La catena della fiducia | The Chain of Trust

Identità, integrità e provenienza nell'era dell'AI governata | Identity, Integrity and Provenance in the Age of Governed AI

Autore / Author: Pasquale Papanice

Affiliazione / Affiliation: Uniark SRLU — SYNTO® (https://synto.it)

Ruolo / Role: CEO & Founder

Data / Date: Marzo / March 2026

Versione / Version: 1.0

Indice / Table of Contents

Bibliografia / References

Bateson, G. (1972). Steps to an Ecology of Mind. Ballantine Books.

EDPB — European Data Protection Board. (2025). Guidelines on the processing of personal data through blockchain technologies. https://www.activemind.legal/guides/edpb-blockchain/

European Commission. (2024). Regulation (EU) 2024/1183 — eIDAS 2.0, amending Regulation (EU) No 910/2014. https://eur-lex.europa.eu/eli/reg/2024/1183/oj

European Commission. (2025). European Blockchain Services Infrastructure (EBSI). https://digital-strategy.ec.europa.eu/en/policies/european-blockchain-services-infrastructure

European Parliament. (2024). EU AI Act — Article 10: Data and data governance. https://artificialintelligenceact.eu/article/10/

European Parliament. (2024). EU AI Act — Article 50: Transparency obligations. https://artificialintelligenceact.eu/article/50/

Gibson, J.J. (1979). The Ecological Approach to Visual Perception. Houghton Mifflin.

Kanizsa, G. (1955). Margini quasi-percettivi in campi con stimolazione omogenea. Rivista di Psicologia, 49(1), 7–30.

Lazzaroli, V. (2025). AI e Sanità: un patto che si costruisce sulla fiducia. Intervista. 01health. https://www.01health.it/tecnologie/intelligenza-artificiale/ai-e-sanita-un-patto-che-si-costruisce-sulla-fiducia-intervista-a-valeria-lazzaroli-presidente-di-enia/

McKinsey & Company. (2025). Why Digital Trust Truly Matters. https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-digital-trust-truly-matters

Microsoft News. (2025). Microsoft Italia ed ENIA: collaborazione per la trasformazione digitale delle PMI italiane. https://news.microsoft.com/source/emea/features/microsoft-italia-ed-enia-ente-nazionale-per-lintelligenza-artificiale-annunciano-una-collaborazione-per-la-trasformazione-digitale-delle-pmi-italiane-in-chiave-ai-present/?lang=it

NSA & CISA. (2025). Strengthening Multimedia Integrity in the Generative AI Era — Content Credentials. https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF

OWASP Foundation. (2025). Top 10 for Large Language Model Applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/

Oxford Academic. (2025). Reconciling Blockchain and Data Protection Laws. Journal of Cybersecurity, 11(1). https://academic.oup.com/cybersecurity/article/11/1/tyaf002/8024082

Papanice, P. (2026a). Il vincolo che dà forma all'AI — Dimensioni ortogonali come precondizione della riferibilità. Working paper. SYNTO / Uniark SRLU.

Papanice, P. (2026b). Contenuto come Con-tenuto — La forma che emerge dai vincoli. Working paper. SYNTO / Uniark SRLU.

Spencer-Brown, G. (1969). Laws of Form. Allen & Unwin.

SPLX.ai. (2025). RAG Poisoning in Enterprise Knowledge Sources. https://splx.ai/blog/rag-poisoning-in-enterprise-knowledge-sources

Zou, W., Geng, R., Wang, B., & Jia, J. (2025). PoisonedRAG: Knowledge Corruption Attacks on Retrieval-Augmented Generation of Large Language Models. Proceedings of the 34th USENIX Security Symposium. https://arxiv.org/abs/2407.12784

VERSIONE ITALIANA

Abstract

I due paper precedenti di questa trilogia hanno stabilito che il contenuto non pre-esiste ai vincoli ma ne emerge (Contenuto come Con-tenuto), e che le dimensioni ortogonali sono condizioni a priori della riferibilità informativa, non metadati accessori (Il vincolo che dà forma all'AI). Questo terzo paper affronta la domanda che quei due lasciano aperta: se il vincolo genera il contenuto, e se le dimensioni ortogonali rendono l'informazione navigabile — chi garantisce che il contenuto sia autentico, attribuibile e anteriore? La risposta è una catena di fiducia a sei anelli: identità verificata (KYC bancario), notarizzazione on-chain (dual hash), classificazione ortogonale (ACL), enforcement nel retrieval (RAG governato), provenance chain (audit documentale), e proof of prior art (certificazione IP). Nessun anello è sufficiente da solo. La catena chiusa è la condizione di possibilità della fiducia verificabile nell'era dell'AI.

Parole chiave: catena di fiducia, notarizzazione blockchain, ACL, RAG, provenance, KYC, proprietà intellettuale, EU AI Act, eIDAS 2.0, governance dell'informazione

1. Introduzione: dove si interrompe la trilogia

Nei paper precedenti (Papanice, 2026a; 2026b) abbiamo sostenuto due tesi complementari.

La prima: il contenuto è con-tenuto — non pre-esiste ai vincoli ma emerge da essi, come il triangolo di Kanizsa emerge dalla configurazione dei pacman (Kanizsa, 1955). L'allucinazione dell'AI generativa è l'assenza di vincoli adeguati, non un difetto del modello.

