HAIA-RECCLIN Data Immunization Protocol (v1.0)

Purpose:  
Establish a transparent, multi-AI governance process that detects, isolates, and prevents poisoned or anomalous data from entering shared datasets, fine-tuning stages, or public repositories.

Maintainer:  
Basil C. Puglisi – Human-AI Collaboration Strategist  
Part of the HAIA-RECCLIN ecosystem for governed multi-AI collaboration.

1. Context & Rationale

Recent research from the UK AI Security Institute, Anthropic, Oxford, and the Alan Turing Institute (Souly et al., 2025) shows that as few as 250 poisoned samples can compromise an LLM regardless of scale. This protocol defines how the HAIA-RECCLIN framework mitigates that risk by using cross-AI redundancy and human arbitration to validate data integrity before it enters any training or fine-tuning pipeline.

2. Objectives

- Detect anomalies and data poisoning attempts across model inputs and outputs.  
- Cross-validate data consistency using independent AI agents.  
- Maintain human arbitration as the final checkpoint.  
- Provide auditable logs for all verification stages.  
- Offer a replicable public standard for open-source AI governance.

3. Governance Structure

Role | Function | AI Agent Example  
------|-----------|------------------  
Researcher | Gathers and verifies raw sources | Perplexity  
Editor | Ensures factual and contextual coherence | Claude  
Coder | Handles transformation and preprocessing scripts | Gemini  
Calculator | Validates metrics, percentages, and quantitative claims | ChatGPT (Math mode)  
Liaison | Cross-references inter-AI outputs and flags inconsistencies | Grok  
Ideator | Generates synthetic test scenarios for data stress testing | Gemini / Mistral  
Navigator (Human Arbiter) | Final approval, dispute resolution, and data release authorization | Basil or designated reviewer

4. Immunization Workflow

Step 1 — Dataset Ingestion  
- All datasets are pulled through at least two independent AI agents.  
- Each AI independently classifies samples as benign, suspicious, or malformed.  
- SHA-256 hashes of raw files are recorded to detect post-analysis drift.  
  
Step 2 — Cross-AI Consistency Scanning  
- Each AI model compares its classification to peer reports.  
- Any disagreement >10% between models flags the batch for human arbitration.  
  
Step 3 — Anomaly Detection Layer  
- Deploy multi-AI pattern detection for linguistic triggers, token repetition, or structural outliers.  
- Apply open-source anomaly tools (e.g., T-Miner 2021; Kolouri 2020) for confirmatory scanning.  
  
Step 4 — Human Arbitration  
- The human Navigator reviews flagged content, evaluates context, and either:  
 1. Confirms false-positive, approves data.  
 2. Confirms risk, quarantines and reports sample.  
- All decisions recorded in /logs/immunization\_decisions.json with timestamp and rationale.  
  
Step 5 — Immunized Dataset Release  
- Only batches cleared through both AI and human arbitration are marked verified: true.  
- Released to public or internal repo with an attached Data Immunization Certificate (digital signature + verification log).

5. Metrics & KPIs

Metric | Target | Description  
---------|---------|-------------  
Cross-AI Agreement Rate | ≥95% | Concordance across independent AIs on data integrity classification  
False-Positive Rate | ≤5% | Percentage of benign data wrongly flagged  
Detection Confidence | ≥0.9 | Probability of correctly identifying poisoned or anomalous inputs  
Human Arbitration Latency | <48h | Turnaround time for flagged data review  
Audit Trail Completeness | 100% | Every dataset change recorded with origin and decision path

6. Current Implementation Status

- Live pilot across three AI systems using identical prompts and comparative scoring (Basil C. Puglisi’s ongoing cross-AI research test).  
- Manual logging of output divergence patterns.  
- Early indicators show cross-AI inconsistency detection at ~93% accuracy.  
- Phase 2 will integrate automated reconciliation scripts and visual dashboards.

The active research environment currently runs identical prompts across ChatGPT, Claude, and Gemini. Divergence analysis is manually logged to test inter-AI reliability and bias exposure. This phase validates the Immunization Protocol as a measurable framework for Human-AI collaborative consistency.

7. Community Participation

Open-source contributors are invited to:  
- Fork and implement the protocol within their multi-AI research environments.  
- Submit pull requests with test data, new anomaly-detection heuristics, or visualization tools.  
- Participate in monthly Integrity Audit Sessions via GitHub Issues.

8. License

This protocol accompanies The Haia Recclin Model (Puglisi, 2025) as part of the HAIA Governance Series. For citation or academic reference, please use: Puglisi, B. C. (2025). HAIA-RECCLIN Data Immunization Protocol (v1.0). HAIA Governance Research Series. MIT License.

Released under MIT License for maximum transparency and reuse. All derivative protocols must preserve attribution to HAIA-RECCLIN Data Immunization Protocol (v1.0).

This protocol accompanies The Haia Recclin Model (Puglisi, 2025) as part of the HAIA Governance Series. For citation or academic reference, please use: Puglisi, B. C. (2025). HAIA-RECCLIN Data Immunization Protocol (v1.0). HAIA Governance Research Series. MIT License.

Together, these mechanisms transform AI safety from reactive correction into structured governance and measurable trust.