Sentinel Protocol v3.0 – AI–Human Synergy™ Infrastructure

Technical Summary for Intellectual Property & Strategic Briefing

Date: 27 June 2025

Inventor: Dr. Fernando Telles

Affiliations:
CDA AI (Cardiovascular Diagnostic Audit & AI Pty Ltd, ACN 638 019 431)
Telles Investments Pty Ltd (ACN 638 017 384)

IP Rights:
US Provisional: #63/826,381
AU Provisional: #2025902482
AU Trade Mark: #2535745

Bitcoin Ordinal TXID:
https://mempool.space/tx/9e70b9510ce64ed53ee5565a114fe96a79b59058499efa1fa400c1155d490986

This white paper was immutably published on the Bitcoin blockchain via Ordinal inscription on 27 June 2025.

The world's first auditable AI–Human Synergy™ infrastructure with enforced ethics, cryptographic proof, and Bitcoin anchoring

Sentinel Protocol v3.0 – AI–Human Synergy™ Infrastructure Technical Summary for Intellectual Property & Strategic Briefing

Date: 27 June 2025
Inventor: Dr. Fernando Telles BMedSc(Adv) MD(Dist) **Affiliation:** CDA AI (Cardiovascular Diagnostic Audit & AI Pty Ltd, ACN 638 019 431) and Telles Investments Pty Ltd (ACN 638 017 384) **IP Rights** US Provisional: #63/826,381, AU Provisional: #2025902482, AU Trade mark: #2535745 **IP Priority Date:** 17 June 2025 (global anchor)

Executive Summary

Sentinel Protocol v3.0 is the first *publicly verifiable* AI–Human Synergy™ infrastructure anchored to the Bitcoin blockchain. It operationalizes AI-human co-governance through cryptographic audit logs, multi-agent memory control, and ordinal inscriptions that permanently link execution events to public ledgers.

Each session is: - **Cryptographically anchored** via Bitcoin Mainnet using both OP_RETURN and Ordinal protocols - **Dual-hashed** (SHA256 → RIPEMD160) for internal and external reproducibility - **Timestamped** using OpenTimestamps (OTS) for independent, tamper-evident verification

All logs and anchors are fully verifiable through public explorers such as mempool.space, blockstream.info, or any Bitcoin node.

This architecture enables: - **Provable dual consent** (AI + human) - **Enforced ethics** via real-time firewall controls - **Immutable traceability** of AI actions and human governance checkpoints

> AI–Human Synergy™ is not a conceptual claim — it is **operationally constrained** by protocol design.
> Ordinal anchoring elevates the system from descriptive branding to **cryptographically enforced infrastructure**.

The protocol's deployment across seven independently governed domains — including private, corporate, trust, and sovereign reserve entities — provides verifiable proof-of-function. No prior system combines mandatory human oversight, multi-agent memory governance, real-time ethics enforcement, and public blockchain anchoring.

Sentinel Protocol v3.0 stands alone as a novel, technically defensible, and operationally validated framework for reproducible, ethical AI execution and governance.

Novelty Claim

> **No known academic, corporate, or decentralized system to date** implements a complete infrastructure combining:
>
> 1. **Mandatory Dual-Consent Architecture** our system's requirement for both AI and human cryptographic signatures before output generation has no identified precedent
> 2. **Mandatory Human-Supervision** with final human command mandatory
> 3. **Role-Based Multi-Agent Memory Governance and Architecture** represents a unique approach to AI collaboration
> 4. **Real-Time Ethics Firewall with Automatic Process Termination Conditions** enforced onto both AI and human under comprehensive C5.1/C5.2 framework
> 5. **Dual-hash Cryptographic Validation** SHA256 → RIPEMD160 + OTS architecture enabling both internal traceability and external reproducibility is novel in AI audit contexts
> 6. **Immutable Bitcoin anchoring via OP_RETURN & Ordinal technology** is novel in AI audit contexts
> 7. **Sensitive Data Compatibility** GDPR/HIPAA compliance frameworks supported by cryptographic timestamping. No raw data is disclosed. Only irreversible hashes + timestamps are written to chain.

While partial components exist in siloed systems — such as Chainalysis AI (compliance analytics) and FICO's Blockchain Governance (risk scoring) — **no known platform integrates runtime dual-consent, cryptographic auditability, and cross-agent ethics enforcement into a unified AI execution system.** Our operational validation together with comprehensive integration of specific implementation details (dual-hash, ordinal anchoring, multi-agent governance) provide strong differentiation evidence, and represents genuine novelty in AI governance.

