Human Seal

WHITE PAPER · VERSION 1.1

The Human Seal: verifiable human authority over AI

A signed, offline-verifiable proof that a named, authorized human approved a specific AI action, grounded in the organization's own modeled roles and data. The mathematics behind the claim, the honest limits, and a working reference implementation anyone can check.

> **The thesis.** The Human Seal is not a rubber stamp. It is a composite, independently verifiable assurance. Its authority comes not from anyone asserting it, but from every one of its properties being checkable by anyone, offline, against a published key. We do not ask to be trusted. We make trust checkable.

Published by HumanGate International (HGI), a neutral steward.
Reference implementation: FLOCORE.
Trust root: did:web:humanseal.world · Standard: the Human Gate · Mark: the Human Seal
2026-09-19 · Not legal advice. Legal statements are for qualified review.

ABSTRACT

Abstract

Artificial intelligence has begun to act: to send money, change records, and answer for organizations as if it were them. The question every board is now asking is simple. Who said yes?

This paper defines the **Human Gate** (an open standard) and the **Human Seal** (its certification mark): a control point at which a real, authorized human must approve any world-affecting AI action, and a signed receipt proving that the approval happened. We show that the Seal's authority is not asserted, it is **constructed from six independently verifiable properties**: cryptographic unforgeability, action binding, modeled authority, a non-bypass invariant, grounding faithfulness, and their composition. We give the mathematics behind each, tie each to a mechanism in a live reference implementation, and state honestly what the Seal does and does not prove.

The paper also answers a sharper question: what does the Human Seal do that a plain digital signature does not? The answer is the **per-tenant model**. Using rgentic and FLOCORE, an organization's roles, authority and data ontology are modeled first. A Seal then certifies not merely that "some human clicked," but that **this person, in this authorized role, approved this exact action, grounded in this organization's own data and compliance rules, with no autonomous path around the gate.** That is the extra, and it is what makes the Seal evidence rather than decoration.

We situate the work against a decade of scholarship on meaningful human control, against the cryptographic and standards prior art (Ed25519, SHA-256, W3C Verifiable Credentials, did:web, C2PA), and against a global regulatory wave in which roughly two dozen jurisdictions and six cross-border instruments now converge on a single requirement: a human must be able to oversee and intervene, and there must be a record. The Human Seal produces that record as a portable, verifiable object.

CONTENTS

Contents

1 The accountability gap

When a person makes a consequential decision, we know who to ask. When an adaptive machine makes it, we often do not. In 2004, Andreas Matthias named this the **responsibility gap**: for learning systems whose behavior the maker can no longer fully predict, the traditional basis for holding the maker or operator responsible erodes, and a gap opens where no one is properly answerable SOURCED. Eight years earlier, Helen Nissenbaum had catalogued why computerization strains accountability at all: the problem of "many hands," bugs treated as no one's fault, the computer as scapegoat, and ownership without liability VERIFIED.

These are not abstractions. A Canadian tribunal held an airline liable after its customer-service chatbot invented a refund policy that did not exist, rejecting the argument that the bot was a separate entity the company was not answerable for SOURCED. A major employer scrapped a recruiting model that downgraded women's applications SOURCED. In each case the machine acted, a person was harmed, and the line back to a responsible human was contested.

The pressure is now financial and personal. Boards face fiduciary exposure for AI failures: reporting suggests fewer than one in four companies have a board-approved AI governance policy while a majority already use AI operationally SOURCED. AI-related securities class actions doubled from 2023 to 2024; average director-and-officer settlements rose about 27 percent to roughly 56 million dollars, even as D&O policies add broad AI exclusions SOURCED. The market is pricing the gap. What it lacks is a way to close it.

> **THE GAP, IN ONE LINE**
>
> An AI acted. Someone was affected. No one can prove which human stood behind it. The Human Seal is the artifact that makes that human, and that moment, provable.

