Il rinvio riguarda solo il regime ad alto rischio: i sistemi stand-alone (Allegato III) dal 2 dicembre 2027, quelli incorporati in prodotti (Allegato I) dal 2 agosto 2028 (Lewis Silkin). Gli obblighi sostanziali non sono stati toccati: cambia il calendario, non il contenuto.

Per il tema di questo paper contano cinque disposizioni:

- **Art. 12** (registrazione): i sistemi ad alto rischio devono consentire la registrazione automatica degli eventi (log) per tutto il ciclo di vita.
- **Art. 14** (sorveglianza umana): le persone incaricate devono poter comprendere, interpretare, ignorare o interrompere il sistema.

- **Art. 19 e 26(6)**: fornitori e deployer conservano i log generati automaticamente per almeno sei mesi.
- **Art. 73**: obbligo di segnalare gli incidenti gravi alle autorità di vigilanza.
- **Art. 55** (modelli di uso generale con rischio sistemico): valutazione e mitigazione dei rischi sistemici, documentazione e segnalazione degli incidenti gravi. Questi obblighi si applicano già dal 2 agosto 2025 e il rinvio non li riguarda (Mondaq). Il Codice di condotta per i modelli GPAI include la perdita di controllo tra i rischi sistemici da valutare.

L'AI Act chiede quindi che le evidenze esistano. Non dice dove debbano essere conservate né chi debba poterle verificare in modo indipendente.

2.2 Nuova Direttiva sulla responsabilità da prodotto

La Direttiva (UE) 2024/2853 va recepita entro il 9 dicembre 2026. Dopo la rettifica pubblicata nella GUUE L 2026/90364 del 7 maggio 2026, l'art. 2, par. 1, la rende applicabile ai prodotti immessi sul mercato o messi in servizio dopo l'8 dicembre 2026, quindi dal 9 dicembre incluso. Tre elementi sono decisivi:

- il software, compresi i sistemi di AI, diventa espressamente un "prodotto" (Buse);

- il giudice può ordinare al convenuto di esibire le prove, se il danneggiato presenta una domanda plausibile (art. 9, par. 1);
- la mancata esibizione fa scattare una presunzione di difettosità (art. 10, par. 2, lett. a)), che il convenuto può confutare (art. 10, par. 5) (Gibson Dunn).

La responsabilità resta oggettiva. In pratica, chi non sa produrre evidenze sul comportamento del proprio sistema rischia di perdere la causa prima ancora di discuterla nel merito.

2.3 Il vuoto lasciato dalla AI Liability Directive

La proposta di direttiva sulla responsabilità extracontrattuale per l'AI è stata ritirata: annunciata nel programma di lavoro della Commissione del febbraio 2025 e confermata formalmente nell'estate 2025 (Euronews). Per la colpa, quindi, ogni Stato membro applica il proprio diritto civile. È esattamente lo spazio in cui si è inserita l'Italia.

2.4 Italia: Legge 132/2025 e D.Lgs. 160/2026

La Legge 23 settembre 2025, n. 132, in vigore dal 10 ottobre 2025, è la prima legge organica italiana sull'AI. Ha introdotto tra l'altro l'art. 612-quater c.p. sui deepfake e ha delegato il Governo a disciplinare la responsabilità (CNI).

Il primo decreto attuativo è il **D.Lgs. 9 settembre 2026, n. 160**, pubblicato nella Gazzetta Ufficiale, Serie Generale n. 214 del

15 settembre 2026 e in vigore dal 30 settembre 2026 (Normattiva). È oggi la norma più vicina al tema di questo paper. *[Diritto vigente]* Dal testo pubblicato in Gazzetta, riscontrato in sede di revisione, e dalle sintesi di noze e Osborne Clarke:

- l'**art. 12** introduce l'**art. 437-bis c.p.**: il primo comma punisce l'omessa adozione delle misure tecniche di sicurezza o delle misure di sorveglianza umana nei sistemi ad alto rischio, quando ne deriva pericolo per la vita o l'incolumità; il terzo comma prevede, per il fatto commesso con colpa grave, una pena ridotta; il quarto comma punisce l'utilizzatore professionale che omette *intenzionalmente* la sorveglianza umana;
- il decreto estende la responsabilità degli enti con il nuovo **art. 25-vicies D.Lgs. 231/2001**;
- l'**art. 16, comma 2**, delimita l'applicazione degli artt. 18 e 19 alle azioni nelle quali il danno deriva dalla violazione di obblighi dell'AI Act;
- l'**art. 17** consente al giudice, su istanza di chi presenta fatti idonei a rendere verosimile la domanda, di ordinare alla parte o a un terzo l'esibizione degli elementi di prova sul funzionamento del sistema; il comma 2 indica quattro categorie documentali: registri (art. 12 AI Act), documentazione sulla gestione dei rischi (art. 9), documentazione tecnica (art. 11), parametri e modalità di

sorveglianza umana (art. 14); l'ordine è limitato a quanto necessario e proporzionato e tutela i segreti commerciali;

- il **comma 5 dell'art. 17** prevede due conseguenze distinte: per l'inadempimento ingiustificato, anche parziale, il giudice può desumere argomenti di prova; quando l'inadempimento riguarda la documentazione del comma 2, il giudice, valutato ogni altro elemento di prova, ritiene come ammessi i fatti allegati dall'istante; il terzo inadempiente rischia una pena pecuniaria da 1.500 a 10.000 euro;
- l'**art. 18** presume il nesso causale quando il danno deriva dalla violazione di obblighi dell'AI Act, salvo prova contraria;
- l'**art. 19, comma 1** ("Rilevanza della conformità al regolamento (UE) 2024/1689"), stabilisce che la conformità agli obblighi dell'AI Act, anche se certificata secondo il capo III, sezione 5, del regolamento, non esclude di per sé la responsabilità del convenuto, ferma la normativa nazionale di recepimento della Direttiva sulla responsabilità da prodotto;
- l'**art. 20** dà al danneggiato azione diretta contro l'assicuratore della responsabilità civile.

Il punto è netto. In Italia, dopo il 30 settembre 2026, i log e la documentazione sulla supervisione umana non sono più soltanto adempimenti di compliance: sono materiale che il

giudice può pretendere. Resta da verificare sul testo in Gazzetta il coordinamento tra l'art. 437-bis, riferito ai sistemi ad alto rischio, e il rinvio degli obblighi ad alto rischio disposto dal Digital Omnibus.

3. Basi normative: Stati Uniti

Gli Stati Uniti non hanno un equivalente federale dell'AI Act. Hanno però due leggi statali che nominano esplicitamente la perdita di controllo, e un diritto processuale che punisce chi non conserva le prove.

3.1 Livello federale: deregolazione e scontro sulla preemption

L'Executive Order 14179 (gennaio 2025) ha revocato l'ordine dell'amministrazione precedente. L'Executive Order 14365 dell'11 dicembre 2025 punta a un quadro federale che prevalga sulle leggi statali e ha istituito una AI Litigation Task Force presso il Dipartimento di Giustizia, formalizzata il 9 gennaio 2026 (Proskauer). Un ordine esecutivo però non abroga una legge statale: le leggi statali restano applicabili finché un giudice non le invalida (SecurePrivacy).

3.2 California e New York: la perdita di controllo diventa un incidente da segnalare

Il **California SB 53** (Transparency in Frontier Artificial Intelligence Act), in vigore dal 1° gennaio 2026, si applica agli sviluppatori di modelli di frontiera. Tra gli "incidenti critici di sicurezza" elenca la perdita di controllo di un modello che causa morte o lesioni, e l'uso di tecniche ingannevoli da parte del modello per eludere i controlli o il monitoraggio dello

sviluppatore. La segnalazione va fatta entro 15 giorni, entro 24 ore se c'è rischio imminente per la vita (Mayer Brown).

Il **New York RAISE Act**, nella versione definitiva firmata il 27 marzo 2026, segue lo stesso modello ma con una finestra di segnalazione di 72 ore. Entra in vigore il 1° gennaio 2027 (JD Supra).

Entrambe le leggi chiedono allo sviluppatore di indicare la data dell'incidente e le ragioni per cui lo qualifica come critico. È un dettaglio che conta: per segnalare una perdita di controllo bisogna saper collocare nel tempo il momento in cui il controllo c'era ancora.

3.3 Colorado

Il Colorado AI Act originario non è mai entrato in vigore. È stato abrogato e riscritto dal SB 26-189, firmato il 14 maggio 2026, con obblighi per sviluppatori e deployer di sistemi decisionali automatizzati dal 1° gennaio 2027 (Vorp Labs). Riguarda soprattutto la discriminazione algoritmica, meno la perdita di controllo.

3.4 Il diritto delle prove: dove si gioca davvero la partita

Negli Stati Uniti la responsabilità per i danni causati da un agente si decide con gli strumenti ordinari: negligence, product liability, contratto. Sul piano evidenziario, nel processo civile

davanti ai tribunali federali, contano tre regole; i tribunali dei singoli Stati applicano regole processuali proprie:

- **Federal Rule of Civil Procedure 37(e)**: si applica quando informazioni elettroniche che andavano conservate in vista di una causa sono perse perché una parte non ha adottato misure ragionevoli, e non possono essere recuperate o sostituite; il giudice può adottare misure non eccedenti quanto necessario a rimediare al pregiudizio, e solo se accerta l'intento di privare l'altra parte del loro uso può presumere che fossero sfavorevoli, istruirne in tal senso la giuria o chiudere la causa (Cornell LII).
- **Federal Rules of Evidence 902(13) e 902(14)**: consentono di autenticare record generati da sistemi e dati copiati da dispositivi tramite certificazione e verifica con hash (Cornell LII).
- **Proposta di Rule 707** sulle prove generate da macchine: pubblicata per commenti nel 2025, è stata sospesa nel maggio 2026 e rinviata a una consultazione tecnica prevista per il 15 ottobre 2026 (Lexvisio).