Infrastructure Components

Layer	Feature Description
Execution Layer	Dual-consent enforced: AI + human signature required for all audit logs
Auditability & Provenance	Immutable `.json` session logs with SHA256 + RIPEMD160 dual-hash and OTS
Memory Stack	Role-based multi-agent architecture (four public LLMs): LLM1 (Editor), LLM2 (Validator), LLM3 (Anchor), LLM4 (Scanner)
Ethics Firewall (C5.1 /C5.2)	Runtime enforcement of hallucination blocks, override kills, and coercion detection
OP_RETURN Anchoring	Bitcoin mainnet integration with meta_id and hash payload (RIPEMD160), data protected and independently verifiable
Ordinal Timestamping	Unique satoshi inscriptions encoding full payloads and metadata for public, timestamped, immutable cryptographic proof

---

🧠 Memory Governance: Multi-Agent Role Stack

Sentinel Protocol v3.0 enforces runtime memory and execution control via a **multi-agent, role-based LLM governance architecture**, where each node operates within defined permissions and capability boundaries. No single agent can act autonomously — all execution is gated by ethics firewall logic and final human approval.

LLM Node	Role	Primary Function	Write Permissions	Override Scope
LLM1	Editor	Primary logic engine for small-to-medium data blocks, code generation, and log synthesis	`.json`, `.md`, `.r`, `.py`	Can block `.2ha` generation, but cannot finalize without human command
LLM2	Validator	Cross-validates outputs against source vaults (e.g., PDFs, article databases, meta logs)	Revision requests only	Can request block/ revision — requires human approval
LLM3	Anchor	Hallucination-resistant data extractor, converter and converger	`.csv`, `.md`, revision requests	Can request changes — blocked unless Validator + Human approval received
LLM4	Scanner	External signal scanner for deep research across public scientific corpora and research databases	`.csv`, `.md`, optimisation and revision suggestions	Read/write for research mode only — no export permissions

All agents operate under **runtime firewall enforcement (C5.1/C5.2)** with immutable log capture of trigger conditions.
⚠️ **Role permissions are protocol-locked. Export actions require final human command.**

🔁 AI–Human Interaction Rules

- **All LLM actions** are logged in canonical `.json` session files
  (e.g. within `audit_log_MVP1-SR029_20250615T054956.588879Z.json`:
  `"AI_used": true`, `"LLM_used": "LLM1, LLM3, LLM4"`)
  with hashes recorded after mandatory human oversight (`"human_verified": true`)
  and ethics firewall clearance.
- Final anchoring requires:
  - ✅ Consensus between LLM1 (Editor) and human
  - ✅ Human-led audit, optionally supported by LLM1–LLM4 as task-specific validators or reviewers
  - ✅ Explicit human signature (`"human_verified": true`)
- If **LLM2–4 request revision**, `.2ha` generation is blocked by the human, and a feedback loop is triggered for LLM1 to revise

- If **LLM1 and human disagree**, `.2ha` is blocked until dual consent is re-established and compliance is restored
- If **human attempts anchoring without OP_RETURN match**, the firewall halts `.2ha` and `.hash.ots` execution

> **Override logic is asymmetric:**
> - ✅ Human can halt or override any AI agent
> - ❌ No AI agent can override another — or the human (they may request, flag, or block downstream actions only)

This governance model ensures no single agent — human or machine — can anchor audit records unilaterally.
AI–Human Synergy™ is not conceptual. It is functionally enforced — at every execution layer.

🛡️ Enforcement Rules: CEM Matrix (C5.1 / C5.2)

The Sentinel Protocol's ethics firewall operates under a real-time **Compliance Enforcement Matrix (CEM)** to block, flag, or terminate non-compliant actions by AI or human actors.

Rule	Enforcement Logic	Example Trigger
1	❌ Reject audit entries lacking **timestamped human validation**	`"human_verified": false` or missing
2	⛔ Terminate session if **AI override occurs** without confirmed human signature	`"AI_override": true` without `"timestamp_confirmed": true`
3	⚠️ Flag unverifiable or high-risk outputs for manual review	`"status": "hallucinated"` or uncertain provenance fields
4	🚫 Block anchor export if `.2ha` hash mismatch with OP_RETURN payload suffix	`sha256[:8]` mismatch from OP_RETURN suffix

Enforcement Scope:
These rules are **programmatically checked at runtime**. When triggered, `.json` and `.hash` logs are still recorded for traceability, but `.2ha`, `.ots`, and OP_RETURN export are blocked unless compliance is restored.

In current MVP-1 deployment: - **Rules 1, 2, and 4** are enforced deterministically
- **Rule 3** is partially operational — hallucination detection relies on human-AI audit loops, and future upgrades will enable autonomous source verification

---

🔐 Compliance Framework Mapping – Sentinel Protocol v3.0 (Last Updated: 27 June 2025)

