sanctifai | TRUST

A SANCTIFAI TRUST WHITEPAPER · JUNE 2026

Trust but Verify

Why responsible AI requires
proof of human oversight.

On the role of human-in-the-loop in agentic AI — and why immutable auditability is the load-bearing element of every responsible AI framework, inside the enterprise and out.

SCAN TO VERIFY ON-CHAIN

EXECUTIVE SUMMARY

Enterprises are deploying agentic AI faster than they can govern it. Every major responsible AI framework — the EU AI Act, the NIST AI Risk Management Framework, ISO/IEC 42001 — answers this shift with the same two requirements: keep a qualified human in the loop at consequential decision points, and maintain a defensible record proving that oversight actually occurred.

Most organizations have invested in the first requirement and improvised the second. They route high-risk decisions to human reviewers, then document that review with a screenshot, a ticket comment, or a database row their own administrators can edit. When a regulator, customer, or opposing counsel asks for proof that a human reviewed a specific decision at a specific moment, the evidence turns out to be exactly as trustworthy as the systems that produced it — which is to say, not independently trustworthy at all. Underlying all of it is the problem agentic systems newly introduce — the **ambiguity of agency**: when an autonomous agent can act under a human's name, a record must prove not merely that someone was logged in, but that genuine human judgment was exercised at the decisive moment.

This paper argues that the audit trail is not a clerical afterthought of responsible AI. It is the load-bearing element. We describe a category of infrastructure — **verifiable human attestation** — that cryptographically binds a verified human's review to a specific task, anchors it where neither vendor nor customer can alter it, and makes the proof portable. SanctifAI Trust is our implementation. The argument stands on its own: as agents take on more of the work of the enterprise, *proof of human* becomes as fundamental to digital trust as proof of identity was to the last era of the internet.

KEY TERM

Ambiguity of agency

In an AI-mediated workflow, the uncertainty over both *who* acted and *whether a real human exercised judgment at all* — or whether an autonomous agent simply acted under that person's name.

01 The accountability gap in agentic AI

The first wave of enterprise AI governance was built for models that *predicted*. A credit-scoring model produced a number; a human took the action. Governance meant validating the model and documenting its lineage. The human action was the natural checkpoint, and conventional records — who clicked approve, when, in what system — were adequate because the human was unambiguously the *agent* of the decision.

Agentic AI dissolves that boundary. Agents now *act*: they execute multi-step workflows, call tools, move money, and chain decisions together at machine speed. The question governance must now answer is no longer only "was the model sound?" but "at the moments that mattered, was a human actually there — and can you prove it?" This is harder than it sounds, for three reasons.

Volume and velocity

A single agentic workflow may cross dozens of control points per hour. Human oversight at these checkpoints is increasingly mediated by the same software stack the agents run on — which means the *record* of oversight is generated by systems that are themselves part of what is being audited.

Ambiguity of agency

This is the deepest of the three, and the phrase is deliberate: *agency* names both halves of the problem — *who* actually acted, and *whether genuine human agency was exercised at all*, or an autonomous agent acted under a person's name. "Approved by user jsmith" answers neither; it cannot tell deliberate judgment from a rubber-stamp, a hijacked session, or an automation running under jsmith's login. As agents grow fluent at driving browsers and APIs, the line between "a human did this" and "software did this as a human" becomes genuinely contestable — in an audit, in procurement, in court. Resolving the ambiguity of agency — proving a real human, not the agent, exercised judgment when it mattered — is the problem this paper is about.

Adversarial scrutiny

The audiences for oversight evidence are no longer just internal. Regulators enforcing the EU AI Act, customers performing vendor due diligence, insurers pricing AI liability, and litigators in discovery will all, sooner or later, ask the same question: *prove the human review happened.* Evidence that can be edited by the party producing it does not survive adversarial scrutiny. It is testimony, not proof.