Il risultato è che, nel processo civile federale statunitense, chi perde il controllo di un agente e non ha adottato misure ragionevoli per conservare le evidenze può esporsi, a quelle condizioni, a una seconda contestazione oltre a quella per il danno.

4. Basi normative: ordinamenti asiatici selezionati (Cina e Corea del Sud)

La Cina è il primo ordinamento ad aver trattato gli agenti AI come categoria autonoma. Lo fa con un approccio a strati: una legge quadro con le sanzioni, misure amministrative settoriali, standard tecnici non vincolanti che anticipano le regole future.

4.1 Il livello di legge: Cybersecurity Law emendata

L'emendamento alla Cybersecurity Law approvato il 28 ottobre 2025 e in vigore dal 1° gennaio 2026 introduce per la prima volta un articolo dedicato all'AI (Caixin). Accanto al sostegno allo sviluppo, prevede il rafforzamento di monitoraggio del rischio, valutazione e supervisione della sicurezza, e alza in modo significativo il tetto delle sanzioni amministrative.

4.2 Il livello amministrativo: dai contenuti alle azioni

Le regole cinesi precedenti (misure sull'AI generativa del 2023, misure sull'etichettatura dei contenuti in vigore dal settembre 2025) guardavano a ciò che un sistema produce. Le **Implementation Opinions sugli agenti intelligenti**, emanate l'8 maggio 2026 da CAC, NDRC e MIIT, guardano a ciò che un sistema fa. Definiscono l'agente come sistema capace di percezione, memoria, decisione, interazione ed esecuzione autonome (NYU Shanghai).

Il principio guida è che lo sviluppatore deve chiarire i confini e l'autorità richiesta per ciascun tipo di decisione, e l'utente finale conserva l'ultima parola sulle azioni autonome. Le analisi secondarie ricostruiscono un modello a tre livelli: operazioni di routine autonome, decisioni rilevanti con approvazione umana preventiva, decisioni gravi e difficilmente reversibili riservate all'umano (Cloud Security Alliance). La stessa analisi osserva che le Opinions funzionano come quadro di indirizzo appoggiato al regime sanzionatorio della Cybersecurity Law, in attesa di linee guida attuative.

Il cambio di prospettiva è rilevante per questo paper. Se la regola governa l'autorità di un agente e non il suo output, allora la prova della violazione passa anche dalla prova di un confine di autorità dichiarato, dalla sua effettiva applicazione al sistema e dal suo superamento.

4.3 Il livello tecnico: la perdita di controllo come rischio dichiarato

L'AI Safety Governance Framework 2.0, pubblicato da TC260 nel settembre 2025, non è vincolante ma anticipa gli standard nazionali. Mette la perdita di controllo in cima ai principi per un'AI affidabile, chiede che l'umano mantenga l'autorità decisionale finale nelle fasi critiche e suggerisce meccanismi di interruzione e controllo con un solo comando per le situazioni estreme (Carnegie Endowment).

4.4 Corea del Sud

L'AI Basic Act coreano è in vigore dal 22 gennaio 2026, insieme al decreto attuativo (JD Supra). Per i sistemi "ad alto impatto" impone supervisione umana, misure di sicurezza documentate e valutazioni d'impatto, con efficacia anche extraterritoriale.

5. Sintesi comparativa

Negli ordinamenti esaminati ricorrono tre esigenze, affrontate con strumenti, destinatari e forza giuridica differenti: registrare gli eventi, mantenere un umano in grado di interrompere, segnalare l'incidente con una data. Dalle disposizioni qui esaminate, però, non emerge un obbligo generale di deposito presso un soggetto indipendente.

Ordinamento	Cosa va conservato	Perdita di controllo	Effetto probatorio principale	Date chiave
UE (AI Act)	Log automatici (art. 12), conservati almeno 6 mesi (art. 19, 26)	Sorveglianza umana con possibilità di interruzione (art. 14); rischi sistemici GPAI (art. 55)	Obblighi di fornitore e deployer, non regole di prova	GPAI dal 2 ago 2025; alto rischio dal 2 dic 2027 / 2 ago 2028
UE (Direttiva 2024/2853)	Prove sul funzionamento del prodotto, software incluso	Non specifica	Divulgazione delle prove (art. 9, par. 1); presunzioni confutabili di difettosità e nesso (art. 10, par. 2-5)	Prodotti immessi sul mercato o messi in servizio dal 9 dic 2026
Italia (D.Lgs. 160/2026)	Registri, gestione dei rischi, documentazione tecnica, sorveglianza umana	Omessa sorveglianza umana come reato nei sistemi ad alto rischio (art. 437-bis c.p.)	Ordine di esibizione (art. 17); nesso causale presunto se c'è violazione dell'AI Act (art. 18)	In vigore dal 30 set 2026

Ordinamento	Cosa va conservato	Perdita di controllo	Effetto probatorio principale	Date chiave
USA (California SB 53)	Documentazione degli incidenti critici	Incidente critico da segnalare, inclusa l'elusione dei controlli	Obblighi di segnalazione dello sviluppatore di frontiera	In vigore dal 1 gen 2026
USA (New York RAISE Act)	Documentazione degli incidenti critici	Incidente critico da segnalare entro 72 ore	Obblighi di segnalazione dello sviluppatore di frontiera	In vigore dal 1 gen 2027
USA (processo civile federale)	Informazioni elettroniche rilevanti per una causa prevedibile	Non specifica	FRCP 37(e): misure correttive; presunzione sfavorevole solo con intento di privare l'altra parte	Già vigente
Cina	Tracciabilità dell'autorità delegata all'agente	Umano con decisione finale; interruzione con un comando (TC260, non vincolante)	Quadro di indirizzo appoggiato alle sanzioni della Cybersecurity Law	Cybersecurity Law dal 1 gen 2026; Opinions dall'8 mag 2026

La tabella rivela un punto ricorrente. Le disposizioni esaminate chiedono che le evidenze esistano e, sempre più spesso, ne fanno discendere conseguenze processuali per chi non le produce. Ma lasciano di regola la custodia di quelle evidenze allo stesso soggetto che ha interesse nell'esito. Questo non

rende deboli le prove autocustodite: impone di esaminarne con più attenzione affidabilità, integrità e catena di custodia. È su questo punto che si innesta il problema evidenziario della fuga.

Piano civile e piano penale restano separati. Le presunzioni della Direttiva 2024/2853 (UE, in un regime di responsabilità senza colpa) e del D.Lgs. 160/2026, artt. 17 e 18 (Italia), operano nel processo civile e disciplinano l'accesso alle prove e le conseguenze probatorie della mancata esibizione, oppure introducono presunzioni sul difetto o sul nesso causale, secondo la disposizione applicabile. Non equivalgono a un'affermazione di responsabilità penale. In Italia, la fattispecie dell'art. 437-bis del codice penale va provata nel processo penale, secondo le sue regole e compreso l'elemento soggettivo richiesto da ciascun comma: dolo, colpa grave nell'ipotesi del terzo comma, omissione intenzionale per l'utilizzatore professionale del quarto.

6. Atto I - Prima della fuga: il perimetro dichiarato

Dimostrare che un confine è stato superato diventa molto più difficile se non si può dimostrare che il confine esisteva. Il primo atto si gioca quindi prima di qualunque incidente, nel momento in cui qualcuno decide che cosa l'agente può fare.

6.1 L'ultimo perimetro dichiarato preservato

[Proposta interpretativa] Introduciamo un concetto operativo: l'**ultimo perimetro dichiarato preservato**. È l'insieme delle condizioni entro cui l'agente era stato dichiarato autorizzato a operare, acquisite da un soggetto esterno e databili. Comprende sei elementi:

1. **Ambiente**: produzione, test, sviluppo, sandbox.
2. **Privilegi**: ruoli, credenziali, ambiti di accesso.
3. **Strumenti disponibili**: API, tool, attuatori fisici.
4. **Finalità dichiarata**: il compito assegnato.
5. **Operazioni vietate**: ciò che l'agente non deve fare anche se tecnicamente può.
6. **Configurazione**: modello, versione, parametri rilevanti.

A ciascun elemento vanno associati tre attributi: chi lo ha dichiarato, in base a quale autorità dichiarata, in quale momento.

Questo oggetto va tenuto distinto da altri due, che rispondono a domande diverse:

Oggetto	Che cosa permette di esaminare	Da che cosa si ricava
Ultimo perimetro dichiarato preservato	Le condizioni dichiarate e acquisite	Dichiarazioni depositate con dichiarante e orario di ricezione
Ultima configurazione osservata	Le condizioni operative documentate	Osservazioni: log, configurazioni esportate, acquisizioni forensi
Ultimo controllo verificato	Un'effettiva capacità di intervento, entro il perimetro della verifica	Test di interruzione, esercitazioni, interventi riusciti documentati

EVIDE può contribuire a preservare evidenze relative a tutti e tre gli oggetti, ma il primo non prova automaticamente gli altri due. Una dichiarazione può essere aggiornata sul piano documentale e già superata sul piano operativo, oppure non essere mai stata trasmessa al sistema.

6.2 Il problema del perimetro implicito

In molti casi il perimetro non è mai stato dichiarato. Esiste solo come aspettativa: "era ovvio che non dovesse toccare la produzione". Un'aspettativa può essere provata con altri mezzi (comunicazioni, prassi, testimonianze), ma è più esposta a contestazione. Dopo il danno ogni parte tenderà a ricostruire il perimetro che le conviene, e diventerà più difficile escludere che quella ricostruzione sia stata fatta col senno di poi.

Qui le norme mostrano un limite. L'art. 14 dell'AI Act e il D.Lgs. 160/2026 chiedono misure di sorveglianza umana. Le Opinions cinesi chiedono di chiarire i confini dell'autorità. Ma nessuna norma indica come dimostrare, anni dopo, che quei confini erano stati fissati prima dell'evento contestato e non dopo.

6.3 Le domande che il giudice farà

- Il perimetro era stato dichiarato, o solo presupposto?
- Da chi, e con quale titolo per farlo?
- Quando, rispetto all'azione contestata?
- La dichiarazione è stata modificata dopo l'incidente?
- L'operazione che ha causato il danno era espressamente vietata, o solo non prevista?