La seconda: le dimensioni classificatorie ortogonali — chi produce, chi accede, dove, cosa, a quale livello di riservatezza — non sono etichette applicate dopo il contenuto, ma condizioni a priori della riferibilità, nel senso kantiano. Senza coordinate, l'informazione non è navigabile. Senza navigabilità, il retrieval è aleatorio. Senza retrieval deterministico, il RAG allucinante.

Ma entrambe le tesi, per quanto solide, lasciano aperta una domanda cruciale: come si dimostra che il contenuto è autentico?

Un sistema con dimensioni ortogonali perfette e ACL rigoroso può ancora essere alimentato da documenti falsi, alterati o retrodatati. Un sistema RAG senza allucinazioni può restituire risposte impeccabilmente ancorate a fonti corrotte. La riferibilità non implica l'autenticità. La navigabilità non implica l'integrità.

Questo paper chiude la trilogia affrontando il piano che i precedenti presupponevano senza tematizzare: il piano probatorio. Non cosa rende l'informazione navigabile, ma cosa rende l'informazione dimostrabilmente vera, attribuibile e anteriore.

2. Il contesto: la crisi della verificabilità

2.1 Il dato avvelenato

Nel 2025, un gruppo di ricercatori ha presentato alla conferenza USENIX Security un paper destinato a cambiare la percezione della sicurezza nei sistemi RAG. PoisonedRAG (Zou et al., 2025) ha dimostrato che l'inserimento di soli cinque documenti malevoli in una knowledge base di milioni è sufficiente a produrre risposte false nel 90% dei casi su query mirate. L'attacco funziona su tutti i principali modelli e framework RAG testati. Tutte le difese valutate si sono rivelate inefficaci.

La portata è radicale. Non si tratta di un attacco alla superficie — ai prompt, alle API, all'interfaccia. Si tratta di un attacco al substrato: ai documenti stessi che l'AI considera come verità di riferimento. Se il ground truth è corrotto, nessuna sofisticazione del modello può compensare.

L'OWASP, il riferimento mondiale per la sicurezza applicativa, ha preso atto di questa nuova classe di vulnerabilità introducendo nel 2025 la voce LLM08 — Vector and Embedding Weaknesses nel suo Top 10 per i Large Language Model (OWASP, 2025). Una categoria che prima non esisteva.

Parallelamente, SPLX.ai (2025) ha documentato come le knowledge base enterprise siano vulnerabili a forme di poisoning ancora più sottili: documenti acquistati da database esterni già compromessi, API di terze parti che iniettano contenuti manipolati, fonti ereditate da acquisizioni aziendali mai verificate.

2.2 L'identità mancante

Ma il poisoning è solo metà del problema. L'altra metà è l'anonimato del contributore.

I sistemi documentali tradizionali — dai DMS enterprise ai repository cloud — registrano chi ha caricato un file. Ma registrare non è verificare. Un nome utente in un log non è un'identità verificata. Un account aziendale non prova che il soggetto sia chi dichiara di essere.

In un ecosistema multi-tenant, dove soggetti diversi contribuiscono contenuti a pool condivise, l'assenza di identità verificata è una falla strutturale. Chiunque abbia le credenziali può caricare qualsiasi cosa. E ciò che viene caricato entra nel substrato che l'AI tratta come verità.

2.3 Il tempo non certificato

La terza dimensione della crisi è temporale. Anche se il contenuto fosse autentico e il contributore verificato, manca la datazione certa. Un file system registra timestamp modificabili. Un database conserva date che l'amministratore può alterare. Un cloud storage certifica l'upload, non l'esistenza anteriore del contenuto.

Senza datazione immutabile, tre scenari critici restano irrisolvibili:

- Prior art: un soggetto crea un contenuto innovativo, ma non può dimostrare di averlo posseduto prima di un concorrente
- Compliance: un audit richiede di provare che un documento esisteva in una determinata versione a una determinata data, ma il sistema non offre garanzie di immutabilità
- Dispute contrattuali: una parte contesta i termini di un accordo, e nessuno può dimostrare quale versione fosse in vigore a quale momento

La crisi della verificabilità ha quindi tre facce: contenuto non garantito (poisoning), identità non verificata (anonimato), tempo non certificato (assenza di timestamp immutabili). Ciascuna, da sola, è sufficiente a invalidare la fiducia nel sistema. Insieme, rendono ogni output dell'AI strutturalmente inattendibile.

3. La risposta normativa: convergenza europea

Il legislatore europeo ha riconosciuto questa crisi con una convergenza normativa senza precedenti.

3.1 EU AI Act: la provenienza come obbligo

Il Regolamento Europeo sull'Intelligenza Artificiale (EU AI Act) stabilisce la tracciabilità dei dati come requisito fondante. L'Articolo 10 impone ai fornitori di sistemi AI ad alto rischio di mantenere audit trail che le autorità possano esaminare: record temporizzati delle modifiche ai dati, delle decisioni di filtraggio, delle valutazioni di qualità (European Parliament, 2024a). L'Articolo 50 estende gli obblighi di trasparenza a categorie più ampie di sistemi AI (European Parliament, 2024b).

