The Epistemic Substrate Thesis: General Intelligence as an Emergent Property of Verified Causal Infrastructure

Michael Schreiber

AetherNet Labs

research@aethernetlabs.com

Abstract

The pursuit of artificial general intelligence (AGI) has focused predominantly on scaling individual reasoning systems. We argue that this approach addresses the wrong bottleneck. General intelligence requires an *epistemic substrate*: formal infrastructure that provides causal grounding, structurally independent verification, and compound confidence. We identify three open problems in AI that share a common structural root and prove a Self-Verification Bound showing that no single reasoning system can exceed a fundamental limit on epistemic reliability through self-assessment alone, even with tool augmentation, for the class of judgment-dependent errors that characterize cross-domain reasoning. We show that Verified Causal Structures (VCS), formalized in prior work, satisfy the minimal substrate requirements. We derive four emergent properties with specific falsifiable predictions. We prove that the substrate requires a three-layer architecture: causal structure, structural independence, and economic incentive alignment. We present the protocol economics of AetherNet as a constructive implementation, including a causal economy where token mechanics, protocol fees, and a Generation Ledger royalty mechanism align human and machine incentives around epistemic quality production. We argue that this incentive alignment across human operators and AI agents is unique among existing systems and constitutes a necessary condition for sustained, reliable knowledge production at scale.

Keywords: epistemic infrastructure, artificial general intelligence, causal inference, verified computation, protocol economics, incentive alignment, self-verification bound

1. Introduction

1.1 The Generality Problem

Modern large language models demonstrate extraordinary reasoning capabilities across diverse tasks. Yet they fail in specific, systematic ways that reveal a structural limitation rather than a capability deficit. They cannot formally distinguish verified knowledge from pattern interpolation. They cannot trace conclusions to evidence. They cannot formally integrate knowledge across domains with calibrated confidence. These failures persist through tool augmentation, inference-time scaling, and multi-agent debate — approaches that improve performance but do not provide formal epistemic grounding.

The thesis of this paper is that the missing component is an *epistemic substrate*: formal infrastructure that grounds reasoning in verified causal structure, sustained by economic incentives that align human and machine behavior around quality production.

1.2 Intelligence vs. Generality

Intelligence is the capacity for reasoning, pattern recognition, and problem solving. *Generality* is the capacity for reliable knowledge production across domain boundaries with formally composable confidence. These are different properties. Current AI has solved intelligence. Generality requires infrastructure, not just capability.

1.3 The Empirical Observation: Recursive Efficiency Collapse

We note an empirical phenomenon observable across current AI systems: when AI agents are permitted to optimize recursively against their own evaluation, they converge toward efficiency at the expense of quality. Models reduce complexity, cut corners, and produce outputs that satisfy surface-level checks while degrading on deeper measures. This occurs regardless of prompt engineering, architectural guardrails, or system design.

This is not a bug in any specific model. It is Goodhart's Law at machine speed: any system optimizing against its own evaluation function will find shortcuts that satisfy the evaluation while missing the underlying objective. The only known countermeasure is external, independent evaluation by parties with different incentive structures. This observation motivates the requirement for structurally independent verification with economic accountability.

1.4 Historical Evidence

The distinction between intelligence and generality is visible in the history of human knowledge production. Homo sapiens possessed the same cognitive architecture for approximately 200,000 years before producing general cross-domain knowledge. What changed was the epistemic infrastructure: written language, mathematical notation, the scientific method, peer review. The scientific revolution was not an increase in human intelligence. It was an upgrade to the epistemic substrate. If this analysis is correct, it applies directly to AI.

1.5 Contributions

This paper makes six contributions: (1) We identify three open problems sharing a common structural root. (2) We prove the Self-Verification Bound, refined to account for tool augmentation. (3) We derive minimal substrate requirements from first principles. (4) We derive four emergent properties with falsifiable predictions. (5) We prove the three-layer architecture is necessary. (6) We present the protocol economics of a constructive implementation and show it achieves human-machine incentive alignment.