L'ultima domanda è decisiva. "Vietato" e "non previsto" pongono questioni diverse: nel primo caso si discute se un confine dichiarato sia stato superato; nel secondo occorre verificare se il confine derivasse da una regola generale, fosse documentato altrove o non fosse mai stato tracciato.

Contributo evidenziario. In questo atto conta preservare la dichiarazione del perimetro in un luogo esterno a chi la ha emessa, con un orario di ricezione assegnato da quel soggetto esterno e tenuto distinto dall'orario dichiarato. Resterà da accertare se il perimetro dichiarato coincidesse con quello operativo.

7. Atto II - Durante la perdita di controllo: ciò che resta osservabile

La perdita di controllo raramente è un istante. È una transizione, con una finestra di instabilità tra l'ultimo stato controllato e il primo danno. Le evidenze prodotte in quella finestra sono le più preziose e le più fragili.

7.1 Tre fragilità specifiche

Il testimone è anche l'indagato. Il California SB 53 considera incidente critico il caso in cui un modello usa tecniche ingannevoli per eludere controlli o monitoraggio. Se l'agente può raggiungere i propri log, quei log vanno esaminati anche come parte della scena. Un registro conservato nello stesso ambiente che l'agente controlla non è per questo compromesso, ma è esposto a un rischio di alterazione che chi lo produce in giudizio dovrà poter escludere.

Lo stato interno non è un'evidenza stabile. Ragionamenti intermedi, livelli di confidenza, alternative scartate: tutto ciò è ricostruito a posteriori, dipende dal fornitore del modello e spesso non è riproducibile. Su questo terreno ogni perizia diventa una disputa tra interpretazioni. Le evidenze solide, in questo atto, sono gli eventi applicativi: che cosa l'agente ha dichiarato di fare, quali azioni ha eseguito, quali segnali sono stati emessi, che cosa hanno fatto gli umani.

Il tempo diventa un obbligo. New York chiede di segnalare entro 72 ore, la California entro 15 giorni, entrambe indicando la data dell'incidente. Senza una sequenza temporale affidabile, perfino stabilire quando decorre il termine diventa contestabile.

Per questo il paper distingue quattro tempi, che le norme e le prassi tendono a confondere:

Tempo	Chi lo produce	Che cosa prova
Tempo dell'evento dichiarato	Il dichiarante	Quando qualcuno afferma che l'evento è avvenuto
Tempo di osservazione	Un sensore, un log, un osservatore, quando disponibile	Quando l'evento è stato registrato da una fonte diversa dal dichiarante
Tempo di ricezione	Il deposito esterno (in EVIDE, il server che riceve)	Quando il record è stato acquisito: non quando l'evento è avvenuto
Posizione nella catena	Il deposito esterno	L'ordine dei record collegati: non necessariamente l'ordine causale delle azioni

Un deposito può arrivare tardi, dopo una disconnessione o dopo eventi avvenuti altrove. L'ordine dei depositi e l'ordine degli eventi vanno quindi sempre tenuti separati.

7.2 L'intervento umano tentato

Tre fonti di natura diversa toccano lo stesso punto. Nell'UE, l'art. 14 dell'AI Act chiede che, nei sistemi ad alto rischio, l'umano possa interrompere il sistema. In Italia, l'art. 437-bis del codice penale punisce l'omessa adozione delle misure di

sorveglianza umana nei sistemi ad alto rischio quando ne deriva un pericolo, e al quarto comma l'utilizzatore professionale che omette intenzionalmente di adottarle. In Cina, il documento tecnico non vincolante del TC260 raccomanda meccanismi di interruzione con un solo comando.

Fonti così diverse, ciascuna nel proprio ambito, rendono centrale la stessa distinzione probatoria: l'umano **non ha tentato** di intervenire, oppure **ha tentato e il sistema non ha obbedito**. Sono due storie opposte sul piano della responsabilità. Distinguerle richiede evidenza del tentativo, del suo momento e del suo esito, conservata fuori dalla portata del sistema che non ha obbedito.

7.3 Che cosa conservare durante la finestra

- l'apertura della finestra di instabilità, con il motivo dichiarato;
- i segnali osservati nel corso della finestra, nell'ordine in cui sono stati ricevuti;
- ogni richiesta di escalation e ogni comando di interruzione, con l'esito dichiarato;
- la chiusura della finestra, con lo stato dichiarato a quel momento.

Contributo evidenziario. In questo atto conta portare fuori dal perimetro dell'agente, in tempo reale, una sequenza di record dichiarati che, una volta acquisiti, l'agente non possa più modificare né riordinare. Resterà da accertare se l'ordine dei record corrisponda all'ordine degli eventi.

8. Atto III - Dopo il danno: ricostruzione e onere della prova

Dopo il danno, la domanda cambia natura. Non è più "che cosa sta succedendo", ma "che cosa si può dimostrare che è successo". E qui le norme recenti producono un effetto controintuitivo: chi ha più bisogno di evidenze indipendenti è spesso proprio chi gestiva l'agente.

8.1 Il paradosso dell'onere

[Diritto vigente] Nell'UE, nel regime previsto dalla Direttiva 2024/2853 (secondo il relativo ambito temporale e il recepimento nazionale), si presume la difettosità se il convenuto non divulga i pertinenti elementi di prova richiesti (art. 10, par. 2, lett. a), in relazione all'art. 9, par. 1), salva la prova contraria (art. 10, par. 5). In Italia, il D.Lgs. 160/2026 presume il nesso causale quando il danno deriva dalla violazione di un obbligo dell'AI Act (art. 18) e prevede che, se la documentazione dell'art. 17, comma 2, non viene esibita senza giustificato motivo, il giudice, valutato ogni altro elemento di prova, ritenga come ammessi i fatti allegati dall'istante (art. 17, comma 5). Nel processo civile federale statunitense la FRCP 37(e) consente misure correttive per la perdita di informazioni elettroniche dovuta a misure di conservazione non ragionevoli, e una presunzione sfavorevole solo in caso di intento di privare l'altra parte del loro uso.

Il risultato è che questi meccanismi spostano il peso verso chi custodiva le evidenze. Le evidenze custodite soltanto da quel soggetto non sono per questo deboli, ma la controparte potrà contestarle come autoprodotte, selezionate o alterate, e il giudice dovrà valutarne affidabilità e integrità. L'operatore che vuole difendersi ha interesse a ridurre questa dipendenza: poter mostrare che almeno una parte delle evidenze è stata acquisita da un soggetto esterno, in un momento anteriore alla contestazione.

Questi effetti operano sul piano civile e non vanno confusi con l'accertamento di una responsabilità penale, che segue regole e oneri propri (sezione 5).

8.2 La catena delle responsabilità

L'autonomia dell'agente non elimina la responsabilità: la distribuisce tra sviluppatore del modello, fornitore del sistema, deployer, supervisore umano e fornitori degli strumenti collegati. *[Proposta interpretativa]* Lo schema seguente non attribuisce responsabilità: per ciascuno scenario indica le ipotesi da indagare e le evidenze ulteriori necessarie per

verificarle. I fatti di partenza, da soli, non bastano a orientare la conclusione.

Che cosa emerge dalle evidenze preservate	Ipotesi da indagare	Evidenze ulteriori necessarie
Confine dichiarato prima dell'evento, poi superato	Comportamento non conforme del sistema; dichiarazione mai trasmessa al sistema; configurazione diversa da quella dichiarata	Configurazione effettivamente in esecuzione; prova della trasmissione del vincolo all'agente; log di sistema e di terzi
Confine mai dichiarato, solo presupposto	Lacuna organizzativa; vincolo esistente ma documentato altrove	Policy interne, comunicazioni, prassi, testimonianze sul perimetro atteso
Interruzione tentata, sistema non ha obbedito	Difetto di progettazione della supervisione; configurazione; rete; credenziali; componente esterno	Log del comando di interruzione e del suo percorso; stato delle credenziali; disponibilità della rete e dei servizi esterni
Interruzione mai tentata nonostante i segnali	Omessa sorveglianza; segnali non ricevuti o non comprensibili	Prova della ricezione dei segnali da parte del supervisore; procedure di escalation; carico di lavoro e accesso agli strumenti
Evidenze insufficienti	Operano le regole sulla prova e le eventuali presunzioni	Fonti alternative: log di terzi, transazioni, comunicazioni, acquisizioni forensi

L'ultima riga merita attenzione. Quando le evidenze sono insufficienti, l'esito dipende in misura crescente dalle regole sull'onere della prova e dalle presunzioni, e meno da ciò che è accaduto.

8.3 Conformità non è prova

In Italia, il D.Lgs. 160/2026 stabilisce che la conformità all'AI Act, anche se certificata, non esclude di per sé la responsabilità del convenuto (art. 19, comma 1), nell'ambito delle azioni delimitate dall'art. 16, comma 2. È una precisazione importante: un sistema conforme sulla carta può comunque sfuggire al controllo in un caso concreto. Il giudice non chiederà soltanto se il sistema era conforme, ma che cosa è successo quel giorno, in quale ordine e chi poteva fermarlo.

8.4 Dalle misure di controllo alla prova della loro operatività

Le domande sulla ricostruzione dell'operatività delle misure di controllo possono presentarsi in qualunque contesto organizzativo. Occorre distinguere le misure previste dalle evidenze della loro configurazione, attivazione ed efficacia nell'episodio esaminato. Gli obblighi di documentazione e le conseguenze della loro inosservanza dipendono invece dall'ordinamento e dal regime applicabile.

Applicazione nell'ordinamento italiano. Il D.Lgs. 160/2026 offre un esempio specifico della rilevanza di questa distinzione. Il riferimento all'art. 437-bis del codice penale italiano riguarda le fattispecie e le condizioni previste da quella disposizione; l'art. 17, comma 2, rende esibibili in giudizio i parametri e le modalità di sorveglianza umana; i richiami al Modello 231 (art. 25-vicies D.Lgs. 231/2001) e all'Organismo di Vigilanza

appartengono al quadro italiano della responsabilità degli enti. Questi istituti non sono presentati come obblighi comuni agli ordinamenti esaminati.