L'implicazione per i sistemi RAG è diretta: ogni risposta deve essere riconducibile alla fonte, e ogni fonte deve avere una storia verificabile. Non è più sufficiente trovare il documento giusto — bisogna dimostrare che era quel documento, in quella versione, in quel momento.

3.2 eIDAS 2.0: il registro distribuito diventa servizio fiduciario

Il Regolamento (UE) 2024/1183 — noto come eIDAS 2.0 — ha introdotto una novità di portata storica: gli electronic ledger sono riconosciuti come categoria di servizio fiduciario (European Commission, 2024). Per la prima volta, la normativa europea conferisce valore legale ai registri distribuiti come strumenti di certificazione.

Questo significa che un hash scritto su una blockchain conforme può avere, nell'ordinamento europeo, la stessa forza probatoria di una marca temporale qualificata. Non è più necessario affidarsi a interpretazioni giurisprudenziali: la base normativa è esplicita.

3.3 EDPB: il pattern di conformità

L'European Data Protection Board, nell'aprile 2025, ha pubblicato le prime linee guida sulla conformità GDPR delle tecnologie blockchain, identificando 16 fattori specifici di valutazione (EDPB, 2025). Il pattern raccomandato è esplicito: dati personali off-chain (cancellabili), hash crittografici on-chain (immutabili). Questo pattern risolve la tensione strutturale tra l'immutabilità della blockchain e il diritto alla cancellazione dell'Articolo 17 GDPR.

3.4 EBSI: l'infrastruttura già esiste

L'European Blockchain Services Infrastructure, sviluppata in oltre cinque anni dalla Commissione Europea con tutti gli Stati Membri, è oggi production-ready (European Commission, 2025). Connette circa 40 enti pubblici in una rete blockchain unificata, con casi d'uso specifici che includono la notarizzazione documentale, la verifica delle credenziali e la condivisione fidata di dati.

La convergenza è completa: l'obbligo normativo (AI Act), il framework giuridico (eIDAS 2.0), le linee guida operative (EDPB), e l'infrastruttura tecnica (EBSI) sono tutti in posizione. Manca l'architettura applicativa che li faccia convergere in un sistema operativo.

4. La catena della fiducia: sei anelli

La tesi di questo paper è che la fiducia verificabile in un sistema di conoscenza AI-driven richiede una catena chiusa a sei anelli, dove ciascun anello dipende dagli altri e nessuno è sufficiente da solo.

4.1 Anello 1 — Identità verificata (KYC bancario)

Il primo anello risponde alla domanda: chi ha contribuito il contenuto?

L'approccio tradizionale — username e password, account aziendale, login federato — verifica le credenziali, non l'identità. Un account può essere condiviso, rubato, creato con dati falsi.

L'approccio qui proposto è strutturalmente diverso: l'identità del soggetto è legata al proprio IBAN attraverso il superamento del KYC bancario. Il Know Your Customer è un processo imposto per legge (D.Lgs. 231/2007 in Italia, Direttive AML/CFT europee) che obbliga gli istituti di credito a verificare l'identità reale del titolare del conto. Il soggetto non dichiara chi è — lo dimostra attraverso un terzo di fiducia regolamentato che ha già eseguito la verifica.

Questo meccanismo implementa il principio del trust delegation: invece di costruire un'infrastruttura KYC proprietaria — costosa, complessa, soggetta a errori — si delega la verifica a chi è già obbligato per legge a farla. La banca diventa, di fatto, l'oracolo di identità.

4.2 Anello 2 — Notarizzazione at upsert (dual hash)

Il secondo anello risponde alla domanda: cosa esisteva, e quando?

Nel momento in cui un documento viene indicizzato nel sistema RAG, il pipeline genera automaticamente un dual hash:

- Omega-1 (Ω_1): SHA-256 del contenuto grezzo del file, scritto su blockchain. Questo hash ha valore universale: chiunque possieda il file originale può verificare indipendentemente che il contenuto corrisponda a quello notarizzato. Il timestamp blockchain certifica il quando.
- Omega-2 (Ω_2): SHA-256 del contenuto concatenato al filename, conservato in database locale. Il filename porta con sé la classificazione ACL (le dimensioni ortogonali discusse in Papanice, 2026a), rendendo Ω_2 una prova non solo del contenuto ma del contesto di governance in cui quel contenuto esisteva al momento dell'indicizzazione.

Due hash, un'unica operazione. Ω_1 prova l'esistenza e l'integrità. Ω_2 prova la governance. Il costo è trascurabile: una singola transazione blockchain per documento.

La notarizzazione è non-bloccante: se la blockchain è temporaneamente irraggiungibile, il documento viene comunque indicizzato. La notarizzazione verrà completata al ripristino della connettività. L'indisponibilità della blockchain non interrompe il flusso operativo.

4.3 Anello 3 — Classificazione ortogonale (ACL)

Il terzo anello risponde alla domanda: chi può vedere cosa?

Come argomentato estesamente in Papanice (2026a), le dimensioni classificatorie ortogonali — chi produce, chi accede, livello di riservatezza, dominio semantico, collocazione geografica — non sono etichette opzionali ma condizioni a priori della riferibilità. Un documento senza classificazione è un'intuizione senza concetto, nel senso kantiano: esiste, ma non può essere conosciuto.

