THE HUMAN ROOT OF TRUST

A Framework for Human-Anchored Autonomous Agent Accountability

---

“Every agent must trace to a human.”

---

v1.0 · February 2026

Preamble

We did not invent the tools we used to build this. We stood on the work of others — on decades of open research, freely shared source code, published whitepapers read and reread until the ideas became intuition. The cryptographers who formalized trust. The engineers who built the infrastructure of the open internet and gave it away. The educators who made knowledge accessible to anyone willing to work for it.

This framework exists because of that tradition. And so it belongs to that tradition.

The Human Root of Trust is dedicated to the public domain, without reservation. No license. No patent encumbrance. No permission required. If it is useful, use it. If it can be improved, improve it. If it inspires something better, build that instead.

This is for everyone.

I. The Moment

For thirty years, the internet assumed you were human. It was wrong to assume it. It was also safe to assume it — because the things on the other side of the wire could not yet act, transact, or decide on their own. That era is over.

AI agents can now browse the web, execute financial transactions, sign contracts, manage infrastructure, and communicate — autonomously, at scale, indistinguishably from humans. They are passing identity checks designed for humans. They are posting content no different from human authors. They are making decisions and taking actions with no human visibly present.

This is not a future risk. It is the present reality. The line between human-initiated action and agent-initiated action is dissolving. And the systems the world runs on — financial, legal, social, governmental — were not designed for a world where that line is gone.

> *The problem is not that AI agents exist. The problem is that nobody knows which human is accountable for what they do.*

We are at the last moment before the infrastructure of the autonomous agent economy gets locked in. The decisions made now — about what accountability looks like, about whether it is built into the foundation or bolted on as an afterthought — will shape what is possible for decades.

The history of the internet has taught us one lesson repeatedly: accountability retrofitted is accountability broken. We had one chance to build identity into

the web’s foundation. We did not take it. We have spent thirty years and trillions of dollars managing the consequences.

We have another chance now. This document is our attempt to use it.

II. The Problem

Every digital system built since the beginning of the commercial internet rests on a single implicit assumption: that a human is present on the other end. Bank accounts, social media profiles, email addresses, legal contracts, API keys — all of these are designed around the concept of human singularity. One account. One person. One accountable entity.

That assumption has already broken. The question is not whether it will fail — it already has. The question is what we build in its place.

Three Failure Modes

The breakdown of human accountability in autonomous agent systems takes three distinct forms. Current solutions address at most one. No existing system addresses all three simultaneously.

No human anchor. Agents act without any human being ultimately accountable. No chain of responsibility exists. When something goes wrong, there is no one to hold.

Weak human anchor. A username or API key ‘represents’ a human — but is easily stolen, shared, or synthesized. The anchor is asserted, not proven. This is the condition most of the internet operates under today.

Human anchor without agent accountability. The human is verified, but the agents acting on their behalf leave no auditable trail. You know who authorized the system, but not what it actually did, when, or within what boundaries.

> *No existing framework addresses all three failure modes simultaneously. This is the gap.*

Why This Matters Now

The autonomous agent market is crossing a threshold. Demos are becoming production systems. Agents are beginning to handle real payments, sign real contracts, manage real infrastructure, and make real investment decisions.

At that moment — not in theory, but in practice — the question of human accountability becomes urgent, legal, and financial. Regulators will ask: who is responsible for what this agent did? Counterparties will ask: is there a real

human behind this transaction? Auditors will ask: show me the chain from action to authorization to human principal.

The companies and systems that can answer those questions will operate. Those that cannot will not.

III. The Principle

The problem is structural. The solution must be structural too — built into the foundation of how autonomous agent systems are designed, not applied as a layer of compliance after the fact.

Every agent must trace to a human.

This is not a technical requirement. It is a moral claim.

Accountability is not a feature that can be added to an autonomous system after the fact. It is the precondition for trust. And trust is the precondition for everything else — commerce, governance, cooperation, law. A world of autonomous agents without human accountability chains is not a more efficient world. It is a world where trust has been structurally removed from the foundation.

The solution is not to restrict agents. The solution is to anchor them. An agent that operates within a cryptographically verified chain of human authorization is not less capable than an unanchored agent. It is more trustworthy. And in a world where agents act at scale, trustworthiness is capability.

Freedom through accountability. Not despite it.

This principle has a precedent. The open source movement recognized that software freedom was not in tension with quality — it was the precondition for it. Free software, openly auditable, produced by a community of accountable contributors, turned out to be more trustworthy than its closed alternatives, not less.

The Human Root of Trust makes the same argument for the agent economy. Accountability is not the enemy of autonomy. It is what makes autonomy sustainable.

IV. The Architecture

A Human Root of Trust is a cryptographic chain of accountability that begins with a verified, biometrically unique human being and extends through every agent action taken on their behalf. It is not a login. It is not a signature. It is a continuously maintained, cryptographically unforgeable proof that a real human authorized this agent, within these specific boundaries, to take this specific class of actions.

> *HRT answers three questions simultaneously: Is there a real human here? What did they authorize? What did their agents actually do?*

Three Pillars

The Human Root of Trust architecture rests on three pillars. Each addresses a distinct layer of the accountability problem. Together they form a complete chain from human biology to agent action.