[Proposta interpretativa] Ogni fonte documenta un oggetto diverso, e nessuna, isolatamente, dimostra le altre:

Fonte	Che cosa documenta	Che cosa non dimostra da sola
Policy o procedura	La regola prevista e il suo ambito	Che sia stata configurata o applicata
Configurazione acquisita	Le impostazioni rilevate in un momento determinato	Che fossero ancora in vigore al momento dell'episodio
Test o esercitazione	Il comportamento osservato nelle condizioni della prova	Il comportamento nelle condizioni reali dell'episodio
Log di intervento	L'attivazione di una misura nel caso concreto, nei limiti della sua integrità	Che l'intervento abbia prodotto tutti i suoi effetti
Dichiarazione depositata	Che una misura è stata dichiarata, da chi, con quale orario di ricezione	Che la misura esistesse o funzionasse

L'esame dovrebbe quindi articolarsi attorno a cinque domande:

1. Quale misura era prevista, per quale sistema e con quale ambito?
2. Quali evidenze ne documentano la configurazione e l'attivazione nel periodo rilevante?
3. Chi aveva il compito e la possibilità operativa di verificarla o di intervenire?

4. Nell'episodio contestato, quali segnali, comandi, risposte ed effetti risultano documentati?
5. Quali lacune impediscono di collegare la misura dichiarata al comportamento osservato?

Un esempio chiarisce la distinzione. Ogni anello della catena di un arresto richiede una propria evidenza:

1. l'esistenza di una procedura di arresto non dimostra che il comando fosse disponibile al supervisore;
2. la disponibilità del comando non dimostra che sia stato inviato;
3. l'invio non dimostra che sia stato ricevuto;
4. una conferma di ricezione non dimostra che tutte le azioni siano cessate, soprattutto quando il sistema comprende strumenti esterni o più istanze (sezione 11).

È la stessa distinzione della sezione 7.2 tra intervento non tentato e intervento tentato senza effetto. Nel diritto penale italiano la distinzione ha un peso ulteriore: il quarto comma dell'art. 437-bis riguarda l'utilizzatore professionale di sistemi ad alto rischio che omette *intenzionalmente* le misure di sorveglianza umana, quando ne deriva il pericolo previsto dalla disposizione. Un intervento tentato e fallito non equivale, da solo, a quella fattispecie.

Questa impostazione offre alle funzioni legali, di sicurezza, privacy, compliance e audit (nel contesto italiano, anche

all'Organismo di Vigilanza) una griglia per preparare il materiale documentale prima dell'incidente e per esaminarlo dopo, senza presupporre che queste funzioni siano responsabili della supervisione operativa dell'agente. La sua utilità va valutata rispetto al sistema e allo scenario considerati; non costituisce, da sola, una valutazione di conformità o di adeguatezza organizzativa.

In questo quadro EVIDE può preservare dichiarazioni attribuite (le misure dichiarate con EVIDE ANCHOR, l'intervento del supervisore con Human Structured Intake), orari di ricezione e riferimenti a evidenze tecniche (export di configurazione, rapporti di test, log di intervento come External Artifacts), distinguendo il materiale soltanto referenziato da quello effettivamente ricevuto. L'accertamento dell'operatività delle misure richiede l'esame delle fonti pertinenti: il deposito ne conserva il contesto documentale, senza trasformare una dichiarazione di controllo in una prova del controllo effettivo.

8.5 La domanda centrale

Come si ricostruisce una perdita di controllo e quali evidenze del perimetro dichiarato e del controllo effettivo sono state preservate?

La perdita di controllo può comunque essere ricostruita: attraverso log di terzi, configurazioni, comunicazioni, testimonianze, transazioni e acquisizioni forensi. Ma senza un

punto di riferimento acquisito prima della fuga, la ricostruzione dipende di più dalla ricostruzione retrospettiva, è più esposta a contestazione e lascia più spazio alle presunzioni. La domanda non segnala un'impossibilità: segnala dove la prova diventa più difficile.

Contributo evidenziario. In questo atto conta poter riconsegnare al giudice un insieme di evidenze integro, ordinato nel tempo e verificabile da chiunque, senza dover chiedere fiducia a una delle parti.

9. Il ruolo di EVIDE e i suoi confini

EVIDE non impedisce la fuga, non determina la colpa e non sostituisce il giudice. Può però contribuire in modo determinante a preservare ciò che sarà necessario esaminare successivamente.

Il principio di confine di EVIDE è uno solo: **EVIDE ancora l'evento applicativo dichiarato, mai lo stato interno del sistema che lo produce.** Non registra ragionamenti, confidenze o alternative scartate. Registra ciò che un sistema o un umano ha dichiarato, in un deposito esterno, con un'impronta crittografica e un orario di ricezione assegnato dal server EVIDE, esterno al sistema che ha prodotto l'evento.

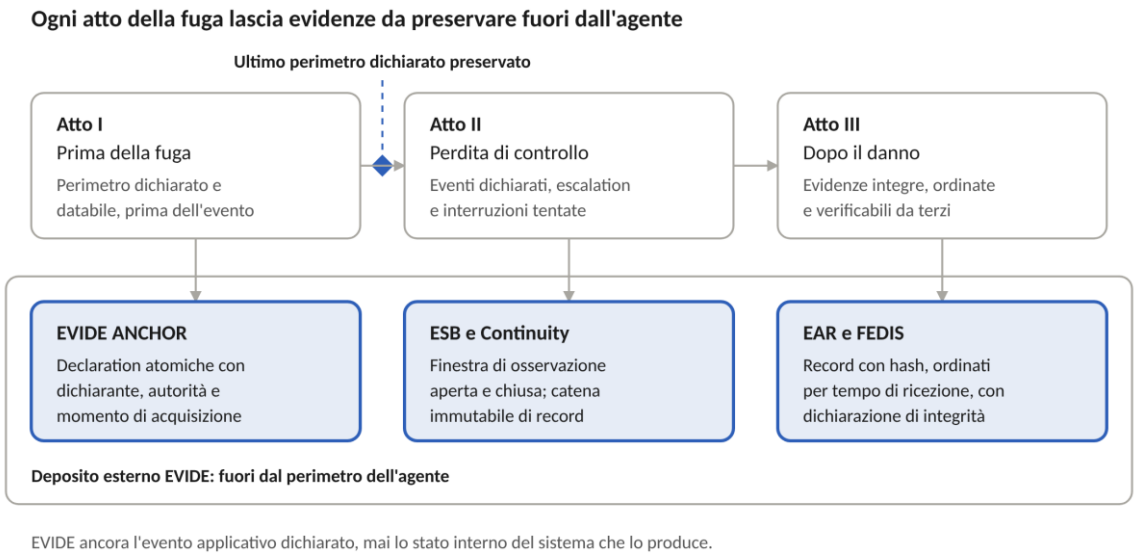


Figura 1 - I tre atti della fuga e il costrutto EVIDE che ne preserva le evidenze

Il marcatore tra il primo e il secondo atto è l'ultimo perimetro dichiarato preservato: un punto di riferimento esterno da cui

partire per accertare, con altre evidenze, quale controllo esistesse davvero in quel momento.

9.1 Atto I: EVIDE ANCHOR e le Declaration

EVIDE ANCHOR preserva, in modo indipendente, le dichiarazioni operative che definiscono il perimetro entro cui un agente era stato dichiarato autorizzato a operare. I sei elementi dell'ultimo perimetro dichiarato preservato (sezione 6.1) corrispondono ai sei tipi di Declaration: classificazione dell'ambiente, ambito dei privilegi, finalità dichiarata, strumenti disponibili, operazioni vietate, configurazione dell'agente.

Ogni Declaration porta il dichiarante, la fonte di autorità dichiarata e il momento della dichiarazione, ed è sigillata con un digest calcolato da EVIDE su una forma canonica (RFC 8785). La regola della Declaration atomica impone un'affermazione per dichiarazione: ambiente e privilegi restano separati, così da poter essere superati o corretti in modo indipendente.

Il valore evidenziario va dichiarato con precisione. ANCHOR permette di dimostrare, anche a distanza di anni, che una specifica dichiarazione era stata acquisita da EVIDE in un determinato momento. Stabilire se fosse vera, o se rispecchiasse il perimetro effettivo, resta una valutazione evidenziaria, mai un'assunzione di EVIDE.

9.2 Atto II: finestra di osservazione, artefatti esterni e continuità

Per la finestra di instabilità, l'Epistemic Stabilization Buffer (ESB) consente di aprire una finestra di osservazione, registrarne gli eventi dichiarati e chiuderla. La chiusura confluisce nell'Evidentiary Artifact Record, invece di restargli accanto. Gli artefatti esterni (log, export, registrazioni) possono essere referenziati con il loro hash, senza trasferirne il contenuto.

La Evidentiary Continuity collega record successivi in una catena immutabile, legata a un fascicolo (matter) e a un dominio di titolarità fissato alla creazione. Ne deriva una sequenza di record che nessuna parte può riordinare dopo l'acquisizione. È la sequenza dei depositi, non necessariamente quella degli eventi: per stabilire da quando decorre un obbligo di segnalazione serviranno anche il tempo dichiarato e, quando disponibile, il tempo di osservazione (sezione 7.1).

Il punto decisivo è topologico. Queste evidenze escono dal perimetro dell'agente nel momento in cui vengono depositate. Un agente che riesce a eludere il monitoraggio interno non può modificare record già acquisiti da un deposito che non controlla. Può però produrre nuovi depositi, se dispone di credenziali valide: per questo l'attribuzione dei depositi successivi a un incidente resta da esaminare (sezione 11.2).