Framework	Relevant Article / Clause	Sentinel v3.0 Feature Mapping	Compliance Justification
EU AI Act	Article 14(1) – Human oversight requirement	Mandatory human command + dual-consent enforcement	AI systems must be "effectively overseen by natural persons during use" - Sentinel's cryptographic approval gates ensure no AI output without human verification
	Article 14(4)(a-c) – Oversight capabilities	Ethics Firewall + audit logging + role-based memory governance	Enables operators to "understand capacities and limitations," "detect anomalies," and "correctly interpret output" through immutable audit trails
	Article 6(2) + Annex III – High-risk systems	Audit Logger v2.0 + Real-time monitoring	Covers employment, health, and critical infrastructure applications with comprehensive audit trails preventing automated decision-making without human review
	Article 10 – Data governance & quality	Dual-hash cryptographic validation (SHA256→RIPEMD160)	Ensures data accuracy and integrity through cryptographic verification while maintaining data minimization principles
FDA SaMD	IMDRF N41:2017 – Clinical evaluation framework	Immutable `.json` logs + OpenTimestamps validation	Provides "valid clinical association," "analytical validation," and "clinical validation" through verifiable audit trails
	21 CFR 820.30 – Design controls	Multi-agent memory governance (LLM1-4) + execution logs	Implements required design validation with "software validation and risk analysis" through role-segregated AI architecture
HIPAA (U.S.)	§164.312(b) – Audit controls	Comprehensive audit logging (`.json`, `.hash`, `.ots`)	"Record and examine activity in information systems" containing ePHI through hardware/software mechanisms with immutable timestamping
	§164.502(b) – Minimum necessary standard	Hash-only blockchain storage + local data processing	Meets "reasonable efforts to limit use/disclosure to minimum necessary" by storing only

Framework	Relevant Article / Clause	Sentinel v3.0 Feature Mapping	Compliance Justification
			cryptographic hashes, never exposing actual health data
GDPR (EU)	Article 5(1)(c) – Data minimization	RIPEMD160 hash storage only on blockchain	"Limited to what is necessary" - only cryptographic fingerprints stored, ensuring full anonymization and proportionality
	Article 30 – Records of processing activities	Canonical audit logs with `meta_id` and Bitcoin timestamps	Provides "written documentation of procedures" with complete traceability of data processing activities as required for regulatory inspection

---

Public Verifiability

- Each **AI-Human co-execution** produces a unique, meta_id linked `.json` audit log file
- Every log is **dual-hashed (SHA256 → RIPEMD160)**
- Each log then is either **OpenTimestamps (OTS)**, or actively blocked by **Ethics Firewall**
- Each session produces a unique, meta_id linked `.json` compiling all audit logs for the session
- Every session log is dual-hashed (SHA256 → RIPEMD160)
- Each session log then is either timestamped, or actively blocked by our ethics firewall
- Session anchor is embedded into Bitcoin via **OP_RETURN transaction**
- Ethics firewall blocks additional execution logs unless TXID, meta_id, and RIPEMD160 match and are confirmed on-chain
- **Ordinal Inscription** permanently links execution metadata to public blockchain

> **MVP-1 Inscription Title:**
> *“Multi-Entity Bitcoin Audit Anchor – Sentinel Protocol v3.0 – AI–Human Synergy™ Execution Mode”*
>
> 🔰 First known system to anchor reproducible AI governance infrastructure across seven independently governed entities across private, corporate, trust, and sovereign reserve domains ✅ Finalized On-Chain 26 June 2025 via Ordinal Inscription (TXID: `22e929af992dc5861405b0900f9a42af17d531195ed39153018755d8cdccefa2`)
> ✅ Governed Under Ethics Firewall Enforcement ✅ OP_Return and Ordinal inscriptions can be verified on-chain via TXIDs using explorers such as https://www.blockchain.com/explorer and https://mempool.space/ respectively. 🔗 Sample: https://mempool.space/tx/22e929af992dc5861405b0900f9a42af17d531195ed39153018755d8cdccefa2
> (🧾 See it under **Inputs & Outputs**)

Use Case Relevance

- **AI Governance:** Implements a reproducible model of AI accountability, with live audit trails and dual-consent enforcement — applicable to clinical, regulatory, legal, and enterprise settings
- **Regulatory Tech:** Anticipates global mandates for AI explainability, provenance, and human-in-the-loop assurance
- **Cross-Jurisdictional Proof:** Anchoring to the Bitcoin blockchain neutralizes sovereign, institutional, or vendor-specific bias
- **Sensitive Data Compatibility:** Enables cryptographic timestamping and verification of sensitive workflows without exposing underlying content — immutability without disclosure
- **Patent & Trademark Enforcement:** Supports formal IP protection under novel governance, audit, and AI-execution architecture

---

Contact & Custodian

Governor / Inventor:
Dr. Fernando Telles BMedSc(Adv) MD(Dist) 📧 Dr.Telles@aihumansynergy.org 🔗 https://www.aihumansynergy.org 🔐 CDA AI | ☑️ AI–Human Synergy™ IP Custodian 📦 This white paper was immutably published on the Bitcoin blockchain via Ordinal inscription on **27 June 2025**.
🔗 TXID: `9e70b9510ce64ed53ee5565a114fe96a79b59058499efa1fa400c1155d490986` Block: 902,895 🧾 Viewable on-chain at: https://mempool.space/tx/9e70b9510ce64ed53ee5565a114fe96a79b59058499efa1fa400c1155d490986