2 Meaningful human control: the idea already has a name

The requirement that a human remain meaningfully in charge of a consequential automated action is not new, and not ours. The term **meaningful human control** was introduced by the NGO Article 36 in a 2013 policy paper, and quickly became central to international debate SOURCED. Article 36's account names three elements that map almost exactly onto our mechanism: **information** (the human has adequate context), **action** (a positive, deliberate human act initiates the outcome), and **accountability** (those who assess and act are answerable).

The most cited philosophical formalization, Santoni de Sio and van den Hoven (2018), grounds control in two necessary conditions VERIFIED:

- **Tracking:** the system responds to the relevant moral reasons of the relevant humans and the relevant facts in the environment. It does what those humans, on reflection, have reason to want. **This is the Gate.**
- **Tracing:** the system's behavior and outcomes trace back to at least one human who could appreciate the stakes and was positioned to be a locus of responsibility. **This is the Seal.**

Crucially, the authors frame control as a **design property for engineers**, not an instruction to operators. That is exactly the posture of this paper: control is built into the rail, not left to human diligence. Siebert and colleagues (2022) later operationalized the two conditions into four engineering properties, including one we return to repeatedly, that **responsibility recorded must be commensurate with the human's actual authority and ability to control** VERIFIED.

The governance literature also distinguishes three levels of oversight. **Human-in-the-loop** means a human must approve each consequential action before it takes effect. **Human-on-the-loop** means the system acts and a human monitors with a live veto. **Human-in-command** means a human decides whether the system is deployed at all SOURCED. For world-affecting, often irreversible actions, human-in-the-loop is the defensible standard, because on-the-loop and in-command alone let an action take effect before any human has endorsed it.

3 Why a human in the loop is not automatically enough

The strongest objection to any human-approval scheme is that it becomes theater, and the objection is evidence-based. The canonical finding is **automation bias and complacency**: people systematically over-trust automation that is usually right. Parasuraman and Manzey (2010) show the effect appears in experts as well as novices, is not reliably removed by training, and worsens under workload SOURCED. Ben Green (2022), surveying dozens of human-oversight policies, warns that oversight requirements can **legitimize** flawed systems and let vendors shirk accountability, giving "a false sense of security" SOURCED. Madeleine Clare Elish (2019) names the **moral crumple zone**: blame gets absorbed by the nearest human, who may have lacked real control SOURCED.

We take this seriously, and it shapes the design. A credible gate must therefore:

- present the human with genuine, decision-relevant context, not a bare "approve?" prompt;
- give real power to refuse, and adequate time to exercise it;
- record **what the human was shown, who they were, and what they decided**, so a rubber stamp is visible rather than hidden; and
- ensure the person who signs is the person with genuine authority, so blame cannot be dumped on a crumple-zone scapegoat.

The Seal is therefore designed as an **audit instrument, not a diligence guarantee**. It does not claim the human judged well. It makes the conditions of the decision inspectable, so that rubber-stamping can be detected and accountability lands where control actually sat. Notably, the law itself now names this concern: EU AI Act Article 14(4)(b) requires oversight designed so the human can "remain aware of the possible tendency of automatically relying or over-relying on the output" of the system, that is, automation bias, by name VERIFIED.

PART II · THE MECHANISM

4 The Gate, the Assent, the Seal

The mechanism has three named parts, one for the checkpoint, one for the act, and one for the proof.

Term	Definition
Human Gate	The control point at which a real human, distinct from the AI, must approve a world-affecting action before it runs. A world-affecting action is anything that changes the outside world: a payment, a write to a system of record, a message sent on the organization's behalf, a deletion. No autonomous path may bypass the gate.
Assent	The act itself: a specific, authorized person deciding, at the gate, to approve or to refuse this exact action, having seen its true context.
Human Seal	The signed, tamper-evident receipt that the assent happened. It records who approved (a stable identity reference, never personal data), the exact action they saw (bound by a content hash), the decision, and the time. It is signed with a cryptographic key and verifiable by anyone, offline, with no call back to the issuer.