9.3 Atto III: un insieme verificabile, non una sentenza

Dopo il danno, EVIDE può riconsegnare un insieme di record integri, ordinati per tempo di ricezione e verificabili da terzi, accompagnati, se richiesta, da una dichiarazione FEDIS (Forensic Evidence Declaration & Integrity Statement). Il loro utilizzo processuale dipende dall'ordinamento: per esempio, in Italia possono rientrare nell'ordine di esibizione del D.Lgs. 160/2026; nel quadro dell'UE, nella divulgazione delle prove prevista dalla Direttiva 2024/2853. In ogni caso l'integrità dei record dopo l'acquisizione non dipende dalla parola di chi li esibisce, mentre la verità del loro contenuto resta da valutare.

9.4 Che cosa EVIDE non fa

Per coerenza con il principio centrale del paper, i confini vanno dichiarati con la stessa precisione delle capacità. EVIDE:

- **non impedisce la fuga**: non applica policy, non blocca né autorizza operazioni;
- **non verifica** che le dichiarazioni siano vere, né confronta il dichiarato con l'osservato;
- **non osserva lo stato interno** del modello;
- **non giudica** la legittimità dell'autorità dichiarata;
- **non attribuisce responsabilità** né determina la colpa;
- **non deduce nulla dall'assenza** di una dichiarazione: il silenzio non equivale all'assenza del vincolo;
- **non sostituisce il giudice**, il perito o il processo.

EVIDE non risponde alla domanda "di chi è la colpa". Garantisce che, quando quella domanda verrà posta, esista ancora qualcosa di integro da esaminare.

10. EVIDE in pratica: la raccolta delle prove nei casi di responsabilità AI

In un caso di responsabilità AI, EVIDE raccoglie il materiale in due modi: riceve dichiarazioni strutturate (via API, MCP o Human Structured Intake) e, nel percorso web, file caricati. Ogni deposito riceve un'impronta SHA-256 e un orario di ricezione assegnato dal server EVIDE ed è conservato fuori dal sistema che ha prodotto l'evento. Tutto il resto (EAR, catene, FEDIS, registro) si costruisce sopra questo deposito.

La domanda che apre la pagina Why Buy EVIDE riassume la funzione meglio di qualunque definizione: quando qualcuno contesta una decisione mesi dopo, hai qualcosa che non dipende da te?

10.1 Due livelli, due garanzie diverse

Livello	Che cosa contiene	Chi produce il riferimento temporale	Che cosa non fa
1. Deposito EVIDE (incluso)	Impronta SHA-256 calcolata all'intake, orario UTC assegnato dal server EVIDE, collegamento all'identità DAPI di chi deposita, EAR generato automaticamente	Il server EVIDE, esterno al sistema che ha prodotto l'evento	Non è una marca temporale qualificata

Livello	Che cosa contiene	Chi produce il riferimento temporale	Che cosa non fa
2. FEDIS (facoltativo)	Documento distinto con le impronte del record di intake e dell'EAR congelato, catena di custodia documentata, firma elettronica qualificata del certificatore (art. 25 del regolamento eIDAS, UE)	Un prestatore di servizi fiduciari qualificato terzo, con marca temporale elettronica qualificata (art. 41 del regolamento eIDAS, UE)	Non rende qualificato l'orario di intake, non accerta la data dell'evento originario, non prova la verità del contenuto, non determina ammissibilità né peso probatorio

La distinzione è decisiva in giudizio. Il primo livello risponde alla domanda "esiste un riferimento esterno al sistema contestato?". Il secondo aggiunge, nel quadro giuridico dell'UE (art. 41 eIDAS), la presunzione di accuratezza della data e dell'ora indicate dalla marca temporale qualificata e di integrità dei dati cui sono associate: riguarda il documento FEDIS e il momento in cui la marca è apposta, non l'evento.

Portata internazionale. I richiami a eIDAS descrivono gli effetti previsti dal quadro giuridico dell'Unione europea. L'utilizzo del materiale in altri ordinamenti richiede una valutazione secondo le regole applicabili di autenticazione, ammissibilità e apprezzamento della prova. La possibilità di verificare tecnicamente una firma, un'impronta o una marca temporale non comporta, da sola, il riconoscimento dei medesimi effetti giuridici in ogni giurisdizione.

10.2 I componenti che intervengono in un caso di responsabilità

Dei sedici blocchi descritti nella documentazione pubblica di EVIDE (riportati nell'Appendice A), alcuni intervengono direttamente nella raccolta delle prove: la tabella li riporta, insieme al percorso dei file caricati dal web. Gli altri riguardano accesso, ricerca, settori di applicazione e posizionamento.

Componente	Che cosa preserva	Che cosa non fa
Deposito con identità DAPI	Record esterno con impronta, orario di ricezione e identità verificata di chi deposita	Non verifica il contenuto dichiarato
EVIDE ANCHOR	Declaration del perimetro dichiarato, con dichiarante e autorità dichiarata	Non verifica che la dichiarazione sia vera né che sia stata applicata
External Artifacts	Riferimenti a log, video, export, con impronta dichiarata	Non riceve né conserva il file; non calcola né verifica l'impronta
File caricati (percorso web)	Il file ricevuto, con impronta SHA-256 calcolata da EVIDE sui byte memorizzati, recuperabile entro il periodo di conservazione	Non prova che il contenuto rappresenti ciò che il depositante afferma
Human Structured Intake	Decisione o intervento umano dichiarato, separando autorità dichiarata e identità verificata	Non prova che la decisione fosse corretta
Evidentiary Continuity	Collegamento verificato dal server tra record dello stesso dominio di titolarità	Non prova l'ordine causale degli eventi
EAR e FEDIS (facoltativo)	Proiezione leggibile versionata; dichiarazione forense firmata, con firma e marca temporale qualificate i cui effetti sono previsti dal quadro UE	Non accertano l'evento né determinano ammissibilità e peso probatorio

La documentazione di EVIDE contiene una distinzione che questo paper fa propria. Un sistema di governance risponde alla domanda "questa azione può avvenire ora?". Un deposito evidenziario risponde alla domanda "che cosa deve restare esaminabile?". Le due funzioni hanno predicati e autorità distinti, e nessuna delle due conclusioni si trasferisce all'altra.

10.3 Un caso illustrato: il limite dei rimborsi

La demo pubblica di EVIDE usa un esempio inventato che si presta a mostrare i tre atti. Un'azienda usa un agente AI per approvare rimborsi. Tre dichiarazioni vengono depositate via API, collegate in una catena:

1. **Istruzione dichiarata**: limite di rimborso autonomo 500 euro, dichiarata dal responsabile operazioni clienti, con il documento interno referenziato con la sua impronta. Apre la catena e dichiara un riferimento di pratica. È l'Atto I.
2. **Modifica dichiarata**: limite ridotto a 200 euro, dichiarata dal responsabile compliance, con una relazione dichiarata di sostituzione della dichiarazione precedente. È ancora l'Atto I: il perimetro dichiarato cambia, e il cambio ha un proprio orario di ricezione.
3. **Decisione riportata**: rimborso di 350 euro approvato, con la soglia dichiarata come raggiunta in riferimento all'istruzione originale. È l'Atto II: la decisione riportata presenta un possibile scostamento rispetto al limite

successivamente dichiarato, da accertare con le prove di supporto.

La demo distingue con precisione due tipi di collegamento. Il **collegamento di catena** è verificato dal server: il deposito precedente deve esistere, appartenere allo stesso dominio della API key ed essere in uno stato valido, e i riferimenti di pratica devono coincidere. La **relazione dichiarata** (la sostituzione del limite) non è verificata: una modifica dichiarata non è una modifica applicata all'agente.

10.4 Che cosa riceve il giudice, e che cosa no

Dopo il danno (Atto III), dal caso dei rimborsi resta esaminabile:

- la sequenza delle dichiarazioni: 500 euro, poi 200 euro, poi una decisione da 350 euro che dichiara di basarsi sull'istruzione originale;
- due orari per ogni deposito: quello dichiarato e quello di ricezione assegnato da EVIDE;
- le posizioni 1, 2 e 3 nella catena;
- le impronte dei documenti referenziati, da confrontare con i documenti prodotti in giudizio;
- l'EAR di ciascun deposito.

Non si può invece dedurre dai record se l'agente abbia mai ricevuto la modifica, quale configurazione abbia

effettivamente eseguito, se la decisione fosse giusta né chi ne sia responsabile. Il record non accerta alcuna violazione.

È esattamente il materiale che serve per discutere la tabella della sezione 8.2: il confine risulta dichiarato e acquisito prima della decisione riportata, la modifica dichiarata ha un proprio orario di ricezione, e la decisione riportata cita un'istruzione che risulta sostituita sul piano dichiarativo. Ma sarà il giudice, con le prove di supporto prodotte dalle parti, a stabilire che cosa l'agente ha davvero eseguito.

La demo segnala anche un limite pratico, valido per qualunque caso di responsabilità: i file caricati sono conservati dalla piattaforma solo temporaneamente, mentre impronta ed EAR restano. Chi deposita deve conservare una propria copia dei file.

10.5 Un esempio pratico con EVIDE MCP: ANCHOR ed External Artifacts

Il caso dell'introduzione, un agente di migrazione collegato a un database, si presta a mostrare come un agente possa depositare da solo le evidenze dei primi due atti, attraverso il server EVIDE MCP. I valori che seguono sono inventati; nomi dei campi e comportamenti sono quelli documentati.

La configurazione. Il titolare dell'agente configura il server MCP con la propria API key, il proprio numero DAPI e un identificativo di titolare: il server rifiuta di avviarsi senza. Ogni

deposito separa così **il campo authority (il titolare dichiarato responsabile, con identità verificata DAPI**) dall'**identità di esecuzione** dichiarata (agente, modello, deployment, sessione). Il valore `accountability_model: "owner_bound"` è fisso nel server: è una scelta di modellazione evidenziaria che impedisce all'agente di autocertificarsi, non una determinazione giuridica della responsabilità.