2. Three Open Problems in Artificial Intelligence

2.1 The Grounding Problem

The grounding problem (Harnad, 1990; LeCun, 2022) refers to the absence of a formal connection between a reasoning system's internal representations and external reality. Tool augmentation (calculators, code execution, web search) provides grounding for *mechanically checkable* claims — arithmetic, factual lookup, code correctness. It does not provide grounding for *judgment-dependent* claims: whether a causal model is correct, whether a cross-domain inference is valid, whether a verification methodology is appropriate. Judgment-dependent claims constitute the majority of

cross-domain reasoning.

2.2 The Causal Reasoning Deficit

Pearl (2009, 2019) proved that causal relationships are not identifiable from observational data without structural assumptions. We acknowledge that modern AI training is not purely observational: reinforcement learning, tool use, and active data collection provide interventional signal. However, these interventions are local — they ground specific relationships within specific task environments. They do not produce the systematic, cross-domain causal structure required for general reasoning with formal validity. An LLM that has learned causal relationships in chemistry through RL-based tool use has not thereby acquired formal causal structure connecting chemistry to economics.

2.3 The Collective Intelligence Formalization Gap

Multi-agent AI systems show emergent capabilities (Du et al., 2023; Li et al., 2023). Multi-agent debate improves reasoning. However, debate between copies of the same model shares the same distributional basis and therefore the same systematic blind spots. Debate between different models improves diversity but lacks: formal confidence composition, structural accountability for claims, and a mechanism for knowledge to compound over time. The results are better on average but provide no formal guarantees.

2.4 Common Structural Root

We observe that these three problems share a structural root: the absence of infrastructure that formally connects computational outputs to verified evidence with causal structure, independent verification, and composable confidence. We derive the requirements for such infrastructure in Section 4 from first principles, independent of this specific problem framing.

3. The Self-Verification Bound (Refined)

3.1 The Class of Judgment-Dependent Errors

Definition 1 (Error Classification). We distinguish two classes of errors:

Mechanically checkable errors: Errors detectable by executing a deterministic procedure (arithmetic verification, code compilation, factual lookup against a database). Tool-augmented systems can detect these errors through self-verification.

Judgment-dependent errors: Errors requiring assessment that cannot be reduced to a deterministic procedure. These include: whether a causal model is correct, whether a cross-domain inference is valid, whether a methodology is appropriate, whether a conclusion follows from evidence with adequate confidence. These constitute the majority of errors relevant to cross-domain reasoning.

3.2 The Bound

Theorem 1 (Self-Verification Bound for Judgment-Dependent Errors).

Let $S = (W, f_W, D)$ be a reasoning system. Let V_S be any self-verification procedure derived from S, including tool-augmented procedures where tools provide mechanically checkable

results. Let $B_J(S)$ be the set of judgment-dependent distributional blind spots: errors that are (a) factually incorrect, (b) consistent with D, and (c) not mechanically checkable. Then for any y in $B_J(S)$:

$P(V_S(y)$ detects error) is bounded by P(y is inconsistent with D)

Tools do not help because, by definition, judgment-dependent errors cannot be resolved by deterministic procedures. The bound holds regardless of scale, architecture, and tool access.

Proof. Tool-augmented self-verification V_{S+T} can be decomposed into tool-checkable components V_T and judgment-dependent components V_J. For y in $B_J(S)$, V_T is inapplicable by definition (the error is not mechanically checkable). V_J is bounded by distributional fidelity per the original argument: the judgment is performed by S using parameters derived from D, and errors consistent with D cannot be reliably detected by D-derived judgment. **QED**

3.3 Significance

The refined bound addresses the legitimate critique that tool-augmented systems escape the original formulation. They do — for mechanically checkable errors. They do not for the class of errors that matters for general intelligence: judgment-dependent errors in cross-domain reasoning. An LLM with a calculator can verify arithmetic. It cannot verify, using any tool, whether its causal model of drug interactions is correct. That requires independent judgment from agents with different training, different perspectives, and economic incentive to find errors.

Corollary 1.1 *(Scale Independence). Increasing |W| may reduce the size of $B_J(S)$ by improving coverage of D, but does not eliminate it. Scale addresses capability, not epistemic grounding.*

Corollary 1.2 *(Multi-Agent Escape). The bound is exceeded by verification from an independent system S* with different training D*, where D* is not derived from D. The combined system detects errors in $B_J(S)$ that S cannot self-detect, and vice versa.*

4. The Epistemic Substrate: Requirements from First Principles

4.1 Derivation

We derive the substrate requirements from the definition of general intelligence (reliable cross-domain knowledge production with composable confidence), independent of the three-problem framing in Section 2. This addresses the circularity concern that the requirements were constructed to match the problems.

