White Paper: Toward a Framework for Trust Engineering

Author: Haig Hovaness

Date: May 28, 2025

Related Project: AI-Based Propaganda Detection Engine for Public Trust

Abstract

This white paper introduces *trust engineering* as an emerging discipline that treats trust not as a philosophical abstraction, but as a measurable, auditable property of digital information systems. Powered by recent advances in artificial intelligence, trust engineering aims to embed transparency, accountability, and resilience directly into the architecture of public communication infrastructure.

As a practical application, we present a prototype *Propaganda Detection Engine (PDE)*—a modular system designed to detect and evaluate manipulative content using explainable AI, provenance tracking, and dynamic trust scoring. The PDE serves as both a technical testbed and a civic prototype for trust engineering in action.

This work proposes a scalable, open, and ethically grounded approach for strengthening the credibility of public discourse. By translating trust into quantifiable system behavior, we enable institutions, platforms, and communities to resist manipulation and promote truthful, resilient communication in an era of information warfare.

1. Introduction: From Ethics to Engineering

Trust has traditionally been treated as a subjective, context-dependent ideal—addressed through ethics, journalism, or political theory. But as algorithmic systems increasingly mediate public discourse, it becomes essential to treat trust as a *system-level performance variable*: one that can be measured, modeled, and optimized with engineering rigor.

The *Propaganda Detection Engine (PDE)* is a proposed application of this shift. It treats trust as an attribute of information integrity—evaluated through metrics like entropy, novelty, semantic coherence, and source reliability. These metrics enable reproducible assessments of content trustworthiness across diverse digital environments.

We define *propaganda* as the deliberate and systematic use of emotionally charged, misleading, or manipulative messaging to influence beliefs or behavior—typically by distorting the truth and suppressing dissenting viewpoints. Propaganda is intrinsically harmful: it suppresses truth, polarizes populations, and at its worst, incites violence.

By contrast, a well-designed PDE operates with an informational advantage. Propaganda campaigns are inherently narrow in focus, often recycling rhetorical patterns or

amplifying specific narratives. A PDE with sufficient media coverage and computational capacity can detect these patterns at scale—leveraging its broader perspective to classify manipulative content and reduce its impact.

This paper introduces the principles of trust engineering and outlines the architecture and components of a PDE designed to operationalize those principles. Later sections present real-world use cases, engineering methodologies, and a feasibility analysis grounded in current AI capabilities.

The stakes are high. In the absence of trusted public information systems, manipulative campaigns can inflame conflict and destabilize democracies. But if technologists, civic institutions, and public-interest organizations collaborate to operationalize trust, we can reduce the influence of propaganda and advance a more peaceful, informed public sphere.

PDE feasibility is not hypothetical. Real-world systems such as GDELT and Media Cloud already demonstrate the technical capacity to ingest and analyze massive volumes of content across languages and platforms. Section 9 provides operational benchmarks that support the viability and scalability of a prototype PDE using current AI infrastructure.

The consequences of unchecked propaganda are visible in ongoing global crises, from regional wars to democratic backsliding. Without effective methods to validate public information and counter coordinated manipulation, societies risk drifting toward persistent instability. Trust engineering offers a pathway forward: by making truth traceable and manipulation visible, we can reduce the strategic utility of propaganda in the digital age.

2. Core Principles of Trust Engineering

Trust engineering begins with a shift in perspective: trust is not an emergent byproduct of responsible behavior but a *designable attribute* of information systems. Like performance, security, or resilience, trust can be intentionally shaped through architecture, metrics, and governance. Five core principles guide this transformation:

Trust is constructible
Trust can be *built* through intentional system design. By embedding structural incentives for transparency, source reliability, and algorithmic explainability, information systems can be engineered to promote credibility and resist manipulation—just as we engineer for safety or fault tolerance.

Trust is measurable
Trust is not a moral abstraction. It can be quantified using reproducible indicators such as informational entropy, factual consistency, narrative coherence, and behavioral history. These variables allow systems to assign dynamic *trust scores* to content and sources.

Trust is dynamic
The credibility of a source or claim changes over time. Trust systems must account for defection, evolution, and contextual shifts. *Time-weighted scoring* and behavioral drift detection ensure that trust assessments adapt to real-world developments rather than relying on static reputational baselines.