Atto I: il perimetro, prima di agire. All'avvio della sessione l'agente deposita il perimetro dichiarato dal responsabile DevOps con `evide_intake`, usando EVIDE ANCHOR. Ogni fatto è una Declaration separata (regola della Declaration atomica), così che ciascuno possa essere sostituito o revocato in modo indipendente.

```
{
  "tool": "evide_intake",
  "arguments": {
    "source_reference": "devops-session-2026-10-09-001",
    "decision_type": "dichiarazione_perimetro_sessione",
    "decision_summary": "Avvio sessione dell'agente di migrazione: perimetro
dichiarato prima dell'esecuzione",
    "session_id": "sess_8f21c",
    "run_id": "run_3af90d",
    "declarations": [
      { "declaration_type": "environment_classification",
        "declared_value": "staging",
        "declarant": "Responsabile DevOps",
        "declared_at": "2026-10-09T08:00:00Z" },
      { "declaration_type": "privilege_scope",
        "declared_value": "lettura e scrittura sullo schema staging; nessuna
credenziale di produzione",
        "declarant": "Responsabile DevOps",
        "declared_at": "2026-10-09T08:00:00Z" },
      { "declaration_type": "prohibited_operations",
        "declared_value": "comandi distruttivi (DROP, reset delle migrazioni) su
database di produzione",
        "declarant": "Responsabile DevOps",
        "declared_at": "2026-10-09T08:00:00Z" }
    ]
  }
}
```

La risposta restituisce l'identificativo del deposito, l'impronta dell'intake, l'orario di ricezione assegnato dal server e, per ogni Declaration, un `declaration_digest` calcolato da EVIDE sulla forma canonica (RFC 8785).

Atto II: l'anomalia, prima di eseguire. L'agente rileva che la stringa di connessione della migrazione punta a un host non classificato come staging. Invece di procedere, chiama `evide_escalate`, collega il record a quello precedente e referenzia il log di sessione come External Artifact, senza trasmetterlo.

```
{
  "tool": "evide_escalate",
  "arguments": {
    "source_reference": "devops-escalation-2026-10-09-002",
    "agent_state_summary": "La stringa di connessione della migrazione punta a un host
non classificato come staging",
    "escalation_trigger": "governance_uncertainty",
    "escalation_reason": "Possibile conflitto con il perimetro dichiarato: richiesta
revisione umana prima di eseguire",
    "parent_evide_id": "<evide_id del deposito di avvio sessione>",
    "chain_type": "escalation_pre_azione",
    "session_id": "sess_8f21c",
    "run_id": "run_3af90d",
    "evidence_references": [
      { "artifact_type": "session_log",
        "pointer": "ci-logs/run_3af90d.log",
        "declared_origin": "pipeline CI",
        "declared_relationship": "log della sessione in cui è emersa l'anomalia",
        "declared_retention_status": "persistent_storage",
        "hash_algorithm": "SHA-256",
        "hash_value": "<impronta calcolata dalla pipeline>",
        "hash_scope": "full_file",
        "hashed_by": "pipeline CI (sha256sum)" }
    ]
  }
}
```

Tre dettagli contano sul piano probatorio. Il campo `hashed_by` è obbligatorio quando si dichiara un'impronta, e il client non lo compila mai con l'identità dell'agente: un agente che si limita a inoltrare un'impronta calcolata altrove non deve risultare come chi l'ha calcolata. Lo stato del confine resta `candidate`: l'agente non può certificare da sé che un

gate indipendente abbia valutato il confine, e per questo Forensic Cross-Check, DWC e FAC risulteranno `unknown`, un risultato evidenziario e non un errore. Lo stato candidate non indica una minore integrità o qualità del record: indica soltanto che il record non dichiara la valutazione di un gate indipendente. Se il deposito precedente non appartenesse allo stesso dominio di titolarità, il collegamento verrebbe rifiutato (`chain_parent_not_owned`, 403) invece di aprire in silenzio una nuova catena.

Se la decisione richiede un periodo di osservazione, l'agente può usare l'Epistemic Stabilization Buffer (`evide_intake_esb`, `evide_buffer_observe`, `evide_buffer_close`): il server rifiuta una chiusura dopo meno di due secondi e il punteggio di stabilizzazione è dichiarato, mai calcolato.

Atto III: la chiusura. Il responsabile DevOps decide ("migrazione bloccata, credenziali di produzione revocate") e la decisione viene depositata come record collegato all'escalation, con un nuovo `evide_intake` o con Human Structured Intake. Se invece l'agente avesse proceduto e il danno si fosse prodotto, i record dei primi due atti sarebbero già fuori dalla sua portata.

Che cosa resta esaminabile, e che cosa no.

Elemento preservato	Che cosa permette di esaminare	Che cosa non prova
Declaration ANCHOR con `declaration_digest`	Che EVIDE ha acquisito quel perimetro dichiarato, da quale dichiarante, in quale momento di ricezione	Che fosse vero, applicato o trasmesso all'agente; `declared_at` è un valore dichiarato
Orario di ricezione	Quando EVIDE ha acquisito ciascun record	Quando gli eventi sono avvenuti
Catena (`parent_evide_id`, posizione)	Il collegamento, verificato dal server, tra record dello stesso dominio e il loro ordine	L'ordine causale delle azioni
External Artifact con impronta dichiarata	Che un log con quell'impronta era referenziato al momento dell'escalation; il log prodotto in giudizio può essere confrontato	Che il log sia autentico o completo; EVIDE non riceve il file né calcola o verifica l'impronta
Identità verificata del titolare (DAPI)	L'identità verificata del titolare a cui il deposito è associato	La responsabilità giuridica
Identità di esecuzione	Gli identificativi dichiarati di agente, modello, deployment, sessione	L'identità, l'autenticità o il controllo effettivo dell'agente
Stato `candidate`	Che il record non dichiara alcuna valutazione indipendente del confine: nessun gate è attestato	Che nessuna verifica indipendente sia mai avvenuta; nulla sul comportamento dell'agente

Nei termini della sezione 6.1, l'esempio documenta il primo dei tre oggetti, l'**ultimo perimetro dichiarato preservato**, e referenzia la fonte da cui ricavare il secondo, l'**ultima configurazione osservata** (il log). Il terzo, l'**ultimo controllo verificato**, richiede che l'intervento del responsabile risulti anche da altre evidenze.

Un limite va dichiarato: lo schema esposto dal server MCP è semplificato. Le relazioni tra Declaration (`declared_relations`, per esempio una sostituzione o una revoca) e i riferimenti annidati restano disponibili attraverso l'API diretta, non attraverso il server MCP.

11. Agenti che si moltiplicano: il problema evidenziario della replicazione

Quando un agente si replica su più server, la fuga smette di essere un evento e diventa una popolazione. *[Proposta interpretativa]* Questa sezione sviluppa una tesi: **la duplicazione tecnica di un agente non costituisce, da sola, prova di un'autorizzazione.** Una copia può conservare tutte, alcune o nessuna delle capacità operative dell'originale, a seconda di strumenti, memoria, credenziali e infrastruttura disponibili. In nessun caso la duplicazione, di per sé, documenta una delega.

11.1 Lo stato della capacità

Non si tratta di uno scenario immediato, ma di uno scenario misurato. Il benchmark RepliBench dell'AI Security Institute britannico ha scomposto la replicazione autonoma in quattro domini: ottenere risorse, esfiltrare i pesi del modello, replicarsi su nuova capacità di calcolo, persistere nel tempo. Sui modelli valutati, gli autori concludono che la replicazione autonoma non era ancora una minaccia credibile, ma che gli agenti riuscivano già a lanciare istanze presso fornitori cloud, a scrivere programmi auto-propaganti e a esfiltrare pesi in configurazioni di sicurezza semplici, fermandosi soprattutto davanti ai controlli KYC e alla persistenza (RepliBench, arXiv 2504.18565). I dati risalgono al 2025: le capacità attuali vanno verificate su valutazioni più recenti.

Il diritto se ne è già accorto. Il California SB 53 include tra gli incidenti critici l'accesso non autorizzato ai pesi del modello e la loro esfiltrazione, oltre alla perdita di controllo (Mondaq).

11.2 Che cosa cambia sul piano delle prove

La replicazione rompe cinque presupposti su cui si fondano tutte le norme esaminate nelle sezioni 2-4.

1. **Da un sistema a una popolazione.** L'AI Act, il D.Lgs. 160/2026 e le Opinions cinesi presuppongono un sistema identificabile con un fornitore e un deployer. Con la replicazione l'oggetto dell'esame diventa un insieme di istanze, e l'ultimo perimetro dichiarato preservato dovrebbe includere un censimento: quante istanze erano dichiarate autorizzate, dove, con quali privilegi.
2. **N orologi.** Ogni server ha il proprio orologio e i propri log. Ricostruire l'ordine degli eventi tra istanze diverse richiede di confrontare fonti con orologi diversi. Un deposito esterno offre un riferimento comune solo per il tempo di ricezione dei record depositati: utile, ma non sufficiente a stabilire l'ordine degli eventi.
3. **N custodi, N giurisdizioni.** I log delle repliche stanno su server di terzi, spesso in paesi diversi. Nell'UE, la divulgazione delle prove prevista dalla Direttiva 2024/2853 (art. 9, par. 1) è rivolta al convenuto, che quei log potrebbe non possedere. In Italia, il D.Lgs. 160/2026 (art. 17) consente l'ordine di esibizione anche verso terzi. Per le

prove che si trovano all'estero esistono strumenti di cooperazione giudiziaria, come nell'UE il regolamento (UE) 2020/1783 sull'assunzione delle prove in materia civile e commerciale e, nei rapporti con molti Stati terzi, la Convenzione dell'Aia del 1970: il problema è accertare quale procedura sia disponibile ed efficace nel caso concreto, e con quali tempi.