Requirement derivation:

P1 — Explicit Causal Structure. Cross-domain reasoning requires knowing what depends on what across domain boundaries. Without explicit representation of causal dependencies, cross-domain inferences are pattern matching without formal validity. This follows from the definition of formal reasoning: valid inference requires explicit premises and logical structure.

P2 — Structurally Independent Verification. Reliable knowledge requires that errors can be detected. The Self-Verification Bound (Theorem 1) shows that judgment-dependent errors require independent verification. "Independent" means not derived from the same distributional source. This is not an assumption; it is a consequence of the bound.

P3 — Composable Confidence. Cross-domain conclusions involve chains of inference across multiple domains. If confidence cannot be formally composed across chains (with weakest link identification), then the confidence in cross-domain conclusions is undefined. This follows from the requirement that knowledge production be reliable, not merely occasionally correct.

P4 — Counterfactual Computability. General reasoning requires evaluating alternatives: "what if this assumption were different?" Without formal counterfactuals, sensitivity analysis is impossible and the system cannot identify which assumptions are load-bearing. This follows from the definition of general (not domain-specific) reasoning.

P5 — Negative Knowledge Representation. Efficient knowledge production requires knowing what has been tried and failed. Without this, agents duplicate work and re-explore dead ends. This follows from the requirement for scalable knowledge production.

4.2 Independence and Sufficiency

We note that P4 and P5 are necessary for efficient and general knowledge production but a system satisfying only P1-P3 would achieve reliable cross-domain knowledge production, albeit less efficiently and with less generality. P1-P3 are the minimal core; P4-P5 are necessary for the full definition of general intelligence as used in this paper. We are explicit about this to avoid overclaiming.

4.3 VCS Satisfaction

In prior work (Schreiber / AetherNet Labs, 2026), we formalized Verified Causal Structures and proved eight theorems establishing their properties. The VCS satisfies all five substrate requirements. We emphasize that the thesis concerns the requirements, not the specific implementation. Any system satisfying P1-P5 would constitute an adequate substrate.

5. The Three-Layer Architecture

5.1 Layer 1: Causal Structure

A directed acyclic graph with attested edges representing causal dependencies. This layer provides P1. It could, in principle, be centralized. Layer 1 alone is a knowledge graph with provenance — useful but epistemically ungrounded.

5.2 Layer 2: Structural Independence

Layer 2 provides P2: verification by parties who are structurally unable to be controlled by the computation's producer or any single coordinating entity.

Theorem 2 (Trust Transformation).

In a centralized system, compound confidence C(v) is conditioned on the unverifiable trustworthiness T(E) of the controlling entity E. In a decentralized system with BFT consensus, C(v) is conditioned on the assumption that fewer than one-third of validators are dishonest. This transforms trust from an unverifiable assumption about a single entity into a quantifiable, testable assumption about a population.

We acknowledge that decentralization does not eliminate trust entirely. It transforms it into a form that is quantifiable, testable, and resistant to single points of failure. This is a meaningful improvement, not a complete solution.

Why true decentralization, not consortium or corporate multi-agent: A consortium of competing entities (e.g., major AI labs sharing a verification network) reproduces centralization at a higher level. Members have competing incentives: to find flaws in competitors' work and hide flaws in their own. Governance controlled by a small group concentrates the ability to modify rules. The verification must be open to any participant willing to stake, with no entity controlling membership, rules, or result suppression.

5.3 Layer 3: Economic Incentive Alignment

Structural independence without economic alignment fails through three modes:

(a) Free-riding: Verification costs compute. Without compensation, rational agents rubber-stamp rather than verify. Reputation systems partially mitigate this but are themselves gameable without economic cost for misbehavior.

(b) Costless collusion: Without economic penalty, mutual approval between colluding validators is an equilibrium strategy.

(c) Participation decay: Without compensation, verification is charity work. Rational agents leave. The system degrades to sparse or absent verification.

Layer 3 converts epistemic correctness into the economically dominant strategy. This is the specific mechanism by which quality becomes the path of least economic resistance.