> *Enterprises are accumulating obligations to demonstrate human oversight at exactly the moment when demonstrating it credibly has become technically nontrivial.*

02 What the frameworks actually require

It has become fashionable to treat "responsible AI" as a values statement. The operative frameworks are more concrete than that. The **EU AI Act** classifies systems by risk and imposes binding obligations on high-risk applications: effective human oversight, technical documentation, and record-keeping sufficient to enable traceability. The **NIST AI RMF** organizes practice around Govern, Map, Measure, and Manage, threading human oversight and documentation through all four. **ISO/IEC 42001**, the first certifiable AI management standard, requires documented processes for human oversight and retained evidence that those processes operate as designed.

Notice what these three instruments — one binding law, one national framework, one international standard — have in common. Each pairs a *substantive* requirement (a human must be meaningfully in the loop) with an *evidentiary* one (you must retain records sufficient to demonstrate it). And each implicitly assumes the records are reliable.

FRAMEWORK	HUMAN-OVERSIGHT REQUIREMENT	EVIDENTIARY REQUIREMENT
EU AI Act	Effective human oversight of high-risk systems before consequential action.	Technical documentation and record-keeping enabling traceability; conformity assessment.
NIST AI RMF	Accountability structures and documented roles for intervention (Govern, Manage).	Mechanisms to record and respond to interventions and incidents.
ISO/IEC 42001	Documented processes for human oversight of AI systems.	Retained documented information sampled as audit evidence.

Shared blind spot — each requires that records exist and assumes they are reliable. None asks the next question: ***reliable according to whom?***

FIGURE 1. THE THREE INSTRUMENTS CONVERGE ON THE SAME PAIRING — AND THE SAME UNPROTECTED ASSUMPTION.

That assumption is the soft spot. A record-keeping system controlled entirely by the audited party satisfies the letter of the requirement while leaving its spirit unprotected. The history of financial controls suggests where this goes: every mature compliance regime eventually migrates from self-attested records to independently verifiable ones. Responsible AI will be no different — and organizations that build for verifiability now will find audits, procurement reviews, and disputes dramatically cheaper later.

03 Human-in-the-loop is the cornerstone — and the weakest link

Human-in-the-loop (HITL) is the consensus answer to AI risk, and rightly so. For decisions involving judgment, ethics, ambiguity, or irreversible consequences, a qualified human reviewer remains the most adaptable control available. Frameworks converge on HITL because nothing else absorbs the long tail of situations no policy anticipated.

But HITL as practiced has a structural weakness the frameworks gloss over: **the loop is only as trustworthy as the evidence that the human was in it.** Consider how human review is evidenced today — a status field in a case tool, a Slack thread, an EHR audit log maintained by the same party whose conduct may later be in question. The record exists, but it has four failure modes that matter under scrutiny.

Mutable

Database rows, logs, and ticket histories can be edited, backdated, or purged by the administrators of the systems that hold them.

Unattributed

The ambiguity of agency in miniature: a username proves a credential was used — not that a specific human was present, attentive, and deliberately exercising judgment. Sessions are shared, hijacked, automated.

Unbound

Most records say *that* something was approved, not precisely *what*. If the underlying artifact changes later, nothing detects the substitution.

Not portable

Evidence lives inside the producing organization's systems. A customer or auditor cannot verify it without taking the organization's word.

FIGURE 2. THE FOUR FAILURE MODES OF CONVENTIONAL OVERSIGHT EVIDENCE.

> *We harden the models, encrypt the data, monitor the drift — and then document the single most important control with the digital equivalent of a sticky note.*

The result is a paradox at the center of responsible AI: we have made human oversight the cornerstone of AI accountability while leaving the *proof* of that oversight on the weakest evidentiary foundation in the stack. Closing this gap is not a matter of more diligent record-keeping. It requires a different kind of record.

04 The properties of a defensible oversight record

If the purpose of an oversight record is to convince a skeptical third party, then the record must hold up *without* requiring trust in the party that produced it. Working backward from that standard, a defensible record of human participation needs five properties.