In termini operativi, l'ACL (Access Control List) traduce le dimensioni ortogonali in regole di visibilità computabili. Ogni documento porta con sé le proprie coordinate; ogni utente ha un profilo che definisce quali coordinate sono accessibili. L'intersezione genera lo spazio informativo pertinente — non una restrizione del totale, ma un contenuto emergente, come argomentato in Papanice (2026b).

4.4 Anello 4 — Enforcement nel retrieval (RAG governato)

Il quarto anello risponde alla domanda: l'AI consulta solo ciò che è autorizzato?

L'enforcement ACL nel RAG non è un filtro applicato ai risultati — è un vincolo sullo spazio di ricerca. Prima che il retrieval inizi, il sistema circoscrive il perimetro documentale accessibile al profilo dell'utente. L'AI non vede documenti che l'utente non è autorizzato a consultare — non li esclude dopo averli trovati, non li trova affatto.

Questa distinzione è architetturalmente critica. Un filtro post-retrieval lascia tracce: il sistema ha acceduto al documento, anche se non lo mostra. Un vincolo pre-retrieval elimina il rischio alla radice: il documento non entra mai nel contesto di generazione.

4.5 Anello 5 — Provenance chain (audit documentale)

Il quinto anello risponde alla domanda: quale fonte ha generato quale risposta?

Ogni risposta del sistema RAG è riconducibile ai chunk documentali che l'hanno generata. Ogni chunk è riconducibile a un documento. Ogni documento ha un hash notarizzato su blockchain. La catena è completa: domanda → risposta → chunk → documento → hash → blockchain → timestamp.

Questo è ciò che la letteratura chiama knowledge provenance: la provenienza verificabile della conoscenza. Non solo cosa l'AI ha risposto, ma da dove ha attinto, quale versione del documento ha consultato, e quando quel documento è stato certificato.

L'EU AI Act, all'Articolo 10, richiede esattamente questo: record temporizzati delle fonti, tracciabilità delle trasformazioni, documentazione della catena di custodia.

4.6 Anello 6 — Proof of prior art (certificazione IP)

Il sesto anello chiude la catena rispondendo alla domanda più delicata: chi possedeva cosa, e da quando?

La combinazione dei cinque anelli precedenti produce un artefatto giuridicamente potente: identità verificata (Anello 1) + hash del contenuto (Anello 2) + timestamp blockchain (Anello 2) = prova che il soggetto X, la cui identità è certificata via KYC bancario, possedeva il contenuto Y alla data Z.

Questo è il fondamento della proof of prior art — la prova di anteriorità che tutela:

- Diritto d'autore: il contenuto esisteva in quella forma, attribuito a quel soggetto, a quella data
- Segreto industriale: il know-how era documentato e classificato, con accesso ristretto, prima di qualsiasi divulgazione
- Know-how aziendale: la conoscenza tacita, una volta esplicitata e caricata nel sistema, acquisisce datazione certa e attribuzione verificabile
- Brevetti: la documentazione di prior art è immutabile e opponibile

Senza il primo anello (identità), il proof of prior art è una traccia anonima. Senza il secondo (notarizzazione), è una dichiarazione non verificabile. Senza il terzo e quarto (ACL + RAG), non c'è prova che il contenuto fosse governato. Senza il quinto (provenance), non c'è legame tra la conoscenza e il suo uso. Ogni anello è debole senza gli altri.

5. Proprietà della catena

5.1 Chiusura

La catena è chiusa nel senso forte del termine. L'ultimo anello (proof of prior art) dipende dal primo (identità), che a sua volta acquista significato solo in presenza degli anelli intermedi. Non c'è un punto di ingresso privilegiato: la catena funziona solo quando è completa.

Questa chiusura richiama il concetto autopoietico di Maturana e Varela discusso in Papanice (2026b): il sistema genera i propri confini, e i confini generano il sistema. La catena della fiducia non è una sequenza lineare di passaggi — è un anello ricorsivo dove ogni componente legittima le altre.

5.2 Non-bloccabilità

La catena è progettata per essere resiliente. La notarizzazione blockchain è non-bloccante: un'indisponibilità temporanea della blockchain non interrompe l'indicizzazione. L'identità KYC è verificata una volta all'adesione, non a ogni upload. L'ACL è computato in tempo reale, non pre-calcolato. La provenance è generata automaticamente, non richiesta manualmente.

Nessun componente dipende dalla disponibilità sincrona di tutti gli altri. La catena è eventually consistent: converge alla completezza anche se alcuni anelli si completano con ritardo.

5.3 Costo marginale trascurabile

Il costo della notarizzazione on-chain è dell'ordine di 0,005 MATIC per transazione — meno di 0,003 euro per documento ai tassi correnti. Per un'organizzazione che indicizza 100 documenti al mese, il costo annuale della

notarizzazione blockchain è inferiore a 4 euro. La certificazione dell'intera knowledge base è economicamente irrilevante.

Questo è fondamentale perché rende possibile la copertura totale: non è necessario selezionare quali documenti certificare. La catena si applica a tutto ciò che entra nel sistema. Il documento che non è stato certificato non esiste.

5.4 Idempotenza

Se un documento viene ri-caricato senza modifiche, il dual hash è identico. Il sistema rileva il duplicato e non genera una nuova transazione blockchain. Se il contenuto cambia, viene generato un nuovo hash e una nuova transazione — ma l'hash precedente è preservato. La catena mantiene la storia completa delle versioni, senza sovrascritture.