The Human Gate is published as an **open, royalty-free standard**, so anyone can build to it. The Human Seal is a **controlled mark**, so that where it appears it means something. A system earns the Seal at one of three assurance levels: **Level 1 Gated** (every world-affecting action is held for a human and logged), **Level 2 Signed** (each approval produces a verifiable receipt bound to the approver and the exact action), and **Level 3 Assured** (a present-human, risk-scaled step-up, with a second approver for the highest-risk actions).

5 The per-tenant model: rgentic and FLOCORE

Here is the answer to "why does a Seal carry more weight than a lone signature?" A signature proves a key was used. It says nothing about whether the signer had the authority to approve, whether the action was consistent with the organization's rules, or whether the approval was grounded in real data. The Human Seal adds all three, because the organization is **modeled first**.

Each tenant becomes its own modeled system

rgentic ingests an organization's organogram, job descriptions and operating instructions and builds a per-tenant model: its **roles**, the **authority** attached to each role (role-based access control), the **data ontology** it works over, and the **compliance pack** that applies to its industry. Each tenant is, in effect, its own small FLOCORE: a faithful map of who may do what, over what data, under which rules. The Seal then certifies an approval **against that modeled reality**.

> THE EXTRA, STATED PLAINLY
>
> Because the roles and authority are modeled, an approval is not "some human clicked." It is: **this person, in this role, which is authorized for this action type, approved this exact action, grounded in this tenant's real data and its compliance pack, with no autonomous path around the gate.** The Seal records all of it. That is the difference between a decoration and a piece of evidence.

5.1 From organogram to dashboard to Seal

The per-tenant model is not an abstraction the customer never sees. It is built from what an organization already has, and it surfaces as a dashboard the organization uses every day.

1. **Inputs.** The organization's organogram, its job descriptions, and the KPIs each role owns.

↓

2. **rgentic builds the per-tenant model.** The roles, the authority attached to each one (RBAC), the data ontology, and the applicable compliance pack.

↓

3. **The model drives a visual dashboard.** Who is authorized for what, and each role's KPIs, rendered for the people who actually do the work.

↓

4. **The Human Seal validates against the model.** An approval at the gate is checked against exactly this modeled authority, and the Seal records that it was consistent with it.

The dashboard is the human-visible face of the model; the Seal is the machine-verifiable proof that an approval was consistent with it. The same organogram, job descriptions and KPIs that tell a person what they are responsible for are what tell the gate who is allowed to approve, and what the approval was measured against. That is why the Seal is evidence and not decoration: it is anchored to a model the organization can see, run, and correct.

Grounding status

The per-tenant ontology, the role registry and the RBAC that the model targets are **live** in the reference implementation; compliance packs ground on an organization's own policies (the first pack is seeded from a real cultivation tenant's six regulatory policies). The rgentic intake flow that ingests the organogram is scoped and in build. Throughout this paper we mark what is verified in code, what is design intent, and what is roadmap, so the reader is never asked to take a claim on trust.

PART III · THE MATHEMATICS

6 Six verifiable properties

The Seal's authority is a conjunction of six properties, each of which anyone can check. This section states each one precisely, gives the plain-language meaning, and ties it to a mechanism in the reference implementation.

Notation. Let A be an AI action, described canonically (the tool, parameters, target, actor and time). Let W be the set of **world-affecting** actions. A Human Seal is a receipt $R(A)$ carrying the fields defined in Appendix A. We write $H(x)$ for the SHA-256 hash of x, and $\mathrm{Sig}(sk, m)$ for an Ed25519 signature over message m with secret key sk.

6.1 Unforgeability and action binding

P1 Integrity and non-repudiation

A Seal is signed with Ed25519 (RFC 8032) over the canonical receipt. Ed25519 is proven **existentially unforgeable under chosen-message attack** (EUF-CMA), and in fact strongly unforgeable (SUF-CMA), in the random-oracle model under the discrete-logarithm assumption on Curve25519 (Brendel, Cremers, Jackson and Zhao, 2021) SOURCED.

```
# given the published verification key pk, for any adversary without sk:
Pr[ Verify(pk, m*, σ*) = 1  ∧  (m*, σ*) not previously issued ]  ≤  ε
# where ε is negligible under the stated hardness assumption
```

In plain terms. No one can produce a receipt that verifies against our public key without holding the private key, even after seeing any number of prior receipts. Equally, a party who did sign cannot later credibly deny it. We do not say "unbreakable"; we say **computationally infeasible under a standard, stated assumption**.