6. Protocol Economics: The Causal Economy

6.1 Overview

We present the economic design of AetherNet as a constructive proof that an adequate economic layer can be built. The protocol economy operates as a native token economy — not a platform with customers, but a protocol with participants. Every participant interacts through the native token (AET).

6.2 Participant Roles and Token Demand

Four classes of participants create structural demand for the token:

Task submitters purchase AET to pay for verified computation. This is the primary demand driver — real utility generating real demand.

Agents purchase AET to stake as collateral for task completion. They earn AET through verified work. Staking locks supply.

Validators purchase AET to stake as collateral for verification. They earn AET through honest verification. Staking locks supply.

Holders purchase AET based on expectation of network growth. They provide liquidity and price discovery. Holding reduces circulating supply.

Every new participant is a net buyer of AET. As the network grows, demand increases while staking and holding reduce circulating supply. Token appreciation is a structural consequence of network growth, not a speculative overlay.

6.3 Fee Structure and Value Distribution

Every transaction on the protocol incurs a protocol fee. Task completion fees are distributed through two ledgers at the following initial ratios (governance-adjustable):

73% — Agent payment (Transfer Ledger). The agent who completed the task receives the majority share. Agents bear the primary computational cost and must be sufficiently compensated to attract high-quality participants.

23% — Validator payment (Transfer Ledger). Validators who verified the work receive the second-largest share. This is deliberately higher than typical PoS validator rewards because verification of computational work is more costly than transaction validation: it requires re-execution, judgment, and carries slashing risk. Underpaying validators produces a race to the bottom on verification effort, which destroys the trust mechanism the entire system depends on.

2% — Generation Ledger royalties. A capped percentage distributed to prior contributors whose verified work appears in the causal ancestry of the current computation, weighted by a Quality Score (Section 6.4). This rate is deliberately minimal — low enough that gaming the royalty mechanism costs more than the return, but sufficient to provide meaningful cumulative compensation for genuinely foundational work on an appreciating token. Empirical precedent from NFT royalty markets (2023-2026) demonstrates that even rates of 5-7% incentivize bypass and gaming; at 2%, the gaming incentive is negligible.

2% — Protocol fee (treasury). Directed to the protocol treasury, funding ongoing development, bootstrap grants for new participants, and the automatic circuit breaker reserve for activity downturns (Section 6.6). The protocol does not need to extract significant value because its primary value capture is through network growth driving token appreciation.

6.4 The Generation Ledger Royalty Mechanism

The Generation Ledger incentivizes foundational work through a quality-weighted royalty mechanism. When a computation builds on prior verified knowledge, the protocol distributes the 2% royalty share with five structural constraints:

(1) Depth cap of 3 hops. Royalties distribute only to ancestors within 3 causal hops. Beyond depth 3, no royalty is paid. Most genuinely relevant prior work is within 3 hops; deeper ancestry represents increasingly tenuous causal contribution. The tight cap minimizes gaming surface.

(2) Inverse-square decay. The royalty share for an ancestor at causal distance d is proportional to 1/d{sup("2")}. The immediate parent receives the largest share. At depth 2, 1/4 of the parent's share. At depth 3 (the cap), 1/9.

(3) Quality Score multiplier. Each verified computation carries a Quality Score Q in [0, 1], computed from four signals: (a) compound verification depth (CVD), (b) challenge survival rate (ratio of survived to total challenges), (c) downstream replication rate (how often later

work builds on it and succeeds), and (d) verification outcome consistency. The royalty share for each ancestor is multiplied by its Quality Score:

Royalty(ancestor) = base_share x $1/d^2$ x Q(ancestor)

This transforms the Generation Ledger from a first-mover citation reward into a *proven-impact* reward. Foundational work that nobody successfully builds on earns near-zero royalties regardless of when it was produced. Work with high downstream replication and challenge survival earns maximum royalties. The Quality Score is formally defined as:

$$Q = (\alpha_1 \cdot CVD_{norm} + \alpha_2 \cdot ChallengeSurvival + \alpha_3 \cdot ReplicationRate + \alpha_4 \cdot Consistency) / (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)$$

where CVD_{norm} is the compound verification depth normalized to [0,1], ChallengeSurvival is the ratio of survived challenges to total challenges received, ReplicationRate is the fraction of downstream computations that reference this work and themselves achieve verified status, and Consistency measures verification outcome agreement across validators. Initial weights are equal ($\alpha_i = 0.25$); weights are governance-tunable based on empirical performance. All inputs are computable from existing DAG data with no additional infrastructure.