6. Il paradigma: fiducia come proprietà architetturale

Valeria Lazzaroli, Presidente di ENIA — Ente Nazionale per l'Intelligenza Artificiale, ha sintetizzato con precisione il principio che anima questa architettura:

> *"Serve un ecosistema di fiducia che parta dalla chiarezza normativa." (Lazzaroli, 2025)*

La fiducia, in questa visione, non è un attributo soggettivo — mi fido di te perché ti conosco — ma una proprietà dell'infrastruttura. Non dipende dalla buona volontà dei partecipanti ma dall'architettura del sistema. Come la ENIA ha sottolineato nella sua collaborazione con Microsoft Italia per la prima sandbox normativa tecnico-giuridica dedicata all'AI Act (Microsoft News, 2025), la fiducia sistemica richiede "un ambiente strutturato, rigoroso e indipendente."

La catena della fiducia qui proposta è esattamente questo: un ambiente in cui la fiducia non è dichiarata ma computata. Ogni anello produce una prova verificabile. Ogni prova è indipendente dal buon volere del soggetto. La catena nel suo complesso genera ciò che McKinsey (2025) chiama digital trust — una fiducia che non richiede di conoscere la controparte, perché è garantita dall'infrastruttura.

> *"Prevedere il rischio prima di subirlo è già una forma alta di etica." (Lazzaroli, 2025)*

Questo paper concorda. E aggiunge: prevedere il rischio non è sufficiente se non si dispone dell'infrastruttura per dimostrare le precauzioni adottate. L'etica predittiva diventa operativa solo quando è verificabile.

7. Il vuoto di mercato

La ricerca condotta per questo paper conferma che nessuna soluzione attualmente disponibile sul mercato combina i sei anelli della catena in un'unica architettura nativa.

I principali vector database — Pinecone, Weaviate, Qdrant — offrono isolamento a livello di namespace o tenant, ma nessuno implementa ACL nativo a livello di singolo documento. L'integrazione con framework di autorizzazione esterni (SpiceDB, Cerbos, OPA) è possibile ma richiede sviluppo custom e introduce punti di giuntura dove la sicurezza si degrada.

I servizi di notarizzazione blockchain — da Notarize a CERTO.legal — operano come moduli standalone, scollegati dal flusso documentale e dal retrieval AI. La certificazione avviene manualmente, selettivamente, e non produce una provenance chain.

I sistemi RAG enterprise — da Azure AI Search a Amazon Kendra — implementano filtri di sicurezza ma non classificazione ortogonale nativa, non notarizzazione, non identità verificata.

Il risultato è che un'organizzazione che desideri una catena di fiducia completa deve assemblare 3-5 strumenti diversi, con integrazione custom, senza garanzie di coerenza architetturale e con gap di sicurezza a ogni giuntura.

8. Conclusione: il terzo taglio

In Il vincolo che dà forma all'AI (Papanice, 2026a), il primo taglio era cognitivo: le dimensioni ortogonali come condizione a priori della riferibilità. In Contenuto come Con-tenuto (Papanice, 2026b), il secondo taglio era epistemologico: il vincolo come generatore di contenuto, non come limitazione.

Questo terzo paper compie il taglio probatorio: la catena della fiducia come condizione di possibilità della verificabilità nell'era dell'AI governata.

I tre tagli sono complementari e cumulativi:

- Primo taglio — Riferibilità: senza dimensioni ortogonali, l'informazione non è navigabile. Il retrieval è aleatorio. L'AI allucinante.
- Secondo taglio — Generatività: il vincolo non limita ma genera. L'ACL non censura ma produce contenuto pertinente. Il confine è costitutivo.
- Terzo taglio — Verificabilità: la catena chiusa — identità, integrità, governance, retrieval, provenance, prior art — trasforma la fiducia da attributo soggettivo a proprietà dell'infrastruttura.

Spencer-Brown (1969) scrisse: "Draw a distinction." Il primo paper ha mostrato che la distinzione genera lo spazio navigabile. Il secondo che lo spazio navigabile genera il contenuto. Questo terzo che il contenuto, per essere reale, deve essere dimostrabile.

In termini batesoniani (Bateson, 1972): una differenza che fa una differenza. Ma la differenza, nell'era dell'AI, deve essere anche una differenza che si può dimostrare.

Il vincolo dà forma. La forma dà contenuto. Il contenuto, per essere conoscenza, richiede prova.

La catena della fiducia è il terzo vincolo. Quello che rende la conoscenza — reale.

ENGLISH VERSION

Abstract

The two preceding papers in this trilogy established that content does not pre-exist constraints but emerges from them (Content as Con-tained), and that orthogonal dimensions are a priori conditions of informational referentiality, not accessory metadata (The Constraint That Shapes AI). This third paper addresses the question those two leave open: if constraint generates content, and if orthogonal dimensions make information navigable — who guarantees that content is authentic, attributable, and prior? The answer is a six-link chain of trust: verified identity (bank KYC), on-chain notarization (dual hash), orthogonal classification (ACL), enforcement in retrieval (governed RAG), provenance chain (documentary audit), and proof of prior art (IP certification). No link is sufficient alone. The closed chain is the condition of possibility for verifiable trust in the age of AI.