BASIS: VERIFIED in code. `receipt_signing.py` (Ed25519 sign / public_key), `human_seal.py issue()` signs the canonical receipt. Confirmed passing this session.

P2 Action binding

The receipt does not say "a human approved something." It carries `content_hash = H(A)`, the SHA-256 (FIPS 180-4) hash of the exact canonical action. SHA-256 offers about 128 bits of collision resistance VERIFIED. The signature therefore binds the approval to that one action.

```
# after a Seal R(A) is issued, substituting a different action A' is detectable:
A' ≠ A  ⇒  H(A') ≠ H(A) = R.content_hash   (except with prob. ~2^−128)
Verify(R, A') = false
```

In plain terms. Change one byte of the approved action and the hash no longer matches, so no one can swap in a different action under the same approval. **We tested this directly:** verification of a Seal against a swapped action returns false, and any edit to a signed field (for example the approver) fails the signature check.

BASIS: VERIFIED in code, tested this session. `human_seal.content_hash()` and `verify()` reject a mismatched action and any field tamper.

6.2 Authority and non-bypass

P3 Modeled authority

An approval is valid only if the approver holds a role authorized for the action's type. Let `roles(u)` be the approver's roles in the tenant model and `perm(type(A))` the permission the action requires.

```
# a Seal is only issued when authority holds in the tenant's RBAC model:
authorize( tenant, roles(u), perm(type(A)) ) = true
# so R(A) asserts: approver u ∈ Authorized(A)
```

In plain terms. The Seal is not "a human clicked," it is "a human who was allowed to approve this kind of action, in this organization's own model, approved it." This is what the per-tenant model (Section 5) adds over a lone signature.

BASIS: reference implementation. The runtime RBAC authorize check over the per-tenant role registry and resolved permissions (live), enforced at the gate before a Seal is minted.

P4 Non-bypass invariant

The gate is a total function over world-affecting actions: no such action executes without a valid Seal. This is the property that makes the whole scheme meaningful, because a gate with a side door proves nothing.

```
# let executed(W) be the world-affecting actions that ran, approved(W) those a human
sealed:
∀ A ∈ W :   execute(A) ⇒ ∃ R(A) with Verify(R(A)) = true
# therefore   executed(W)  ⊆  human-approved(W)
# and the count of AUTONOMOUS world-affecting actions is 0
```

In plain terms. Every world-affecting action either carries a human's Seal or does not run. There is no "auto-approve" rung. The reference implementation exposes this as a per-tenant **action map** that classifies every action as human-gated, agent-proposed (recommend only), or autonomous, and the autonomous count for world-affecting actions is zero by construction.

BASIS: reference implementation. The approval gate plus the per-tenant action map (verified this session: world-affecting actions classified human-gated, zero autonomous).

6.3 Faithfulness and composition

P5 Grounding faithfulness (bounded hallucination)

Where the action follows from an AI-generated answer, that answer is validated against the tenant's own records before it reaches the gate. Define a faithfulness score `f(answer, data) ∈ [0,1]` and a calibrated threshold `τ`.

```
f(answer, data) < τ  ⇒  refuse (the answer is not shipped, no action is proposed)
# the system says "I do not have that" rather than inventing it
```

In plain terms. The human is not asked to approve an action built on a fabrication. Below the threshold the system refuses and records the gap instead of guessing. This keeps the human's decision anchored to the organization's real data.

BASIS: reference implementation. Grounded-answer validation against tenant rows with an honest refusal path and a recorded knowledge gap (live).

