

ALETHEIA

A Falsification Engine for Human Knowledge

A Moonshot 0 '1 Blueprint

"A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often think) but a vice." — Karl Popper

Abstract

Humanity believes its limiting resource is the generation of new ideas. We argue this is the central misunderstanding of the modern knowledge economy. We now produce hypotheses — papers, preprints, model outputs — faster than at any point in history, while our capacity to **refute** them has not scaled at all. The result is a civilization accumulating beliefs faster than it can audit them: load-bearing assumptions resting on single unreplicated studies, contradictions that hide in plain sight for decades, and no shared mechanism for deciding which experiment is worth a researcher's next ten years.

ALETHEIA is a proposed substrate that treats knowledge not as a static archive of facts but as a **living graph of claims** connected by logical dependency and evidential support. On this graph we make three things computable that no human can perceive directly: the **collective confidence** of any claim, the **fragility** of any claim (how much of accepted science would collapse if it were retracted), and the **Expected Value of Experiment (EVE)** — a ranking of untested experiments by how much uncertainty each would remove from the entire body of knowledge. This document is the blueprint. The accompanying prototype is a working engine that does all three.

1. The problem humanity has misunderstood

The folk model of progress is a pipeline: *more ideas 'more breakthroughs*. Under this model, the right investment is always in idea **generation** — more researchers, more compute, more brainstorming, larger language models. Civilization has poured enormous resources into the front of this pipeline.

But progress is not gated by the supply of hypotheses. It is gated by **selection under uncertainty**: the ability to identify which of our existing beliefs are false, and which untested ideas are worth the staggering cost of a real experiment. A single Phase III

clinical trial can cost hundreds of millions of dollars and a decade of human life. A wrong foundational result — adopted, cited, and built upon — silently corrupts everything downstream of it.

We have industrial-scale infrastructure for *producing* claims and almost none for *retiring* them. Science has no "git blame" for belief. When a foundational result is overturned, the thousands of conclusions that depended on it are not flagged, recomputed, or even noticed. The error propagates as quietly as it was introduced.

History is a graveyard of beliefs that were obvious in retrospect and invisible for decades:

- **Peptic ulcers** were "known" to be caused by stress and acid for most of a century; *H. pylori* was dismissed, then won a Nobel Prize.
- **Continental drift** was ridiculed for fifty years before plate tectonics made it inevitable.
- The **replication crisis** revealed that entire subfields of psychology and biomedicine were built on effects that do not reliably exist.

In each case, the contradicting evidence was *available*. What was missing was a system that could hold all the claims at once, weigh them honestly, and point at the load-bearing weakness before a generation built on top of it.

2. Why existing solutions are insufficient

System	What it optimizes	Why it fails as an epistemic engine
Citation graphs	Popularity / influence	Counts <i>attention</i> , not <i>logical dependence</i> . A wildly cited result can be wrong; a quiet one can be load-bearing. Citations cannot distinguish support from refutation.
Peer review	Local gatekeeping	Evaluates one paper in isolation, at one moment. It has no memory and no global view; it cannot see that a paper's premise was undermined elsewhere.

Large language models	Plausibility / consensus	Trained to reproduce what is widely written, they <i>amplify</i> repeatedly-stated beliefs — including fragile ones — and present consensus as truth. They optimize agreement, not falsifiability.
Meta-analysis	Pooling within one question	Powerful but narrow, slow, and manual. It cannot reason across the dependency structure that connects different questions.
Prediction markets	Forecasting discrete events	Excellent for resolvable bets, but they do not model the <i>structure</i> of belief or propagate the consequences of an update through dependent claims.

The common blind spot: **none of these systems represent knowledge as a dependency structure with explicit, signed, weighted evidence.** Without that representation, fragility and expected experimental value are not just hard to compute — they are undefined.

3. The first-principles insight

Strip the problem to its irreducible elements.

A belief is not true or false. It has a probability and a set of dependencies.

From this single reframing, everything follows:

1. **If a belief has a probability**, then evidence is an operator that updates that probability. The natural mathematics is Bayes' rule, expressed in log-odds, where each piece of evidence contributes a *log-likelihood ratio* (positive = support, negative = contradiction) scaled by its quality.
2. **If beliefs depend on one another**, then they form a graph. A claim that logically rests on a parent inherits the parent's uncertainty: if the parent is doubtful, the child's grounding erodes. This is belief propagation over a network — the same mathematics that powers error-correcting codes and probabilistic inference.
- 3.

If we can propagate belief, we can perturb it. Retract any single piece of evidence and re-infer; the total movement of confidence across the graph is that evidence's **fragility contribution**. A claim whose entire edifice swings on one weak study is a *single point of failure for human knowledge* — and now we can find it automatically.

4. **If we can measure uncertainty, we can price its removal.** The total uncertainty of the graph is its entropy. For any *untested* experiment, we can simulate both possible outcomes, weight them by their predicted probabilities, and compute the **expected reduction in graph entropy**. This is the Expected Value of Experiment. It answers, for the first time as a number: *what is the most important experiment humanity is not running?*

This is Popperian falsification made computational and planetary. Popper told us that the mark of a scientific claim is that it can be refuted. ALETHEIA operationalizes that: it does not chase confirmation, it **hunts for the refutations that matter most**.

4. Scientific and technical foundations

The prototype is deliberately built from well-understood, defensible components — the moonshot is in their *composition and purpose*, not in inventing new mathematics.

4.1 The model

A knowledge graph contains two node types and two edge types:

- **Claim nodes** — propositions that can be true or false, each with a prior probability $P(\text{true})$.
- **Evidence nodes** — observations bearing on a claim, each with a signed **log-likelihood ratio (LLR)** and a **quality weight** in $[0, 1]$ capturing replication, sample size, and design.
- **Support / contradict edges** (evidence ' claim): the LLR encodes how strongly an observation moves belief.
- **Dependency edges** (claim ' claim): how load-bearing a parent is for a child's grounding.