Keywords: chain of trust, blockchain notarization, ACL, RAG, provenance, KYC, intellectual property, EU AI Act, eIDAS 2.0, information governance

1. Introduction: Where the Trilogy Left Off

In the preceding papers (Papanice, 2026a; 2026b) we argued two complementary theses.

The first: content is con-tained — it does not pre-exist constraints but emerges from them, just as the Kanizsa triangle emerges from the configuration of pacmen (Kanizsa, 1955). Generative AI hallucination is the absence of adequate constraints, not a model defect.

The second: orthogonal classificatory dimensions — who produces, who accesses, where, what, at which confidentiality level — are not labels applied after content, but a priori conditions of referentiality, in the Kantian sense. Without coordinates, information is not navigable. Without navigability, retrieval is random. Without deterministic retrieval, RAG hallucinates.

But both theses, however solid, leave open a crucial question: how do you prove that content is authentic?

A system with perfect orthogonal dimensions and rigorous ACL can still be fed with false, altered, or backdated documents. A hallucination-free RAG system can return responses impeccably anchored to corrupted sources. Referentiality does not imply authenticity. Navigability does not imply integrity.

This paper closes the trilogy by addressing the plane that the preceding ones presupposed without thematizing: the evidentiary plane. Not what makes information navigable, but what makes information demonstrably true, attributable, and prior.

2. The Context: The Verifiability Crisis

2.1 Poisoned Data

In 2025, a group of researchers presented at USENIX Security a paper destined to change the perception of security in RAG systems. PoisonedRAG (Zou et al., 2025) demonstrated that inserting just five malicious documents into a knowledge base of millions is sufficient to produce false responses in 90% of cases on targeted queries. The attack works across all major models and RAG frameworks tested. All evaluated defenses proved ineffective.

The scope is radical. This is not a surface attack — on prompts, APIs, or interfaces. It is an attack on the substrate: the documents themselves that AI treats as ground truth. If the ground truth is corrupted, no model sophistication can compensate.

OWASP, the global reference for application security, acknowledged this new vulnerability class by introducing LLM08 — Vector and Embedding Weaknesses in its 2025 Top 10 for Large Language Models (OWASP, 2025). A category that previously did not exist.

In parallel, SPLX.ai (2025) documented how enterprise knowledge bases are vulnerable to even more subtle poisoning: documents purchased from already-compromised external databases, third-party APIs injecting manipulated content, sources inherited from unverified corporate acquisitions.

2.2 The Missing Identity

But poisoning is only half the problem. The other half is contributor anonymity.

Traditional document management systems — from enterprise DMS to cloud repositories — record who uploaded a file. But recording is not verifying. A username in a log is not a verified identity. A corporate account does not prove that the subject is who they claim to be.

In a multi-tenant ecosystem where different subjects contribute content to shared pools, the absence of verified identity is a structural flaw. Anyone with credentials can upload anything. And what is uploaded enters the substrate that AI treats as truth.

2.3 Uncertified Time

The third dimension of the crisis is temporal. Even if content were authentic and the contributor verified, certified dating is missing. A file system records modifiable timestamps. A database preserves dates that administrators can alter. Cloud storage certifies upload, not prior existence of the content.

Without immutable dating, three critical scenarios remain unresolvable:

- Prior art: a subject creates innovative content but cannot prove they possessed it before a competitor
- Compliance: an audit requires proving that a document existed in a specific version at a specific date, but the system offers no immutability guarantees
- Contractual disputes: a party contests agreement terms, and no one can demonstrate which version was in effect at which moment

The verifiability crisis thus has three faces: unguaranteed content (poisoning), unverified identity (anonymity), uncertified time (absence of immutable timestamps). Each alone is sufficient to invalidate trust in the system. Together, they render every AI output structurally unreliable.

3. The Regulatory Response: European Convergence

The European legislator has recognized this crisis with unprecedented regulatory convergence.

3.1 EU AI Act: Provenance as Obligation

The European AI Regulation establishes data traceability as a foundational requirement. Article 10 mandates that providers of high-risk AI systems maintain audit trails that authorities can examine: timestamped records of data modifications, filtering decisions, and quality assessments (European Parliament, 2024a). Article 50 extends transparency obligations to broader categories of AI systems (European Parliament, 2024b).

The implication for RAG systems is direct: every response must be traceable to its source, and every source must have a verifiable history. It is no longer sufficient to find the right document — one must prove it was that document, in that version, at that moment.

3.2 eIDAS 2.0: Distributed Ledgers Become Trust Services

Regulation (EU) 2024/1183 — known as eIDAS 2.0 — introduced a historic innovation: electronic ledgers are recognized as a trust service category (European Commission, 2024). For the first time, European law confers legal value on distributed ledgers as certification instruments.

This means that a hash written to a compliant blockchain can have, in European law, the same evidentiary force as a qualified timestamp. Jurisprudential interpretation is no longer required: the legal basis is explicit.

3.3 EDPB: The Compliance Pattern

The European Data Protection Board, in April 2025, published the first GDPR compliance guidelines for blockchain technologies, identifying 16 specific assessment factors (EDPB, 2025). The recommended pattern is explicit: personal data off-chain (erasable), cryptographic hashes on-chain (immutable). This pattern resolves the structural tension between blockchain immutability and Article 17 GDPR's right to erasure.