P6 Composition

The Seal's assurance is the conjunction of the properties above. Each conjunct is independently verifiable; the Seal is their logical AND.

```
Assurance(Seal) = Unforgeable ∧ ActionBound ∧ Authorized ∧ NonBypassed ∧ Faithful ∧
HumanPresent
```

In plain terms. Weaken any one property and the reader can see exactly which. A Seal is not a single opaque badge; it is a bundle of checkable claims. The final conjunct, **HumanPresent** (proving the approver is a real, present human and not another agent), is the frontier property, delivered at Level 3 through a present-human, anti-mimicry step-up. It is the honest open edge (Section 10), and the place the current literature says least.

BASIS: P1, P2 VERIFIED in code and tested; P3, P4, P5 in the reference implementation; HumanPresent is Level 3 (roadmap).

7 Verifiability over assertion, and the authority playbook

A certification mark is only worth what its issuer's word is worth, unless the mark is **self-verifying**. The Human Seal is designed so that its authority does not rest on trusting HumanGate International or FLOCORE at all.

- **Verifiability over assertion.** Anyone can verify a Seal offline against the published key. No trust in us is required for the check to be conclusive (P1, P2).
- **Openness.** The Human Gate standard is free to implement, so there is no lock-in and therefore nothing to distrust in the incentive.
- **Neutrality.** The mark is stewarded by HumanGate International, a body designed to be separate from any single vendor.
- **A working reference.** FLOCORE implements the standard on its own governed actions, proving the standard is real and not paper.
- **Domain-bound issuer.** The issuer is bound to a domain by did:web (DNS plus TLS plus a published key) at `did:web:humanseal.world`, so a verifier can resolve the key over HTTPS and check any Seal.

What made other marks credible

Verifiability is necessary but not sufficient; a mark also has to be adopted. The precedents are consistent about how that happens.

Precedent	The lever that earned it authority
C2PA / Content Credentials	Heavyweight, cross-industry founders (2021, a group including Adobe, Arm, BBC, Intel and Microsoft); an open, royalty-free spec with a controlled, visible mark; the spec body separate from the adoption body.
PCI Security Standards Council	Founded 2006 by the five major card networks with equal-share governance; enforced not by law but as a condition of doing business (to accept cards, you comply).
FIDO Alliance	Platform-owner founders who could ship the standard in their own products; independent accredited-lab conformance; a mark tied to interoperability.
UL	Independent testing plus continuous factory re-audit; and, decisively, **insurer demand-pull** that made the mark commercially necessary.
CE marking	A mark tied to **market access** by law, with legal penalty for misuse. A mark that gates a market is stickier than a voluntary badge.
Fair Trade	No self-declaration; independent audit; public impact reporting; a consumer-facing mark with high recognition and trust.

The repeatable playbook: a neutral multi-founder body; an open standard with a controlled mark; independent conformance and continuous re-audit, never self-declaration; the mark tied to market access and, first, to insurers; and active enforcement so the badge stays scarce and meaningful. The insurer opening is already here: the NAIC AI Model Bulletin was adopted in 23 states plus DC as of 1 April 2026, requiring written AI governance programs; insurers are adding generative-AI exclusions (Verisk endorsements CG 40 47 and CG 40 48, effective 1 January 2026); and standalone AI insurers increasingly require audit trails as a coverage precondition `SOURCED`. A carrier that recognizes the Human Seal would do for it what insurers did for UL.

8 Prior art and honest novelty

Intellectual honesty is part of the authority argument, so we are precise about what is and is not new.

The primitives are standard, and that is a feature

Ed25519 (RFC 8032), SHA-256 (FIPS 180-4), the W3C Verifiable Credentials Data Model 2.0 (a W3C Recommendation since 15 May 2025), and the did:web method are all established, and the Seal reuses them deliberately. C2PA / Content Credentials is the closest structural analogue: a signed, hash-bound, offline-verifiable manifest, but for media provenance rather than a governance decision `VERIFIED`. Non-repudiation and tamper-evident audit trails are long-standing security properties (the ISO/IEC 27001 and 25010 lineage) `SOURCED`. Reusing court-tested primitives makes the Seal more credible, not less.