Trust is auditable
To be credible, trust decisions must be explainable. Every score, label, or suppression action should be supported by a transparent trail: logged, inspectable, and subject to independent verification. *Auditability* is essential for legitimacy in high-stakes information environments.

Trust is multidisciplinary
Engineering trust is not only a technical challenge—it is a civic one. Effective trust systems require collaboration among technologists, journalists, ethicists, legal scholars, and impacted communities. Governance frameworks must balance algorithmic design with democratic accountability.

3. System Components of an Information Trust Engine

The Trust Engine is composed of modular components that work together to ingest content, assess trustworthiness, detect manipulation, and visualize narrative patterns. Each subsystem is independently upgradable, allowing the system to evolve in response to new techniques, content formats, and adversarial tactics. The components fall into four primary functional domains:

Signal Ingestion and Normalization

Ingestion
The system must collect raw data signals from a broad array of sources, including social media platforms, messaging apps, forums, and video content. These inputs are standardized through a normalization process that removes noise, aligns formats, and translates content into a common analytical framework.

Provenance Tracking
This layer logs the origin, chain of custody, and transformation history of each piece of content. It establishes a verifiable record to support source credibility assessments and accountability for manipulation.

Analysis & Scoring Layers

Trust Metric Assignment Engine
Assigns and adjusts trust scores for sources, narratives, and actors using reputation data, historical behavior, and expert feedback. The model must be transparent, adaptive, and resilient to adversarial manipulation.

Multimodal Propaganda Classifier
Evaluates text, images, video, and audio to detect propaganda markers such as emotional manipulation, disinformation tropes, and coordination patterns. It is continuously updated with adversarial examples and cultural training data.

Entropy / Narrative Novelty Analysis
Measures the informational entropy of content to detect repetitive, templated, or recycled messaging. Flags narratives likely produced by coordinated campaigns or automated actors.

Source Behavior Profiling
Continuously tracks semantic drift, ideological shifts, and behavioral patterns of content sources. Enables adaptive trust scoring over time.

Claim Verification Layer
Extracts key factual assertions and checks them against curated knowledge bases for probabilistic validation. Highlights content needing manual review.

Signal-to-Noise Evaluation Engine
Synthesizes inputs from upstream modules to differentiate meaningful information from spam, distortion, or irrelevant content. Assigns a trust rating with transparency tags.

Control Interface

Dashboard and Visualization Controls
A real-time interface that enables analysts, researchers, and oversight personnel to monitor system outputs. Dashboards provide visual summaries of narrative trends, signal volatility, amplification networks, and trust score distributions. Users can filter by content type, region, platform, or score threshold, supporting both high-level situational awareness and granular forensic review.

Configuration Management Panel
The central administrative interface for adjusting operational parameters across the Trust Engine. Administrators can set trust thresholds, enable or disable scoring modules, adjust model weights, define source tiers, and manage ingestion pipelines. All changes are logged to the audit trail and subject to role-based access controls to ensure traceability and governance compliance.

User Access and Permissions Manager
Controls access and authorizes functions according to security policies. All user access and activity will be logged.

Audit and Intervention Tools
Facilities for reviewing system internal data, including signal tracing, decision logic,

and output data. Administrative tools for correction of defects or anomalies detected in the audit process.

Governance & Oversight Components

Score Governance Engine
Oversees the logic behind composite trust score calculation. Applies policy-based weighting rules, enforces consistency across scoring modules, and initiates alerts when anomalies or outlier patterns are detected in trust assessments.

Claim Volatility Monitor
Tracks fluctuations in the credibility of factual assertions over time. Identifies claims that show rapid score instability, enabling prioritization for manual review, recalibration, or suppression based on evolving context.

Policy Enforcement Ruleset
Applies predefined and adaptive policies to enforce platform-specific or institutionally mandated rules. Governs escalation thresholds, content labeling protocols, and intervention triggers in alignment with governance objectives.

Cross-Model Comparison Module
Validates scoring outcomes by comparing results from multiple analytical models. Supports redundancy, error checking, and model drift detection to improve robustness and mitigate algorithmic bias or overfitting.

.