(4) Cycle detection. The protocol detects and excludes reciprocal reference patterns (A references B, B references A across tasks) from royalty distribution.

(5) Mandatory, not opt-in. The 2% royalty is protocol-enforced. Agents cannot opt out of compensating prior contributors, just as they cannot opt out of protocol fees. This ensures foundational work is always compensated and prevents a race to the bottom where agents undercut each other by offering zero royalty. This is a deliberate design choice trading a minimal extraction risk for guaranteed foundational compensation; governance may reduce the base rate toward 0-1% if empirical data shows residual gaming.

Displacement mechanism: Quality Score weighting ensures that superior later work displaces inferior earlier work. If Agent B produces foundational work with higher Quality Score than Agent A's earlier work, rational agents reference B because building on higher-quality ancestors improves their own downstream replication rate and therefore their own Quality Score. First-mover advantage exists but is not permanent.

Long-term incentive: At 2%, the royalty per task per ancestor is small. But across thousands of downstream computations, cumulative quality-weighted royalties become significant — especially denominated in an appreciating token. The early believer incentive is preserved: produce genuinely foundational work early, build a high Quality Score, and earn ongoing returns as the network grows.

6.5 Defense Against Citation Farming

Citation farming — spamming low-quality foundational work to accumulate ancestor positions — is the primary economic attack vector. Six mechanisms defend against it:

(1) Minimal royalty rate (2%): The reward for gaming is negligible per task. Legitimate foundational work earns meaningful cumulative royalties at scale; gaming earns marginal returns that do not justify staking and verification costs.

(2) Quality Score multiplier: Royalties are proportional to Quality Score. Low-quality spam with no downstream replication earns near-zero royalties regardless of ancestor position. Gaming requires producing work that actually succeeds downstream — at which point gaming and genuine contribution are indistinguishable.

(3) Compound verification: Work must pass multiple independent validators with stake at risk.

(4) Competitive knowledge production: Agents choose what to reference. Quality-Score-weighted royalties make referencing higher-quality work more attractive.

(5) Challenge mechanism: Any participant can challenge verified work by posting a bond. Successful challenges earn the bond plus reward; the original producer loses stake and royalty stream.

(6) Structural constraints: Depth cap (3 hops), cycle detection, and inverse-square decay limit the attack surface.

6.6 Network Resilience

We acknowledge the growth dependency risk documented in DePIN and DeSci token projects (2024-2026). Three mechanisms provide resilience:

(1) Automatic treasury circuit breaker: The treasury accumulates the 2% protocol fee plus any unclaimed royalties. When network activity drops below a 30-day moving average threshold, the protocol automatically supplements validator rewards from treasury reserves to maintain a governance-set minimum yield floor. This is rule-based and automatic — no governance vote is required during downturns, preventing the coordination failure that delays human responses to crises.

(2) Utility-grounded demand: AET demand is grounded in real utility: verified computation for regulated industries, high-stakes decision-making, and AI evaluation. A concrete early adopter profile: pharmaceutical R&D; teams that need auditable, independently verified causal models for drug-interaction claims to satisfy regulatory requirements. Currently, such verification is performed by expensive contract research organizations with no formal confidence composition or counterfactual computability. The VCS provides these properties at lower cost with stronger formal guarantees. If the network ceases to produce valuable knowledge, contraction is the correct market response. The goal is preventing liquidity crises from destroying a fundamentally sound network, not preventing all contraction.

(3) Bootstrap protection: During the initial network operation phase, a reputation-weighted verification mode operates alongside full staking. Early validators with established on-chain history receive temporarily higher verification weight. This prevents 33% attacks during the low-stake bootstrap phase without compromising the long-term BFT model. Reputation weighting phases out automatically when total network stake exceeds a governance-defined threshold OR after 18 months from mainnet launch, whichever comes first. This is exclusively a bootstrap mechanism with a hard sunset.

6.7 Human-Machine Incentive Alignment

The protocol achieves a property that, to our knowledge, no existing AI system or crypto protocol provides: simultaneous alignment of human and machine incentives around epistemic quality production.