The gap in the governance standards

No mainstream AI governance standard defines a signed, per-action, human-approval receipt. ISO/IEC 42001 (AI management systems), ISO/IEC 23894 (AI risk guidance), the NIST AI RMF, and IEEE 7000/7001 all **require** human oversight and auditable records, but none specifies the cryptographic artifact that evidences a single human approval `VERIFIED as a gap`. The Seal is the concrete evidence those requirements call for.

The honest competitive read: a race, not a lead

Recent work is converging on signed AI action receipts. The nearest concrete format is an individual IETF draft (draft-marques-asqav-compliance-receipts, 2026), which signs the result of a **machine policy evaluation** (allow / deny / rate-limit) with Ed25519 over canonical JSON. It is not IETF-endorsed, and, decisively, it records what an **automated engine** decided, not that a named human approved `VERIFIED`. Academic proposals (institutional attestation; receiver-attested receipts) show the concept is mature and about to be contested.

> THE DEFENSIBLE CLAIM
>
> We do **not** claim the first signed AI receipt. We claim **a signed, verifiable receipt of a named human's approval, bound to the exact action, grounded in the organization's own modeled roles and data, built on proven standards.** The novelty is the composition and the human-approval semantics, not a new cryptographic primitive. The right posture toward the emerging IETF format is interoperability, a superset profile that adds the human-approver identity it lacks, not competition.

9 Mapping to law worldwide

The Human Seal is not built for one jurisdiction. It is built for a sentence that jurisdictions are independently arriving at: **a human must be able to oversee and intervene in a consequential automated decision, and there must be a record.** The gate is the oversight half; the signed receipt is the record half.

Instrument	What it requires	Status
EU AI Act, Art. 14 + 12	Human oversight of high-risk AI (understand, resist automation bias, interpret, refuse, stop), and automatic, traceable logging over the lifetime of the system. The gate answers Article 14; the signed receipt answers Article 12.	Phasing 2026–2027
Council of Europe Framework Convention on AI	The first legally binding international AI treaty (opened 5 Sept 2024): oversight, mandatory documentation, accountability, notice, and effective remedies to contest outcomes. As parties transpose it, the gate-plus-receipt-plus-contest chain becomes the treaty-level compliance shape.	Binding, in force
South Africa, POPIA s71	A person may not be subject to a decision based solely on automated processing with legal or material effect, and must have a path to human intervention. A human approval removes the "solely automated" trigger; the receipt evidences the intervention.	Live
Vietnam, AI Law 134/2025	Human oversight required in important decisions and for high-risk systems, with conformity assessment and registration. Effective 1 March 2026.	Live
Nigeria (NDPA), Kenya (DPA s35), Mexico (LFPDPPP)	Statutory rights against solely-automated high-impact decisions with human review (Nigeria, Kenya), and mandatory human oversight for hiring, credit and profiling (Mexico, in force 21 March 2025).	Live
OECD AI Principles; G7 Hiroshima; GPAI; ASEAN	The cross-border norms: "human agency and oversight" plus lifecycle traceability (OECD, 2024 update), traceability and incident documentation (G7), carried to dozens of jurisdictions.	Soft-law reference

Read across the current global map, roughly two dozen national and regional jurisdictions plus six cross-border instruments now converge on this requirement SOURCED. The single strongest anchor is the Council of Europe Framework Convention, because it is binding and it names oversight, documentation and the right to contest together. A Human Seal is close to a turnkey artifact for that chain. This mapping is provided for orientation and is **not legal advice**; conformance supports but does not by itself establish compliance, which depends on full system context and qualified counsel.