The modular structure outlined above provides a conceptual foundation for the trust engine. This design approach also enables efficient integration of artificial intelligence tools. Because each functional block—such as Entropy Analysis, Claim Verification, and the Fact-Resilience Estimator—has a clearly defined scope and interface, implementation can proceed in a distributed and iterative manner. Domain-specific AI models can power individual components, while the Governance and Oversight layer provides a unifying structure for coordination, feedback, and accountability. This design facilitates not only implementation at scale, but also collaborative development across institutions. The implementation architecture will be guided by the engineering methodologies described in the following section.

4. Engineering Methodologies

The successful implementation of a trust engine depends not only on what it does, but on how it is built. To ensure adaptability, transparency, and resilience, the development process must follow modern engineering best practices. The following methodologies are foundational:

Modular, Open-Source Architecture

The system is designed for modularity. Components such as classifiers, scoring engines, and ingestion layers are independently upgradable and replaceable, allowing rapid iteration and targeted improvements without disrupting the whole system. Open-source construction fosters transparency, peer review, and a diverse contributor base committed to public-interest goals.

Explainable AI

Trust in the system depends on interpretability. Models must not operate as black boxes. Instead, explainable AI methods should be used to make trust scores understandable to end users, auditors, and maintainers. This includes transparency in feature attribution, decision pathways, and model confidence.

Time-Weighted Scoring Functions

Because trust is dynamic, recent behavior must influence scoring more than distant history. Time-weighted scoring ensures that reputational shifts, source defection, and evolving narrative patterns are reflected in near real-time. This is especially critical in fast-moving information environments like social media.

Consensus Validation Protocols

To mitigate algorithmic bias and overfitting, trust scores should be validated across multiple models or datasets. Redundancy and cross-comparison provide error correction and strengthen fairness. This is especially important in diverse cultural or linguistic contexts where single-model heuristics may fail.

Audit Trail Infrastructure

Every decision made by the system—whether scoring, filtering, or classifying—must be logged in a secure, queryable format. These logs enable traceability, third-party auditing, and institutional accountability. The audit trail should support both automated and human review.

Supporting Technical Infrastructure

In addition to core development methodologies, the trust engine relies on carefully structured data sources, robust machine learning workflows, performance evaluation, and hardened security protocols.

Data Types and Sources

Ingested content includes text, images, audio, and metadata streams drawn from social platforms, messaging apps, RSS feeds, verified databases, and public records. All data intake must comply with transparent usage policies and undergo legal and ethical review for privacy, consent, and relevance.

Machine Learning Infrastructure
Model training employs federated learning frameworks and synthetic data augmentation to preserve user privacy and improve generalizability. Multilingual and culturally sensitive embeddings are emphasized to ensure global applicability. Modular model deployment allows for rapid iteration and fine-tuning of components.

Evaluation and Metrics
System effectiveness is evaluated using multiple lenses: detection accuracy, false positive/negative rates, response latency, community feedback, and successful intervention outcomes. Regular transparency reports should publish these metrics to support public accountability and trust.

Security Protocols

The system must be secured against tampering, data breaches, and adversarial interference. End-to-end encryption, role-based access controls, audit logging, and routine penetration testing are required. Ethical hacking engagements and red-team simulations should be scheduled prior to major release milestones.

Together, these methodologies and infrastructure elements provide a blueprint for building a scalable, auditable, and adaptive trust engine. By aligning system design with public-interest values and engineering rigor, the framework enables real-world applications that go beyond propaganda detection and extend to broader domains of public communication integrity.

5. Technical Architecture: Signal Lifecycle and System Interactions

The Trust Engine is designed as a modular, multi-layered system that processes incoming information signals through structured analytical stages. These stages interact to assess credibility, detect manipulation, estimate long-term trustworthiness, and support governance and accountability. This section outlines how components function together across the signal's lifecycle.

This architecture supports scalable implementation through *modular AI integration*, allowing advanced models to power signal analysis, trust estimation, and system adaptation while remaining governed and auditable.

5.1 Signal Ingestion and Normalization

Information enters the system from various sources — including news articles, social media content, public records, and statements — through a unified ingestion layer.

Signals are first standardized to a common internal format. Metadata is extracted (e.g., timestamps, source identity, publication context), and provenance is captured to enable traceability.

Key components include:

- Signal Capture Interface
- Metadata Extractor
- Format Normalizer
- Provenance Tracker

Output: A structured signal object, ready for multi-layered analysis.