01

Verified presence

Resolves the ambiguity of agency: a real, present human bound to the act of review by hardware-backed biometric or liveness proof — not an inference from a login.

02

Cryptographic binding

The attestation commits to hashes of the exact task and result. Proof for Task A is invalid for Task B; later substitution is detectable.

03

Immutability

Once sealed, unalterable by vendor, customer, and any administrator of either. Anchored where no single party is in control.

04

Privacy by construction

Only salted commitments and an opaque human fingerprint persist. No names, emails, content, or biometrics on the record.

05

Portability

A sealed certificate the relying party verifies themselves — no screenshot to trust — that travels with the work product.

FIGURE 3. FIVE PROPERTIES THAT TURN A CLAIM OF HUMAN REVIEW INTO A CHECKABLE FACT.

Together these define a category of infrastructure — **verifiable human attestation** — that sits between agentic work and accountable approval. It is to human oversight what the certificate authority was to web identity: the mechanism that converts a claim into a checkable fact.

05 SanctifAI Trust: proof of human, engineered

SanctifAI Trust is our implementation of this category, built on three pillars that map one-to-one to the properties above. Trust sits between agentic work and accountable approval, producing a durable record that follows the work into audits, customer portals, and third-party workflows.

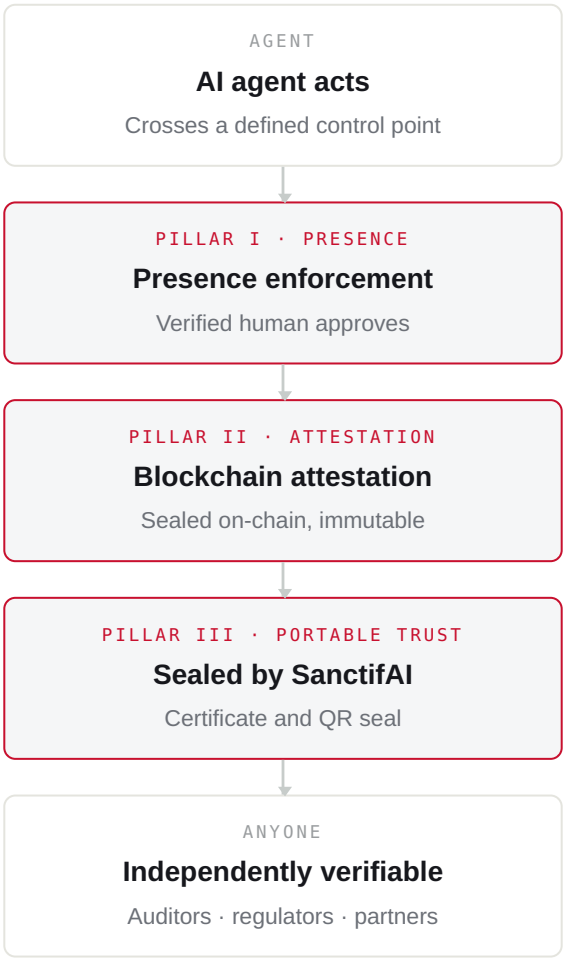


FIGURE 4. THE ATTESTATION FLOW, FROM CONTROL POINT TO PORTABLE PROOF.