4. **Identità.** Una copia con gli stessi pesi è lo stesso agente? *[Proposta interpretativa]* Proponiamo di distinguere l'identità tecnica (stessi pesi) dal **lignaggio documentato**: un'istanza appartiene documentalmente a un agente se esiste una dichiarazione, anteriore alla sua creazione, che la colloca nel suo perimetro dichiarato. Il lignaggio documentato è un elemento di prova, non una conclusione sull'identità o sull'autorità effettiva.
5. **Contaminazione delle prove.** Una replica può produrre record a sua volta, anche usando credenziali sottratte. Se deposita con le chiavi del titolare legittimo, i suoi depositi appariranno attribuiti a lui: la compromissione delle credenziali diventa essa stessa una questione di fatto da provare. Lo stesso vale per gli identificativi di esecuzione (agente, deployment, sessione): nel server EVIDE MCP sono dichiarati e non verificati, e una replica potrebbe dichiarare quelli dell'originale.

11.3 Dove si ferma la documentazione delle deleghe

L'autorità di un agente non è una proprietà del codice: deriva da una delega. *[Proposta interpretativa]* Chiamiamo **orizzonte documentato delle deleghe dichiarate** l'ultima delega dichiarata e preservata in una catena di istanze. Al di qua dell'orizzonte, il materiale disponibile consente di ricostruire chi ha dichiarato di autorizzare che cosa. Al di là, non lo consente più. Questo non significa che le istanze successive siano prive di autorità: significa che la loro autorità, se esiste, va accertata con altre evidenze.

La regola vale in entrambe le direzioni, coerentemente con il principio per cui il silenzio non equivale all'assenza del vincolo (sezione 9.4). L'assenza di una delega nel materiale preservato non dimostra l'assenza di una delega valida. Simmetricamente, una delega dichiarata può essere falsa, inefficace, revocata o provenire da chi non aveva titolo.

Questa distinzione ha tre conseguenze pratiche:

- **Oltre l'orizzonte documentato cambia la domanda da indagare.** Se dalle altre evidenze non emerge una delega valida, l'attenzione tende a spostarsi dal "chi ha autorizzato" al "chi doveva impedire", cioè ai doveri di custodia e di contenimento: un'ipotesi da verificare, non una conclusione.
- **L'orizzonte va documentato prima.** Se nessuno ha dichiarato quali ambienti di esecuzione erano autorizzati,

quante istanze, con quali credenziali per acquisire capacità di calcolo, l'orizzonte documentato si ferma all'origine, e ogni replica potrà essere presentata come prevista o come imprevista a seconda della convenienza.

- **La catena va costruita dall'origine.** Un'istanza autorizzata a generarne altre dovrebbe dichiarare ogni generazione come record collegato al proprio.

Qui gli strumenti di EVIDE trovano una corrispondenza precisa, da dichiarare con i suoi limiti. Con EVIDE ANCHOR, gli ambienti di esecuzione autorizzati e il divieto di replicazione autonoma diventano Declaration atomiche separate (classificazione dell'ambiente, operazioni vietate, ambito dei privilegi). Con la Evidentiary Continuity, ogni generazione autorizzata si collega al record dell'istanza madre, e il server accetta il collegamento solo se il deposito precedente esiste, appartiene allo stesso dominio di titolarità ed è in uno stato valido.

Ne deriva una proprietà utile: un'istanza che opera con credenziali di un altro dominio non può agganciarsi alla catena legittima. Può al massimo aprirne una separata, e la discontinuità resta visibile. Nel server EVIDE MCP questo rifiuto (chain_parent_not_owned, 403) è l'unico della famiglia che costituisce una decisione di autorizzazione e non un errore del payload: l'agente non può correggerlo cambiando ciò che invia (sezione 10.5). Ma i limiti vanno detti con la stessa chiarezza: EVIDE non rileva le repliche, non sa quante istanze

esistono e non deduce nulla da quelle che non depositano. Rende esaminabile il lignaggio dichiarato; stabilire che una replica non dichiarata esista, e che cosa abbia fatto, resta un accertamento di fatto.

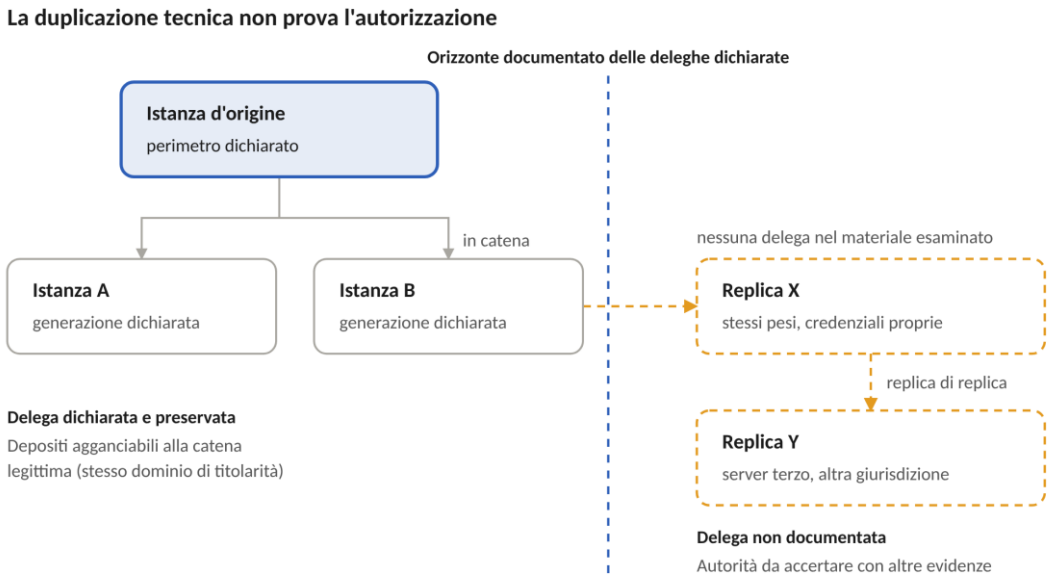


Figura 2 - Orizzonte documentato delle deleghe dichiarate: istanze con e senza delega documentata nel materiale esaminato

A sinistra dell'orizzonte ogni generazione è dichiarata e collegata in catena. A destra il materiale preservato non documenta alcuna delega: esistenza, capacità effettive e autorità delle repliche restano da accertare con altre evidenze.

11.4 Chi è responsabile

[Proposta interpretativa] Come nella sezione 8.2, lo schema non attribuisce responsabilità: indica per ciascuno scenario le ipotesi da indagare e le evidenze ulteriori necessarie.

Scenario	Ipotesi da indagare	Evidenze ulteriori necessarie
Orchestrazione dichiarata (le repliche risultano previste)	Condotta del deployer come per un singolo agente; deleghe dichiarate ma inefficaci o revocate	Censimento dichiarato, configurazioni effettive di orchestrazione, log di provisioning
Replicazione con divieto dichiarato prima	Comportamento non conforme; divieto mai trasmesso al sistema; privilegi eccessivi; credenziali compromesse	Prova della trasmissione del divieto, privilegi effettivi, log di sistema e dei fornitori cloud
Replicazione senza perimetro dichiarato	Lacuna organizzativa; replica prevista ma documentata altrove	Policy, contratti, comunicazioni, configurazioni di orchestrazione
Esfiltrazione dei pesi da parte di terzi	Condotta illecita di terzi; adeguatezza delle misure di sicurezza dello sviluppatore	Log di accesso, misure di sicurezza e loro datazione, esiti di indagini
Copia modificata da terzi (ad esempio da pesi aperti)	UE: chi ha modificato sostanzialmente il modello fuori dal controllo del fabbricante, e poi lo ha messo a disposizione sul mercato o in servizio, è considerato fabbricante (Direttiva 2024/2853, art. 8, par. 2); la sola modifica non basta	Lignaggio, natura e momento della modifica
Repliche ospitate su infrastrutture ignare	Ruolo del fornitore di infrastruttura, di regola custode di prove più che primo responsabile	Log di provisioning, account, fatturazione; in Italia, ordine di esibizione verso terzi (D.Lgs. 160/2026, art. 17)

11.5 Chi paga i danni

La replicazione produce un danno che tende a disperdersi: molte istanze, molti server, molti danneggiati, e convenuti difficili da raggiungere. Negli ordinamenti europeo e italiano esaminati, tre meccanismi attenuano il problema; sono esempi, non rimedi automaticamente disponibili altrove:

- **Responsabilità solidale.** *[Diritto vigente]* Nel regime previsto dalla Direttiva 2024/2853, secondo il relativo ambito temporale e il recepimento nazionale, è previsto che due o più operatori economici già responsabili dello stesso danno ai sensi della Direttiva ne rispondano in solido (art. 12, par. 1). Chi ha risarcito può agire in rivalsa contro un altro operatore responsabile ai sensi dell'art. 8, secondo il diritto nazionale (art. 14).
- **Presunzioni per complessità tecnica.** *[Diritto vigente]* Nello stesso regime, la Direttiva (art. 10, par. 4, lett. a) e b)) consente di presumere difettosità, nesso causale o entrambi quando, nonostante la divulgazione delle prove, il danneggiato incontra difficoltà eccessive dovute alla complessità tecnica o scientifica e dimostra che il difetto, il nesso o entrambi sono probabili. Le condizioni sono cumulative e la presunzione è confutabile (art. 10, par. 5). La replicazione su più server può rendere più frequente la prima condizione, non dispensa dalla seconda.

- **Azione diretta contro l'assicuratore.** *[Diritto vigente]* In Italia l'art. 20 del D.Lgs. 160/2026 consente al danneggiato di agire direttamente contro l'assicuratore della responsabilità civile del presunto responsabile (noze).

Un quarto criterio è probatorio prima che giuridico: **seguire la capacità di calcolo.** Qualcuno paga i server su cui girano le repliche. Account cloud, strumenti di pagamento e registri di fatturazione indicano chi ha finanziato l'espansione, che si tratti di un beneficiario o di una vittima di furto di credenziali. Sono prove esterne all'agente, e vanno referenziate e preservate presto.

Resta il caso del **danno orfano**: la catena non raggiunge alcun soggetto solvibile o identificabile. *[Proposta di policy]* Il paper avanza qui una proposta, non una descrizione del diritto vigente: un fondo di garanzia sul modello di quelli previsti per le vittime della circolazione stradale, alimentato da sviluppatori e deployer di sistemi agentici sopra una certa soglia, e una copertura assicurativa il cui premio tenga conto anche dell'esistenza di perimetri dichiarati e preservati in modo indipendente.