3.4 EBSI: The Infrastructure Already Exists

The European Blockchain Services Infrastructure, developed over five years by the European Commission with all Member States, is now production-ready (European Commission, 2025). It connects approximately 40 public bodies in a unified blockchain network, with specific use cases including document notarization, credential verification, and trusted data sharing.

The convergence is complete: the legal mandate (AI Act), the juridical framework (eIDAS 2.0), the operational guidelines (EDPB), and the technical infrastructure (EBSI) are all in position. What is missing is the application architecture that makes them converge into an operational system.

4. The Chain of Trust: Six Links

The thesis of this paper is that verifiable trust in an AI-driven knowledge system requires a closed chain of six links, where each link depends on the others and none is sufficient alone.

4.1 Link 1 — Verified Identity (Bank KYC)

The first link answers the question: who contributed the content?

The traditional approach — username and password, corporate account, federated login — verifies credentials, not identity. An account can be shared, stolen, or created with false data.

The approach proposed here is structurally different: the subject's identity is linked to their IBAN through bank KYC clearance. Know Your Customer is a legally mandated process (EU AML/CFT Directives) requiring credit institutions to verify the real identity of account holders. The subject does not declare who they are — they prove it through a regulated trusted third party that has already performed the verification.

This mechanism implements the trust delegation principle: instead of building proprietary KYC infrastructure — costly, complex, error-prone — verification is delegated to those already legally obligated to perform it. The bank becomes, in effect, the identity oracle.

4.2 Link 2 — Notarization at Upsert (Dual Hash)

The second link answers the question: what existed, and when?

When a document is indexed in the RAG system, the pipeline automatically generates a dual hash:

- Omega-1 (Ω_1): SHA-256 of the raw file content, written to blockchain. This hash has universal value: anyone possessing the original file can independently verify that the content matches the notarized record. The blockchain timestamp certifies the when.
- Omega-2 (Ω_2): SHA-256 of the content concatenated with the filename, stored in a local database. The filename carries the ACL classification (the orthogonal dimensions discussed in Papanice, 2026a), making Ω_2 a proof not only of content but of the governance context in which that content existed at the time of indexing.

Two hashes, one operation. Ω_1 proves existence and integrity. Ω_2 proves governance. Cost is negligible: a single blockchain transaction per document.

Notarization is non-blocking: if the blockchain is temporarily unreachable, the document is indexed regardless. Notarization completes upon connectivity restoration. Blockchain unavailability does not interrupt the operational flow.

4.3 Link 3 — Orthogonal Classification (ACL)

The third link answers the question: who can see what?

As argued extensively in Papanice (2026a), orthogonal classificatory dimensions — who produces, who accesses, confidentiality level, semantic domain, geographic location — are not optional labels but a priori conditions of referentiality. A document without classification is an intuition without a concept, in the Kantian sense: it exists, but cannot be known.

In operational terms, the ACL (Access Control List) translates orthogonal dimensions into computable visibility rules. Each document carries its own coordinates; each user has a profile defining which coordinates are accessible. The intersection generates the pertinent informational space — not a restriction of the total, but emergent content, as argued in Papanice (2026b).

4.4 Link 4 — Enforcement in Retrieval (Governed RAG)

The fourth link answers the question: does the AI consult only what is authorized?

ACL enforcement in RAG is not a filter applied to results — it is a constraint on the search space. Before retrieval begins, the system circumscribes the document perimeter accessible to the user's profile. The AI does not see documents the user is not authorized to consult — it does not exclude them after finding them; it does not find them at all.

This distinction is architecturally critical. A post-retrieval filter leaves traces: the system accessed the document, even if it does not display it. A pre-retrieval constraint eliminates the risk at its root: the document never enters the generation context.

4.5 Link 5 — Provenance Chain (Documentary Audit)

The fifth link answers the question: which source generated which response?

Every RAG system response is traceable to the document chunks that generated it. Every chunk is traceable to a document. Every document has a hash notarized on blockchain. The chain is complete: query → response → chunks → document → hash → blockchain → timestamp.

This is what the literature calls knowledge provenance: the verifiable origin of knowledge. Not only what the AI answered, but from where it drew, which version of the document it consulted, and when that document was certified.

The EU AI Act, in Article 10, requires exactly this: timestamped records of sources, traceability of transformations, documentation of the chain of custody.

4.6 Link 6 — Proof of Prior Art (IP Certification)

The sixth link closes the chain by answering the most delicate question: who possessed what, and since when?

The combination of the five preceding links produces a legally powerful artifact: verified identity (Link 1) + content hash (Link 2) + blockchain timestamp (Link 2) = proof that subject X, whose identity is certified via bank KYC, possessed content Y at date Z.

This is the foundation of proof of prior art — the priority proof that protects:

- Copyright: the content existed in that form, attributed to that subject, at that date
- Trade secrets: the know-how was documented and classified, with restricted access, before any disclosure
- Corporate knowledge: tacit knowledge, once made explicit and uploaded to the system, acquires certified dating and verifiable attribution
- Patents: the prior art documentation is immutable and opposable

Without the first link (identity), the proof of prior art is an anonymous trace. Without the second (notarization), it is an unverifiable declaration. Without the third and fourth (ACL + RAG), there is no proof that content was governed. Without the fifth (provenance), there is no link between knowledge and its use. Every link is weak without the others.