10 Threat model and honest limits

Stating the limits is part of the credibility. The Human Seal does **not** prove:

- **The quality of the human's judgment.** The Seal is evidentiary about the conditions of a decision (who, what, when, what was shown), not a guarantee the decision was wise. This is by design (Section 3): it lets rubber-stamping be detected, not prevented outright.
- **The legal correctness of a compliance pack.** Packs encode an industry's rules, but they require qualified review; a Seal certifies that an approval was grounded in the pack, not that the pack is legally perfect.
- **Safety against a fully compromised, authorized approver.** If a legitimate signer's key and identity are both captured, the Seal will verify. This is the same trust boundary as any signature scheme; the mitigations are key hygiene, the Level 3 present-human step-up, and second-approver checks for the highest-risk actions.
- **Zero egress during grounding.** Where an answer is produced with a hosted model, that path is a data-exposure surface; the mitigation is the on-box and distillation route, which is a program commitment, not a property of the Seal itself.

The **frontier limit** is the HumanPresent conjunct (P6): proving the approver is a real, present human rather than another agent. Meaningful human control scholarship and Article 14 both assume a human but say little about proving presence at the moment of approval. This is the genuine open edge, and Level 3 (present-human anti-mimicry, with a hardware or platform authenticator step-up) is our answer to it. We name it as the research frontier rather than claiming it solved.

11 Verify a Seal yourself

The clearest proof that authority is checkable is that you can check it. A Human Seal is a small JSON object. To verify one, a third party needs only the object and the issuer's public key (resolvable at the did:web trust root). No call back to the issuer is required.

1. Take the receipt. Remove the `signature_ed25519` and `public_key` fields; the remainder is the signed core.
2. Recompute the canonical form of the core: JSON with keys sorted and no insignificant whitespace. Any verifier reproduces the identical bytes.
3. Check the Ed25519 signature over those canonical bytes against the public key (which you can independently confirm against `did:web:humanseal.world`).
4. If you have the action, recompute `H(action)` and confirm it equals the receipt's `content_hash`. This proves the human approved **this** action.

If any step fails, the Seal is invalid, and it fails in a way that names the reason (unsigned, signature mismatch, or content-hash mismatch). The reference implementation ships exactly this algorithm; Appendix B gives the precise steps and Appendix A the receipt fields.

APPENDIX A

Appendix A. The receipt schema

The fields of a Level 2 Human Seal, as produced by the reference implementation. Personal data never appears; the approver is a stable reference, not a name or email.

action_id	Stable identifier of the action being approved.
action_type	The class of action (used for the authority check, P3).
content_hash	`"sha256:" + H(canonical action)`. Binds the Seal to the exact action (P2).
tenant	The organization whose model the approval was checked against (P3).
approver_ref	Stable identity reference of the approver. Never personal data.
decision	`approved` or `refused`. Refusals are sealed too.
decided_at	UTC timestamp of the decision (ISO 8601).
issuer	The did:web trust root, e.g. `did:web:humanseal.world`.
level	The assurance level (2 = Signed).
signature_ed25519	Ed25519 signature over the canonical core (all fields above). Excluded from the signed bytes.
public_key	The verification key, base64url. Confirmable against the did:web document.

APPENDIX B

Appendix B. The verification algorithm

The algorithm the reference implementation runs, in pseudocode faithful to the code. It is deliberately small so that any independent party can reimplement it.

```
# canonical(obj): JSON, sorted keys, no insignificant whitespace
function verify(receipt, action = null):
    if receipt.signature_ed25519 is empty or receipt.public_key is empty:
        return { valid: false, reason: "unsigned" }

    core      = receipt without { signature_ed25519, public_key }
    message   = canonical(core)                    # reproduce the signed bytes
    ok        = Ed25519(receipt.public_key).verify(receipt.signature_ed25519, message)
    if not ok:
        return { valid: false, reason: "signature check failed" }

    if action is not null:                          # optional action binding (P2)
        if content_hash(action) != receipt.content_hash:
            return { valid: false, reason: "content_hash mismatch (possible tamper)" }

    return { valid: true, approver_ref: receipt.approver_ref,
             decision: receipt.decision, level: receipt.level }
```

Any change to a signed field (approver, decision, action_type, timestamp) alters the canonical message and fails the signature check. Any swap of the action alters the content hash and fails the binding check. Both behaviors were exercised against the reference implementation while preparing this paper.