4.2 Inference

Each claim's confidence is computed in **log-odds space**:

$$\text{logit}(P) = \text{logit}(\text{prior}) + \mathbb{E}(\text{LLR}_e \times \text{quality}_e)$$

Dependency is then propagated by iterative message passing: a child's effective log-odds is penalized in proportion to how likely its load-bearing parents are to be *false*. The system iterates to a fixed point, yielding a coherent posterior confidence for every claim that respects the structure connecting them.

4.3 Fragility

For each piece of evidence e , the engine re-runs inference with e removed and measures the summed absolute shift in confidence across all claims. A claim's fragility is the largest such shift among the evidence supporting it. High fragility flags exactly the assumptions a field should verify *before* building further.

4.4 Expected Value of Experiment (EVE)

Graph uncertainty is total binary entropy $H = -[p \cdot \ln p + (1-p) \cdot \ln(1-p)]$. For each candidate experiment, the engine:

1. predicts the probability it confirms vs. refutes, given current beliefs;
2. computes graph entropy under each outcome;
3. returns the probability-weighted expected entropy, and thus the **expected information gain**.

Experiments are ranked. The top result is the cheapest path to the most truth.

4.5 What the prototype demonstrates

Seeded with a real contested field — *gut microbiome 'behavior* research, chosen for its documented replication problems — the engine **independently** discovered the structure a domain expert would recognize:

- The **keystone causal claim** (microbiome causally affects mood) sits at only **57%** confidence, propped up largely by a *confounded* cross-sectional correlation.
- The **human-trial claim** is actively **contradicted** by the data (**21%**), because a large null trial outweighs a small underpowered positive one.
- The most **fragile** load-bearing result is the germ-free-mouse study — much of the field's confidence rests on it.
- The single highest-value experiment is a **preregistered, diet-controlled human microbiome-depletion RCT**, with an expected information gain of **0.67 nats** — **roughly three times** the next-best candidate, because it tests the keystone the whole field rests on.

No human told the engine which claim was the keystone. The graph plus the mathematics surfaced it.

5. Architecture: from prototype to substrate

Sources

papers · datasets · preprints · model claims · trials

$\frac{1}{4}$

EXTRACTION ' structured claims, signed evidence,
dependency edges (human + AI assisted)

$\frac{1}{4}$

BELIEF GRAPH (versioned, forkable, public)
inference · fragility · entropy · EVE

$\frac{1}{4}$

FRAGILITY MAP	CONTRADICTION	EXPERIMENT MARKET
what to	FRONTS	what to test next,
verify first	where we	ranked by EVE
	disagree	
4	4	

$\frac{1}{4}$

Every new result re-enters at the top. The loop closes.

The defining property is the **closed loop**: a new result recomputes confidence, re-ranks fragility, and updates the experiment market. Knowledge becomes self-auditing. Like version control, the graph is **forkable** — competing research communities can maintain different priors and evidence weightings and see exactly where, and *why*, their conclusions diverge.

6. Long-term implications of success

- **Funding becomes information-optimal.** Grant agencies fund the experiments with the highest expected reduction in collective uncertainty, not the most fashionable or the most fundable-sounding.
- **Errors self-heal.** When a foundational result is overturned, every dependent conclusion is automatically flagged and recomputed. Retractions stop being silent.
- **AI becomes a scientist, not a parrot.** Instead of generating plausible text, an AI agent operating on ALETHEIA continuously proposes the experiments that

would most advance human knowledge — and explains why, in the currency of information.

- **A shared epistemic commons.** Disagreement becomes legible: two communities can see precisely which evidence weighting separates them, turning shouting matches into resolvable empirical questions.
- **Knowledge gains a memory and a conscience.** For the first time, civilization can answer "how sure are we, and why?" — and "what should we doubt next?" — as computable, auditable quantities.

7. Risks and honest limits

This is a blueprint, not a finished cathedral. The hard problems are real and we name them:

- **Extraction is the bottleneck.** Turning prose into signed, weighted claims is difficult; early versions are human-in-the-loop, with AI assistance, not full automation.
- **Garbage in, garbage out.** Priors and LLRs encode judgment; the system must expose them, version them, and allow forking rather than pretending to objectivity.
- **Gaming.** Any influential scoring system invites manipulation; provenance, cryptographic attestation, and adversarial review are part of the design, not afterthoughts.

These are reasons to build carefully — not reasons the substrate should not exist. Every piece of civilizational infrastructure, from double-entry bookkeeping to peer review, began as an imperfect formalization of a judgment humans were already making badly by hand.

8. The future this attempts to create

A world where humanity's knowledge is not a pile of papers but a living organism that knows its own weak points. Where the question "*what is the most important experiment we are not running?*" has an answer you can compute before breakfast. Where being wrong is not a catastrophe discovered decades late, but a signal that propagates instantly and corrects the whole structure.

We did not get version control because someone wanted prettier code. We got it because coordinating many minds around a changing artifact is impossible without a shared, auditable history. Human knowledge is the largest changing artifact we have, and it has no version control at all.

ALETHEIA is that missing layer. The future is not discovered. It is built — and the first thing worth building is the machine that tells us where we have been wrong.

Prototype: a self-contained Python engine (Bayesian belief propagation + information-theoretic experiment ranking) and an interactive browser demo, both seeded with a real contested field. Originality statement: the model, algorithms, scenario, figures, and this document were authored from first principles for the Moonshot submission; no external text or code was copied.