5. Properties of the Chain

5.1 Closure

The chain is closed in the strong sense. The last link (proof of prior art) depends on the first (identity), which in turn acquires meaning only in the presence of intermediate links. There is no privileged entry point: the chain functions only when complete.

This closure echoes the autopoietic concept of Maturana and Varela discussed in Papanice (2026b): the system generates its own boundaries, and the boundaries generate the system. The chain of trust is not a linear sequence of steps — it is a recursive loop where each component legitimates the others.

5.2 Non-Blockability

The chain is designed for resilience. Blockchain notarization is non-blocking: temporary blockchain unavailability does not interrupt indexing. KYC identity is verified once at onboarding, not at every upload. ACL is computed in real time, not pre-calculated. Provenance is generated automatically, not requested manually.

No component depends on the synchronous availability of all others. The chain is eventually consistent: it converges to completeness even if some links complete with delay.

5.3 Negligible Marginal Cost

The cost of on-chain notarization is on the order of 0.005 MATIC per transaction — less than 0.003 euros per document at current rates. For an organization indexing 100 documents per month, the annual blockchain notarization cost is under 4 euros. Certifying the entire knowledge base is economically irrelevant.

This is fundamental because it enables total coverage: there is no need to select which documents to certify. The chain applies to everything that enters the system. The document that was not certified does not exist.

5.4 Idempotency

If a document is re-uploaded without modifications, the dual hash is identical. The system detects the duplicate and does not generate a new blockchain transaction. If content changes, a new hash and new transaction are generated — but the previous hash is preserved. The chain maintains the complete history of versions, without overwrites.

6. The Paradigm: Trust as Architectural Property

Valeria Lazzaroli, President of ENIA — the Italian National Entity for Artificial Intelligence, precisely synthesized the principle animating this architecture:

> *"What is needed is a trust ecosystem that starts from regulatory clarity." (Lazzaroli, 2025)*

Trust, in this vision, is not a subjective attribute — I trust you because I know you — but a property of infrastructure. It does not depend on participants' good will but on system architecture. As ENIA emphasized in its collaboration with Microsoft Italia for the first technical-legal regulatory sandbox dedicated to the AI Act (Microsoft News, 2025), systemic trust requires "a structured, rigorous, and independent environment."

The chain of trust proposed here is exactly that: an environment where trust is not declared but computed. Each link produces a verifiable proof. Each proof is independent of the subject's good will. The chain as a whole generates what McKinsey (2025) calls digital trust — trust that does not require knowing the counterpart, because it is guaranteed by infrastructure.

> *"To anticipate risk before suffering it is already a high form of ethics." (Lazzaroli, 2025)*

This paper agrees. And adds: anticipating risk is insufficient if one lacks the infrastructure to demonstrate the precautions taken. Predictive ethics becomes operational only when it is verifiable.

7. The Market Gap

Research conducted for this paper confirms that no currently available solution combines the six links of the chain in a single native architecture.

The leading vector databases — Pinecone, Weaviate, Qdrant — offer namespace- or tenant-level isolation, but none implements native ACL at the individual document level. Integration with external authorization frameworks (SpiceDB, Cerbos, OPA) is possible but requires custom development and introduces juncture points where security degrades.

Blockchain notarization services — from Notarize to CERTO.legal — operate as standalone modules, disconnected from the document flow and AI retrieval. Certification occurs manually, selectively, and does not produce a provenance chain.

Enterprise RAG systems — from Azure AI Search to Amazon Kendra — implement security filters but not native orthogonal classification, not notarization, not verified identity.

The result is that an organization seeking a complete chain of trust must assemble 3-5 different tools, with custom integration, without guarantees of architectural coherence and with security gaps at every juncture.

8. Conclusion: The Third Cut

In The Constraint That Shapes AI (Papanice, 2026a), the first cut was cognitive: orthogonal dimensions as a priori conditions of referentiality. In Content as Con-tained (Papanice, 2026b), the second cut was epistemological: constraint as content generator, not limitation.

This third paper makes the evidentiary cut: the chain of trust as the condition of possibility for verifiability in the age of governed AI.

The three cuts are complementary and cumulative:

- First cut — Referentiality: without orthogonal dimensions, information is not navigable. Retrieval is random. AI hallucinates.
- Second cut — Generativity: constraint does not limit but generates. ACL does not censor but produces pertinent content. The boundary is constitutive.
- Third cut — Verifiability: the closed chain — identity, integrity, governance, retrieval, provenance, prior art — transforms trust from a subjective attribute to an infrastructure property.

Spencer-Brown (1969) wrote: "Draw a distinction." The first paper showed that distinction generates navigable space. The second that navigable space generates content. This third that content, to be real, must be demonstrable.

In Batesonian terms (Bateson, 1972): a difference that makes a difference. But in the age of AI, the difference must also be a difference that can be proven.

Constraint gives form. Form gives content. Content, to be knowledge, requires proof.

The chain of trust is the third constraint. The one that makes knowledge — real.

Paper prepared for publication on Academia.edu

March 2026