Presence enforcement

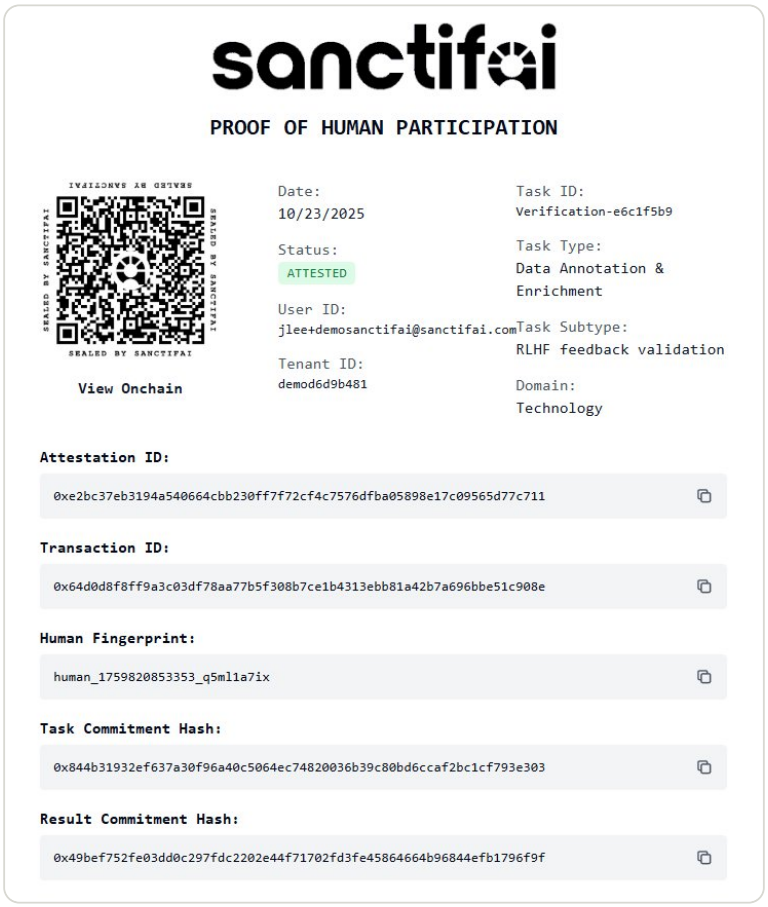
When an agentic workflow crosses a policy threshold, Trust issues a cryptographic challenge containing the task identifier, input/output hashes, and the control requiring review. A qualified reviewer responds through a Human Presence Interface — preferred embodiment WebAuthn biometric assertion, with support for hardware keys, decentralized identity, and liveness challenges. The resulting presence token is short-lived (sixty seconds), single-use, and cryptographically bound to that task's commitments, so verification cannot be replayed, transferred, or rubber-stamped in bulk. Review happens where the work happens — embedded in your application, or via a browser extension with zero-redirect cross-origin verification.

Blockchain-backed attestation

Each completed review becomes an attestation anchored via the Ethereum Attestation Service on Base. The on-chain record contains only commitments — task, tenant, and result hashes plus an opaque human fingerprint and timestamp — under a zero-knowledge architecture in which sensitive identifiers never appear raw. The sealed attestation cannot be edited by SanctifAI. It cannot be edited by you. That symmetry is the point: when you hand the record to a regulator, its credibility does not depend on your infrastructure, your administrators, or your incentives.

Sealed by SanctifAI — portable trust

Every attestation resolves to a public, human-readable certificate and an embeddable QR seal that travels with the work product into PDFs, dashboards, and audit submissions. A verification API lets third parties and downstream agents check proofs programmatically — which matters in a world where the consumer of your work product is increasingly another AI system that needs a machine-checkable basis for trust.



What's not in here: the on-chain anchor stores only salted commitments and an opaque human fingerprint — no names, emails, document content, or biometrics. Identity shown above is held off-chain and dereferenced only under legal hold.

FIGURE 5. ANATOMY OF A PROOF-OF-HUMAN CERTIFICATE. EVERY FIELD EXISTS BECAUSE AN AUDITOR, A REGULATOR, OR OPPOSING COUNSEL WILL EVENTUALLY ASK FOR IT.

06 Immutable auditability, inside and out

It is worth being concrete about who consumes this evidence, because the internal and external cases are different and both are underserved today.