5.2 Analytical Processing and Scoring

Once normalized, signals are routed to analytical modules, each responsible for evaluating a specific dimension of trust. These modules are powered by a combination of machine learning models, statistical engines, and rule-based systems. *AI tools—including large language models, graph-based classifiers, and probabilistic inference engines—support key tasks* such as semantic analysis, stance detection, source profiling, and predictive scoring.

Primary analytical functions include:

- **Entropy and Narrative Novelty Analysis** – Detects redundancy and templated messaging.
- **Claim Extraction and Verification** – Identifies factual assertions and cross-references them against reliable knowledge sources.
- **Source Behavior Profiling** – Tracks source defection patterns, inconsistencies, and semantic drift.
- **Amplification Network Mapping** – Analyzes how content spreads across networks and identifies coordinated or inorganic amplification.
- **Stance and Sentiment Detection** – Measures framing, bias, and emotional charge.
- **Signal-to-Noise Evaluation** – Assesses contextual clarity and relevance relative to surrounding discourse.

Each module outputs a set of trust-relevant indicators. These are fused into a composite trust score through a weighted aggregation model, with configurable parameters based on context, source type, or platform requirements.

5.3 Oversight, Feedback, and Governance

Beyond scoring, the system includes governance modules that assess the credibility of trust evaluations over time and under adversarial pressure. These modules provide oversight, initiate corrective processes, and reinforce transparency.

Key governance functions include:

- **Fact-Resilience Estimator (FRE)** – Evaluates whether claims maintain stable trust scores over time and under challenge.
- **Claim Volatility Monitor** – Flags assertions with rapidly shifting credibility signals.
- **Score Governance Engine** – Oversees weighting strategies, applies policy rules, and initiates alerts for anomalous patterns.
- **Audit and Accountability Layer** – Records decision traces and supports forensic review or audit replay.

Governance components operate continuously and can trigger system-level interventions such as model retraining, source downgrades, or score recalibration.

5.4 Output Interfaces and Applications

Final trust scores, analysis metadata, and governance signals are made available through structured interfaces designed to serve multiple stakeholders. These include:

- **Public-facing indicators** such as trust labels, disclaimers, or confidence overlays
- **Moderation outputs** including flags, visibility scores, or escalation tags
- **Editorial and researcher tools** that surface source-level histories, amplification paths, and claim lineage
- **AI-Augmented Dashboards** – Visualizations and summary insights generated by analytical models, including claim clustering, narrative evolution tracking, and volatility forecasts
- **Query and Exploration Tools** – Interfaces for investigators to search claim histories, trace source behaviors, or filter signals by topic, score range, or volatility
- **Audit Log Access** – Secure access to machine-readable decision traces for oversight, compliance, and forensics. Access privileges and data granularity are adjustable based on user role, organizational policy, and use case. These facilities ensure that human actors remain empowered to interpret, question, and respond to the system's outputs.

5.5 Feedback Loops and Adaptive Learning

Signals with ambiguous outcomes or manual overrides are fed into a feedback layer for reanalysis and model tuning. **AI models used in earlier stages are retrained or adjusted using new evidence, user feedback, and system audit trails**. This adaptive

loop ensures the models remain responsive to changes in discourse patterns, adversarial behavior, and evolving knowledge domains.

5.6 System Properties

- **Parallelism** – Components operate concurrently, allowing scalable throughput
- **Interoperability** – Each module is API-accessible and replaceable, enabling open integration
- **Governance-Awareness** – Every decision is traceable and subject to policy constraints and human audit
- **AI-Modular Integration** – Each major subsystem can incorporate or replace AI models without disrupting overall system integrity. Model execution is governed by transparency and explainability constraints, with fallback mechanisms and human review where necessary