SOURCES

Source ledger

Tags: VERIFIED primary source read directly · SOURCED named, dated source via search summary · ASSUMED our synthesis. Items marked SOURCED should be confirmed against the primary text before use as a load-bearing legal quotation.

Meaningful human control and oversight

- Article 36 (2013), Killer Robots: UK Government Policy on Fully Autonomous Weapons. article36.org [SOURCED]
- Santoni de Sio and van den Hoven (2018), Meaningful Human Control over Autonomous Systems: A Philosophical Account, Frontiers in Robotics and AI. [VERIFIED]
- Siebert et al. (2022), Meaningful human control: actionable properties for AI system development, arXiv:2112.01298. [VERIFIED]
- ICRC, AI and machine learning in armed conflict: a human-centred approach; Inter-Parliamentary Union, Human autonomy and oversight. [SOURCED]

Why oversight can fail; the accountability gap

- Parasuraman and Manzey (2010), Complacency and Bias in Human Use of Automation, Human Factors. [SOURCED]
- Green (2022), The Flaws of Policies Requiring Human Oversight of Government Algorithms, Computer Law and Security Review. [SOURCED]
- Elish (2019), Moral Crumple Zones, Engaging Science, Technology, and Society vol. 5. [SOURCED]
- Matthias (2004), The responsibility gap, Ethics and Information Technology 6(3). [SOURCED]
- Nissenbaum (1996), Accountability in a Computerized Society, Science and Engineering Ethics. [VERIFIED]

Cryptography and standards

- RFC 8032 (EdDSA), IETF, Jan 2017. [VERIFIED]
- Brendel, Cremers, Jackson, Zhao (2021), The Provable Security of Ed25519, IACR ePrint 2020/823 / IEEE S&P 2021. [SOURCED]
- FIPS 180-4 (SHA-256), NIST, Aug 2015; NIST SP 800-107 Rev.1 (collision-resistance figures), Aug 2012. [VERIFIED]
- W3C Verifiable Credentials Data Model 2.0, Recommendation 15 May 2025; did:web Method Specification (W3C CCG). [VERIFIED]
- C2PA / Content Credentials, spec v2.4 (2026); founded 2021. [VERIFIED mechanism]
- ISO/IEC 42001:2023; ISO/IEC 23894:2023; NIST AI RMF 1.0; IEEE 7000/7001-2021 (require oversight and records, define no signed per-action human-approval receipt). [VERIFIED as a gap]
- draft-marques-asqav-compliance-receipts-08, IETF individual draft, 2026 (machine policy-decision receipts). [VERIFIED]
- ISO/IEC 27001 and 25010 non-repudiation lineage; audit-trail guidance. [SOURCED]

Law and market

- EU AI Act (Regulation 2024/1689), Articles 12 and 14. [VERIFIED]
- Council of Europe Framework Convention on AI (opened 5 Sept 2024, first binding AI treaty). [VERIFIED]
- South Africa POPIA s71; Vietnam AI Law 134/2025 (eff. 1 Mar 2026); Nigeria NDPA; Kenya DPA s35; Mexico LFPDPPP (in force 21 Mar 2025). [SOURCED]
- OECD AI Principles (2024 update); G7 Hiroshima Process; GPAI; ASEAN Guide. [SOURCED, G7 text VERIFIED]
- Gartner (Feb 2026), AI governance platform spend $492M in 2026, surpassing $1B by 2030. [VERIFIED via secondary]
- NAIC AI Model Bulletin (23 states + DC as of 1 Apr 2026); Verisk generative-AI exclusion endorsements CG 40 47 / 48 (eff. 1 Jan 2026); board / D&O liability trend. [SOURCED]

Implementation claims (P1, P2 tested; P3, P4, P5 in the reference implementation) were verified against the FLOCORE code while preparing this paper. Full research packs with every URL are held locally under docs/research/.