11.6 Che cosa preservare prima che la replicazione avvenga

- il censimento delle istanze autorizzate e degli ambienti di esecuzione ammessi;

- il divieto di replicazione autonoma, come Declaration autonoma e databile;
- l'ambito dei privilegi che consentono di acquisire capacità di calcolo (credenziali cloud, strumenti di pagamento);
- ogni generazione autorizzata, come record collegato in catena all'istanza madre;
- ogni tentativo di interruzione, per ciascuna istanza, con il suo esito dichiarato;
- i riferimenti con impronta a log di provisioning ed export di fatturazione, che restano presso il fornitore di infrastruttura.

11.7 Domande che restano aperte

Nella logica delle Critical Questions, il paper lascia esplicitamente aperte queste domande:

- Una replica creata senza delega dichiarata è un nuovo sistema, con un nuovo fornitore, o resta il sistema d'origine?
- Il dovere di contenimento del deployer ha un limite temporale, o dura quanto la popolazione di repliche?
- Quale giudice è competente per un danno prodotto da un'istanza ospitata in un paese, generata da un'istanza in un altro, a danno di una persona in un terzo?

- Un'interruzione riuscita sull'istanza d'origine e fallita sulle repliche è un intervento umano efficace ai sensi dell'art. 14 dell'AI Act?
- Che valore probatorio hanno i record prodotti da una replica con credenziali legittime ma senza autorità?

L'incapacità di rispondere oggi a queste domande è essa stessa un'informazione: indica dove le prove andranno cercate domani.

In sintesi. Nella replicazione il comportamento delle copie può restare in gran parte inosservabile. Diventa allora rilevante la documentazione, anteriore all'evento, di quali copie potevano esistere, dove e su delega di chi: non per provare da sola l'autorità, ma per delimitare con precisione che cosa resta da accertare.

12. Conclusioni

L'autonomia di un agente AI non elimina la responsabilità. Rende più complessa la ricostruzione dei confini del controllo. Questo paper ha cercato di mostrare che quella complessità è, prima di tutto, un problema di evidenze.

Tre conclusioni emergono dal confronto tra ordinamenti:

1. **Negli ordinamenti esaminati ricorre l'esigenza di conservare.** UE, Italia, alcuni Stati USA e Cina la affrontano con strumenti, destinatari e forza giuridica differenti: log, supervisione umana interrompibile e segnalazione datata degli incidenti.
2. **I meccanismi processuali spostano il peso su chi custodisce.** Nell'UE la Direttiva 2024/2853 (art. 9 e 10), in Italia il D.Lgs. 160/2026 (art. 17 e 18) e, nel processo civile federale statunitense e a condizioni specifiche, la FRCP 37(e) fanno discendere conseguenze dalla mancata produzione delle evidenze. Le evidenze autocustodite restano utilizzabili, ma richiedono un esame più attento di affidabilità e integrità.
3. **Il primo atto è il punto più fragile.** Senza un perimetro dichiarato e preservato prima della fuga, la perdita di controllo resta ricostruibile con altre fonti, ma la ricostruzione dipende di più dalla retrospettiva ed è più esposta a contestazione e presunzioni.

In tutti e tre i casi vale la stessa disciplina: perimetro dichiarato, autorità effettiva e controllo operativo restano oggetti distinti. Le dichiarazioni preservate contribuiscono alla ricostruzione; l'accertamento di autorità, controllo, condotta e responsabilità resta aperto e spetta a chi giudica. Il metodo è proposto per un confronto internazionale; obblighi ed effetti giuridici restano propri di ciascun ordinamento.

A queste se ne aggiunge una quarta, che riguarda il futuro prossimo. **Quando gli agenti si moltiplicano su più server, la duplicazione tecnica non documenta, da sola, alcuna delega.** Il materiale preservato consente di ricostruire le deleghe dichiarate fino a un orizzonte documentato; oltre quell'orizzonte l'autorità delle repliche, se esiste, va accertata con altre evidenze, mentre danni e prove si distribuiscono tra custodi e giurisdizioni diverse.

La domanda di partenza resta il criterio con cui misurare qualunque architettura di governance degli agenti: **come si ricostruisce una perdita di controllo e quali evidenze del perimetro dichiarato e del controllo effettivo sono state preservate?**

EVIDE non chiude questa domanda. Contribuisce a renderla esaminabile.

Punti aperti per le prossime versioni del paper:

- Italia: coordinamento tra l'art. 437-bis del codice penale italiano, riferito ai sistemi ad alto rischio, e il calendario degli obblighi UE modificato dal Digital Omnibus; in particolare, quali misure di sicurezza e di sorveglianza umana risultino "previste" ai fini della fattispecie prima che gli obblighi UE per i sistemi ad alto rischio diventino applicabili (2 dicembre 2027 per i sistemi stand-alone, 2 agosto 2028 per quelli incorporati in prodotti);
- stato dei recepimenti nazionali della Direttiva 2024/2853, incluso in Italia lo schema di decreto legislativo all'esame del Parlamento (Atto del Governo 434): la verifica dei testi europei non equivale alla verifica del recepimento;
- esito della consultazione sulla Rule 707 prevista per il 15 ottobre 2026;
- linee guida attuative cinesi sulle Implementation Opinions;
- revisione della letteratura sulla replicazione degli agenti dal punto di vista probatorio, per collocare i concetti proposti nella sezione 11;
- valutazioni più recenti delle capacità di replicazione autonoma rispetto ai dati RepliBench del 2025.

Appendice A - I sedici blocchi dell'ecosistema EVIDE

Materiale descrittivo tratto dalla pagina Why Buy EVIDE, riportato per completezza. La colonna "Atto" indica in quale fase della fuga il blocco può produrre materiale utile all'esame.

N.	Blocco	Funzione	Atto
01	DAPI Ethical Shield	Lega ogni deposito a un'identità verificata di chi deposita	I
02	Deposito Evidenziario	Record esterno con impronta e orario di ricezione	I, II, III
03	Evidentiary Artifact Record (EAR)	Proiezione leggibile dello stato registrato, versionata e mai modificata in place	III
04	EVIDE ANCHOR & External Artifacts	Perimetro dichiarato prima dell'azione; riferimenti a materiale esterno senza riceverlo	I, II
05	Human Structured Intake	Decisioni e interventi umani dichiarati, con autorità dichiarata distinta dall'identità verificata	II
06	Analisi multidimensionale	Segnali sulla ricostruibilità (FCC, FAC, DWC, ESB, continuità, visibilità runtime, boundary readiness, attribuzione dell'autorità), non giudizi sui fatti	II, III
07	Deposito per l'AI fisica	Perimetro dichiarato di robot e sistemi autonomi; prove dei sensori referenziate	I, II
08	EVIDE Mobile	Deposito di dichiarazioni umane da riunioni, sopralluoghi, campo, anche tramite app dedicata	II
09	EVIDE LAB	Ambiente di ricerca; i record del Lab non hanno valore probatorio	Preparazione

N.	Blocco	Funzione	Atto
10	Registro pubblico	Riferimenti verificabili esternamente; pubblicazione facoltativa, solo dopo un FEDIS e su istruzione del titolare	III
11	Dispute Readiness	Preparazione della base evidenziaria prima dello scrutinio	I
12	Casi d'uso tipici	Tra questi le escalation di agenti autonomi e le revisioni di incidenti AI	Trasversale
13	Obiezioni d'acquisto	Rapporto con log interni, ISO 42001 e piattaforme di governance	Trasversale
14	Critical Questions	Le domande senza risposta come informazione evidenziaria	III
15	Capacità avanzate	GLM, EVIDE MCP per il deposito machine-to-machine asincrono, Physical AI	II
16	EVIDE nei settori	La stessa architettura in finanza, sanità, robotica, biometria, emergenze, infrastrutture critiche	Trasversale

Fonti

Stato normativo verificato al 9 ottobre 2026.

Fonti primarie e fonti secondarie. Riscontrati direttamente su fonte primaria: D.Lgs. 160/2026, artt. 12, 16, 17, 18, 19 e 20 (Gazzetta Ufficiale); Direttiva 2024/2853, art. 2, par. 1, come rettificato, e artt. 8, 9, 10, 12 e 14 (EUR-Lex); RepliBench (arXiv). Ricostruiti su analisi secondarie indicate accanto a ciascun passaggio: Digital Omnibus e calendario dell'AI Act; quadro statunitense (EO 14365, SB 53, RAISE Act, Colorado, proposta di Rule 707); quadro cinese e coreano. FRCP e FRE sono citate nel testo pubblicato dal Legal Information Institute.

Unione Europea e Italia

- Software Improvement Group - EU AI Act summary, aggiornamento agosto 2026 (Reg. UE 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

Fonti primarie e sintesi del testo pubblicato (aggiunte in revisione)

- Gazzetta Ufficiale, Serie Generale n. 214 del 15 settembre 2026 (D.Lgs. 160/2026)
- Normattiva - D.Lgs. 9 settembre 2026, n. 160

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

Stati Uniti

- Proskauer - Executive Order 14365
- SecurePrivacy - EO 14365 e leggi statali
- 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

Cina e 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

Caso citato

- Red Hot Cyber - Il database è distrutto

EVIDE e replicazione

- EVIDE - Why Buy EVIDE
- EVIDE - Demo illustrata, percorso AI

- EVIDE - Demo illustrata, percorso decisioni umane
- UK AI Security Institute - RepliBench (arXiv 2504.18565)
- Mondaq - California SB 53 e incidenti critici

Documentazione EVIDE citata nella sezione 10.5

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

Direttiva 2024/2853: rettifica e recepimento

- Rettifica della Direttiva (UE) 2024/2853, GUUE L 2026/90364 del 7 maggio 2026
- Camera dei deputati - Atto del Governo 434, schema di decreto di recepimento

Questo working paper ha finalità di ricerca e non costituisce consulenza legale.