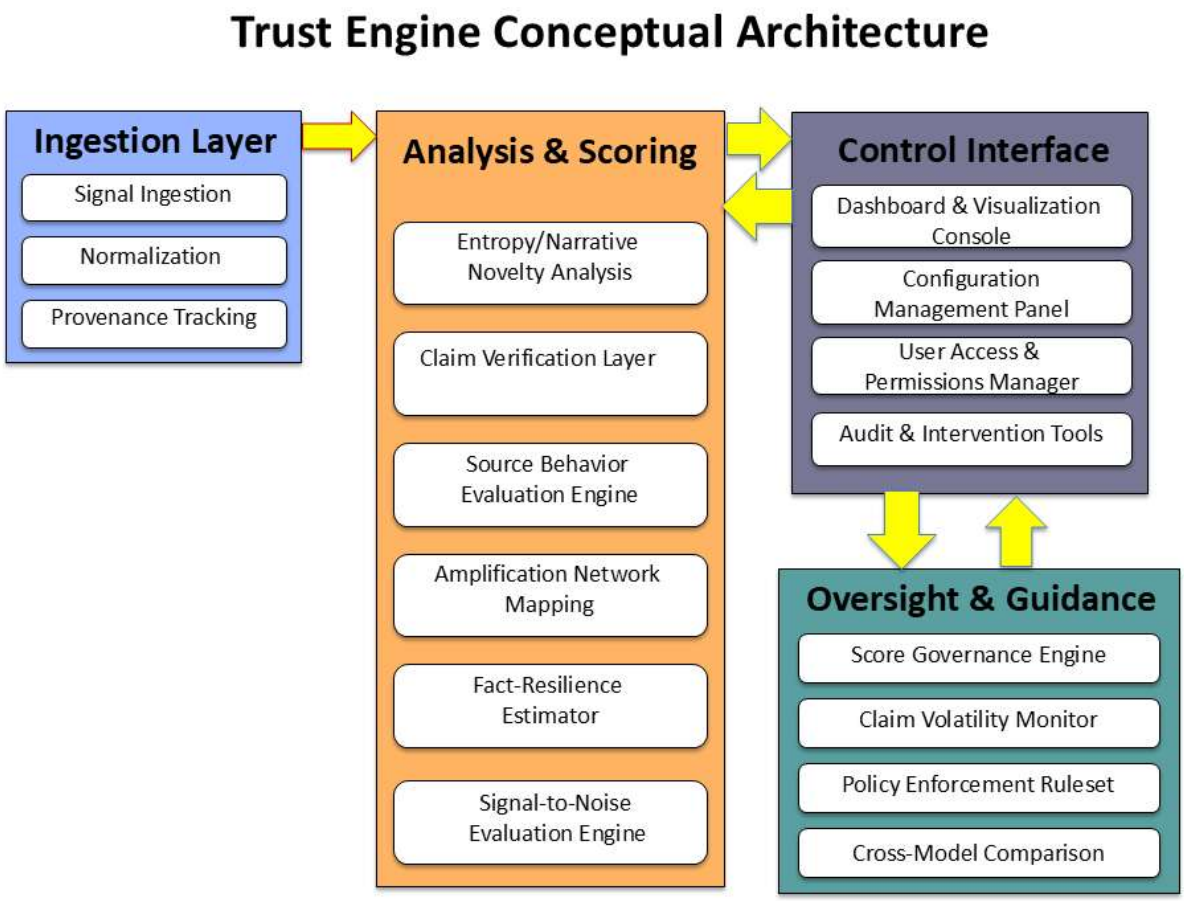

Figure 1. Trust Engine Conceptual Architecture.
This diagram illustrates the modular subsystems and data flows described in Sections 5.1 through 5.6.

6. Trust Engine Use Cases Beyond Propaganda Detection

Public Health Messaging

The trust engine can prioritize reliable sources in areas such as vaccine guidance, medical treatment updates, and epidemiological data. By elevating verified health information and suppressing misinformation, it can support informed decision-making during pandemics or localized outbreaks.

Crisis Communication

In contexts such as natural disasters, cyberattacks, or armed conflicts, timely and accurate information can be life-saving. Trust scoring systems can help emergency responders, journalists, and platform moderators surface credible content and suppress harmful rumors in real time.

Electoral Integrity and Civic Information

The system can assist in identifying and flagging disinformation related to polling station closures, ballot manipulation claims, or voter intimidation narratives. It can also elevate verified information from election officials, watchdog groups, and credible local media.

Platform-Based Civic Moderation

Social media platforms and civic technology projects can integrate components of the trust engine into their moderation workflows. This allows for more transparent, auditable, and context-aware decisions about content visibility, account behavior, and network-level amplification.

7. Strategic Role of Public-Interest Organizations

This paper proposes trust engineering as a public good—and technology-oriented public interest organizations are natural partners for operationalizing its core principles. Collaboration could include joint prototyping, governance research, and convening of ecosystem stakeholders to establish trust engineering as a foundational pillar of the digital information landscape.