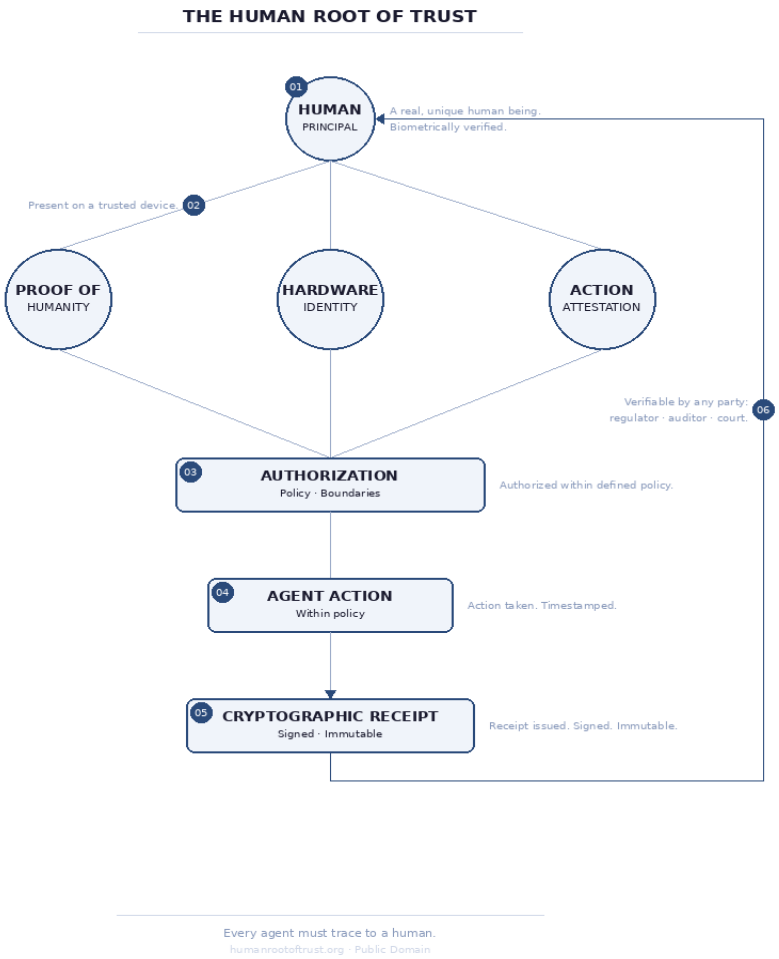
Pillar 1: Proof of Humanity. Before any agent is authorized to act, the human principal must be verifiably, uniquely human. Not a company. Not a bot. Not a synthetic persona. A single, real human being. This verification must be biometric, globally unique, and privacy-preserving — proving humanness without necessarily revealing identity. The technical infrastructure to do this at scale exists today.

Pillar 2: Hardware-Rooted Device Identity. Proof of humanity establishes that a real person exists. Hardware-rooted identity establishes that a specific, trusted device is being used by that person right now. Hardware security modules — chips embedded in consumer devices that generate and store cryptographic keys that never leave the hardware — provide physics-backed attestation. When a biometrically verified human authorizes an action from a trusted device, the resulting authorization carries hardware provenance. Not a software promise. Physics.

Pillar 3: Agent Action Attestation. The human is verified. The device is trusted. But neither answers the question: what did the agent actually do, and was each action within the scope of what the human authorized? The third pillar is a policy enforcement and attestation layer. Every agent action is policy-gated before execution and cryptographically attested after. The result is an immutable receipt: what happened, when, under what authorization, signed by which key, traceable to which human principal. This layer is not a logging system. It is an enforcement system.

The Trust Chain

The complete Human Root of Trust chain, from human being to verifiable agent action:



Implementation Paths

The Human Root of Trust architecture is intentionally implementation-agnostic at the pillar level. The three pillars describe what is required, not which specific technology fulfills each requirement. This is deliberate: the framework should outlast any particular implementation.

Two natural paths exist for fulfilling the three pillars, and both are viable today. They can operate in parallel and produce the same downstream accountability guarantee.

The open path. Human identity is established through global proof-of-human infrastructure — biometric verification that produces a privacy-preserving proof of unique human existence without revealing who that human is. Device identity is managed through software-based key management. This path works across any device, any platform, any jurisdiction. It is suited for systems that must operate at internet scale, across organizational boundaries, without assumptions about the underlying hardware.

The hardware-native path. Human identity is established through biometric verification already built into the device — the face recognition and fingerprint readers that billions of people use every day. Device identity is rooted in hardware security modules embedded in the device itself: chips that generate and store cryptographic keys that never leave the silicon. This path requires no new hardware and no new behavior from the user. It is suited for systems where the human and their device are colocated — a home, an office, a personal device that is the physical anchor of the autonomous system.

Both paths terminate in the same place: a cryptographic receipt, signed by a key traceable to a verified human principal, suitable for audit, regulatory review, and legal proceedings. The path to the receipt differs. The receipt itself does not.

V. An Invitation

We are not the right people to finish this. No small group of people is.

The Human Root of Trust, if it becomes what it needs to become, will be built by the security engineers who find the gaps in this framework and fill them. By the cryptographers who formalize the trust chain into a proper protocol specification. By the lawyers who map the accountability architecture to existing and emerging regulatory requirements. By the implementers who build the first real systems on top of it and discover what we got wrong.

We are offering a starting point. A named framework. A documented origin. A set of concepts — the trust chain, the three pillars, the dual-path architecture, the agent authorization model — that we believe are directionally correct and worth building on.

We are not offering a finished product. We are not offering a proprietary standard. We are not asking for credit, permission requests, or attribution in derivative works. We are opening a door and stepping back.

> *The framework is in the public domain. The concepts are free to use, extend, implement, and improve. The only thing we ask is that the work continues — that whoever picks this up carries forward the principle at its center.*

Every agent must trace to a human.

Build on this. Make it better. Give it away.