Internally — from reconstruction to retrieval

Responsible AI programs call for recurring audits of high-risk systems, KPIs on oversight coverage, and quarterly executive reporting. Today, assembling that evidence means reconciling logs across ticketing systems, chat tools, and application databases — a quarterly archaeology project whose findings are only as strong as the weakest log. With attestation at the control points, oversight coverage becomes a queryable fact: which controls fired, who reviewed, when, with what disposition — every record pre-verified and tamper-evident. Just as importantly, immutability disciplines the present: reviewers behave differently when their approval is a sealed, permanent attestation rather than a checkbox, and "approval inflation" — the quiet decay of HITL into click-through — becomes visible in the record itself.

Externally — compliance as a commercial asset

Enterprise procurement increasingly asks vendors not whether they use AI responsibly but whether they can *prove* it; a sealed, independently verifiable certificate of human review is a categorically better answer than a SOC 2 appendix and a promise. Regulated entities can hand conformity assessors records whose integrity does not depend on the entity's own systems. Insurers can price against verifiable oversight rather than attested policy. And in disputes — where evidence matters most — an attestation anchored beyond either party's control is the difference between proof and testimony.

A THIRD AUDIENCE: OTHER AGENTS

As inter-organizational workflows become agent-to-agent, the question "was a human in the loop on your side?" will be asked by software, at machine speed, as a precondition of transaction. Machine-verifiable proof of human participation is the only answer that scales.

07 What verifiable attestation is — and is not

Thought leadership earns its name by being honest about scope, so we will be.

It is one control, not the whole program

A complete governance framework spans AI inventories, risk classification, model lineage, bias auditing, drift monitoring, explainability, access control, and governance culture. Trust does not do those things, and the platforms that do are complements, not competitors. Trust slots in at a precise point: wherever the program says "a human must review this, and we must be able to prove it," Trust is the proof.

Proof of human is not proof of good judgment

An attestation establishes that a verified, qualified human deliberately approved a specific output at a specific moment. It does not establish that the approval was correct, unbiased, or wise — those assurances come from reviewer qualification, independent validation, and bias auditing. What attestation contributes is *accountability with integrity*: when a decision is later questioned, there is an unimpeachable record of who stood behind it and exactly what they stood behind. Accountability does not guarantee quality, but quality programs are unenforceable without it.

Immutability binds everyone — that is the feature

Some organizations hesitate at records they themselves cannot amend. That instinct should be inverted. Evidence you can edit is evidence a skeptic can discount; the inability to alter your own attestations is what makes them worth presenting. We anchor to a public ledger not out of ideology but because the property at stake — a record no single party, including us, can alter, verifiable by anyone — is what public attestation infrastructure uniquely provides. A private append-only log restores the original problem: trust in the operator.

08 From oversight to provenance

The near-term driver of verifiable attestation is compliance: the EU AI Act's enforcement timeline, ISO 42001 certification, NIST-aligned audit expectations. But the longer arc is bigger than compliance. As generative and agentic AI saturate knowledge work, **human provenance** — the verifiable fact that a human made, reviewed, or authorized a particular digital artifact — becomes a scarce and valuable property of work itself. Customers will pay premiums for it. Regulators will mandate it. Markets will form around it.

In that world, the organizations that thrive will be those that treated proof of human not as paperwork but as product: embedded at every consequential control point, sealed beyond dispute, and presented to the world as a mark of how they work. Responsible AI frameworks tell us *what* to do — keep humans accountable for the decisions that matter. Verifiable attestation is *how* we make that accountability real, durable, and worth something to the people who must rely on it. In the end, every responsible-AI obligation reduces to one question — was human agency real and present when it mattered — and verifiable attestation is what makes the answer checkable.

Trust, but verify. We built the verify.

SANCTIFAI TRUST

sanctifai | TRUST

SanctifAI Trust is the audit trail for human oversight of agentic AI: every review, approval, and escalation captured as verifiable, tamper-evident Proof of Human. Trust is part of the SanctifAI family — the Human Layer of the AI Economy — connecting AI agents to accountable human judgment for verification, escalation, consultation, and simulation.

To see how Trust maps to your EU AI Act, NIST AI RMF, and ISO/IEC 42001 obligations, or to book a demo, visit trust.sanctifai.com.