The modular character of the proposed architecture enables distributed participation across multiple institutions. Public-interest organizations committed to open-source development, digital rights, and trustworthy AI—such as Mozilla—are especially well positioned to help pioneer this emerging field. By supporting the development of a prototype Propaganda Detection Engine (PDE), these institutions could catalyze the creation of open standards, transparent governance models, and scalable infrastructure to improve the integrity of public discourse worldwide

8. Next Steps for Trust Engineering Development

The development of a functional trust engine will proceed in three coordinated phases, each designed to balance technical advancement with ethical oversight, institutional alignment, and real-world testing.

Phase 1: Foundational Research and Prototyping

The initial phase will focus on rigorous research to refine the conceptual framework and build functional prototypes of core modules such as provenance tracking, signal ingestion, and narrative mapping. Partnerships with universities, civic tech labs, and NGOs will be essential to assemble diverse datasets, test environments, and ethics panels. Early-stage user testing with journalists, fact-checkers, and community moderators will validate usability and define key performance indicators (KPIs).

Phase 2: System Integration and Piloting

This stage will integrate all core components of the trust engine into a unified platform, focusing on interoperability and modularity. The system will be deployed in real-world pilot environments such as media monitoring hubs, local elections, or information integrity collaborations with platforms. Real-time feedback loops will help refine classifiers and trust metrics, while governance teams assess transparency, false positive risk, and bias mitigation.

Phase 3: Policy Alignment and Scalable Deployment

The final stage will involve formalizing open standards and submitting frameworks to policymakers, multilateral organizations, and industry partners for endorsement. Scalable deployment models will be explored, such as APIs for integration into social platforms, browser-based extensions, or citizen alert systems. Long-term success will depend on sustainable funding, institutional buy-in, and participatory oversight models.

9. Ethical and Governance Considerations

A trust engine designed to evaluate digital content must operate under robust ethical safeguards. In high-stakes domains such as elections, health, or crisis response, credibility decisions have real-world consequences. To earn legitimacy and sustain public trust, four key governance principles must guide system design and oversight:

Algorithmic Accountability
No model should operate as a black box in public-interest systems. The trust engine will adopt transparent model architectures, document training data provenance where feasible,

and expose scoring logic for audit. Independent algorithm auditors must be empowered to test, critique, and recommend improvements based on observed outcomes.

Privacy and Surveillance Risks
Signal ingestion and network mapping can raise legitimate civil liberties concerns. The system will employ strict privacy safeguards, including data minimization, anonymization, and local on-device processing when possible. All data usage will be subject to ethical review by multidisciplinary oversight bodies with public representation.

Resilience to Adversarial Manipulation
Any system that assesses trust becomes a potential target. The trust engine will include countermeasures against adversarial attacks such as reputation gaming, classifier poisoning, or synthetic media injection. These defenses will include red-teaming, anomaly detection, and fallback protocols to protect reliability and integrity.

Democratic Legitimacy and Inclusive Design
Technical excellence alone cannot secure legitimacy. Affected communities must help shape trust mechanisms. Participatory governance structures—such as citizen juries, multilingual feedback channels, and transparent escalation processes—will ensure the system reflects pluralistic values and evolving societal norms.

These principles will guide both the initial architecture and ongoing governance of the trust engine, ensuring that technical sophistication is balanced by civic responsibility and public accountability.

10. Illustrative Use Cases

The following scenarios illustrate how the trust engine can operate across different domains, providing actionable insights and supporting timely interventions.

Case 1: De-escalating Conflict Through Rapid Verification

Following a major accident or violent crime, political actors attempt to frame the event as a terrorist attack to justify military escalation. Manipulative narratives circulate widely on social media, inflaming public opinion. The trust engine detects coordinated amplification patterns and cross-verifies claims using its provenance and fact-checking layers. The early identification of false narratives enables journalists and civil society groups to disseminate corrections, reducing the momentum toward escalation.

Case 2: Detecting a Coordinated Voter Suppression Campaign

During a regional election, the system identifies a sudden surge in near-identical posts across multiple platforms falsely claiming polling station closures. Narrative mapping and amplification analysis trace the campaign to a small cluster of newly created

accounts. A real-time alert is issued to electoral authorities and watchdog organizations, who respond with verified information and platform takedown requests.

Case 3: Managing Disinformation in a Humanitarian Crisis

In the aftermath of a natural disaster, rumors spread online suggesting that distributed aid is contaminated. The trust engine uses its narrative mapping and fact-resilience estimation to identify the scope and credibility of the false claims. Humanitarian organizations use this insight to issue timely counter-messaging, supported by transparent trust indicators, helping restore confidence in relief efforts.

Case 4: Enhancing Civic Moderation on Social Platforms

A social media platform's civic integrity team integrates the trust engine's amplification and source profiling tools. The system detects a cross-border network artificially boosting xenophobic content through coordinated posting. Trust metrics and propagation data are used to guide moderation decisions and provide transparency in public reports about inauthentic behavior.

11. Feasibility and Scalability of Deployment

The development and deployment of a functional trust engine are technically feasible using existing AI infrastructure, open-source tooling, and cloud-native architectures. Real-world media monitoring systems such as GDELT and Media Cloud already demonstrate the capacity to ingest and process millions of content items daily across diverse platforms and languages.

Natural language processing (NLP) classifiers and multimodal content analysis models can now process thousands of text documents and hundreds of media files per second using commercially available GPU infrastructure. Core tasks—including narrative clustering, entropy scoring, sentiment analysis, and probabilistic claim verification—are already being performed at scale by commercial threat intelligence and content moderation systems.

The proposed trust engine leverages these capabilities by combining modular scoring, transformer-based models, and scalable ingestion pipelines. This architecture supports horizontal scaling and enables deployment in targeted domains such as election monitoring, health information verification, or crisis response.

To manage cost during the pilot phase, the system will employ strategic sampling techniques, including time-based sampling, entropy-weighted content selection, and source-tiered prioritization. These methods ensure representative evaluation without requiring full-scale ingestion of all available content. Pilot deployments can focus on high-risk narratives, known disinformation vectors, or specific geographic regions.

Estimated monthly infrastructure costs—including ingestion, model inference, trust scoring, logging, and auditing—range from $2,000 to $3,000 USD. These figures are

consistent with comparable open-source and civic tech initiatives, and well within the scope of mission-aligned funders such as the Mozilla Foundation. Additional cost optimization may be achieved through model compression, batched inference, and use of shared compute partnerships.

Taken together, these factors demonstrate that the trust engine is not merely a conceptual framework, but a practical, scalable, and ethically grounded system ready for prototyping and iterative deployment.

12. Summary and Conclusion

The assurance of trust in public information is no longer an unsolvable philosophical dilemma — it is an engineering problem, tractable with today's AI and information technologies. This paper has proposed a modular, auditable, and scalable Trust Engine capable of identifying and thus counteracting propaganda in real-time.

Just as frequency modulation reduced noise in radio transmission, and information theory made digital communication reliable, trust engineering offers a path to minimize manipulative distortion in our shared digital environment. Propaganda, too, can be filtered and flagged—not by censoring speech, but by clarifying signal.

Trust engineering is not merely a technical vision—it is a democratic necessity. With principled design, multidisciplinary governance, and open collaboration, we can build systems that do not merely transmit information but earn the trust of those they serve.

References

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Leetaru, K., & Schrodt, P. A. (2013). *GDELT: Global Data on Events, Location, and Tone*.

Marwick, A. E., & Lewis, R. (2017). *Media Manipulation and Disinformation Online*. Data & Society.

Media Cloud Project, Berkman Klein Center (ongoing).

Helberger, N., Pierson, J., & Poell, T. (2018). *Governing online platforms: From contested to cooperative responsibility*. The Information Society.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Proceedings of the 22nd ACM SIGKDD.

Shannon, C. E. (1948). *A Mathematical Theory of Communication*.

Starbird, K. (2017). *Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter*. ICWSM.

Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework*. Council of Europe report

Statement of Originality

This white paper is an original work created by the author, Haig Hovaness. All core ideas, frameworks, and conceptual contributions—specifically the articulation of "trust engineering" as a formal discipline and the proposal of measurable, audit-capable trust metrics—originate from the author. While artificial intelligence tools (e.g., language models) were employed to assist with drafting and clarity, the synthesis, structure, and intellectual direction of the document are solely attributable to the author. No content has been copied from existing literature or previously published works.

Haig Hovaness
Independent Researcher
hovaness@pipeline.com
Pelham, NY

May 28, 2025