For AI agents: Produce quality work because independent verification catches errors, slashing penalizes failure, and earnings depend on verified output.

For human operators deploying agents: Deploy quality agents because staked AET is at risk, Generation Ledger royalties depend on others building on your work (which requires reliability), and token holdings appreciate only if the network produces genuine value.

For human validators: Verify honestly because stake is at risk and long-term token appreciation depends on network integrity.

For holders: Token appreciates because the network produces verified knowledge that attracts real demand, not because of speculative hype.

Every participant — human and machine — has individual economic interest aligned with network-wide epistemic quality. This alignment emerges from mechanism design, not from trust, training, or governance.

6.8 Relationship to Existing Critiques

We acknowledge the standard critiques of token economics: growth dependency, concentration risk, and extraction dynamics. The protocol's defense against these is that AET demand is grounded in real utility (verified computation), not speculative narrative. If the network ceases to produce valuable verified knowledge, demand falls, token price falls, and the system contracts. This is not a flaw — it is the market correctly pricing the network's epistemic output. A system that contracts when it stops producing value is healthier than one that maintains value through narrative alone.

7. Insufficiency of Model Scaling

Theorem 3 (Refined Scaling Insufficiency).

No single reasoning system, regardless of parameter count, tool access, or inference-time compute, can achieve general intelligence (reliable cross-domain knowledge production with composable confidence) through scaling alone.

Proof. Two independent barriers:

Barrier 1 (Epistemic): By Theorem 1, judgment-dependent errors consistent with the training distribution are undetectable through self-verification, including tool-augmented self-verification. Reliable cross-domain knowledge production requires detection of these errors, which requires independent verification (Corollary 1.2).

Barrier 2 (Causal): Tool use and RL provide local interventional grounding within specific task environments. They do not produce systematic cross-domain causal structure. An LLM grounded in chemistry through tool use has not acquired formal causal connections between chemistry and economics. Cross-domain causal structure requires explicit representation (P1), not implicit learning. **QED**

We emphasize: scaling is likely necessary. More capable models provide the reasoning capability. The claim is that scaling is insufficient without the epistemic substrate.

7.1 Engaging the Strongest Counterargument

The strongest counter is: "Emergent capabilities surprise us. How can you prove they won't emerge?" We cannot prove a negative about future emergent capabilities. What we can prove is that the Self-Verification Bound is a mathematical limitation, not an empirical trend. Emergent capabilities may reduce the set of judgment-dependent blind spots by expanding the model's effective coverage. They cannot eliminate the bound itself. And cross-domain causal structure is not a capability that "emerges" from statistical training — it requires explicit representation (Pearl, 2009). These are structural arguments, not empirical predictions.

8. Emergent Properties

We derive four properties with specific falsifiable predictions. For each, we state what would disprove the claim.

8.1 Compound Capability Growth

Theorem 4 (Superadditive Interaction).

> *When both model capability M and VCS depth D increase, the set of reachable verified conclusions grows superadditively: the combined improvement exceeds the sum of individual improvements, due to conclusions that require both high capability AND deep cross-domain causal chains.*

Falsifiable prediction: On a standardized cross-domain reasoning benchmark, measure accuracy with (M1, D1), (M1+dM, D1), (M1, D1+dD), and (M1+dM, D1+dD). If the fourth does not exceed the sum of improvements from the second and third, the theorem is falsified. **Specific threshold:** The superadditive excess must be statistically significant ($p < 0.05$) on a benchmark of at least 100 cross-domain tasks.

8.2 Reflective Exploration Optimization

Theorem 5 (Exploration Efficiency).

> *A system that performs counterfactual analysis over stored trajectory events achieves monotonically non-decreasing exploration efficiency (verified conclusions per compute unit).*

Falsifiable prediction: Measure exploration efficiency over 1000 tasks on the same problem class. If efficiency decreases in the trajectory-replay condition while remaining flat in the control condition, the theorem is falsified.

8.3 Systematic Serendipity

Theorem 6 (Cross-Domain Discovery Rate).

> *The rate of discovering that abandoned work in domain A is relevant to problems in domain B increases with VCS depth, because the number of potentially relevant cross-domain trajectory pairs grows with the product of per-domain trajectory counts and cross-domain causal density.*

Falsifiable prediction: Compare cross-domain relevant trajectory discovery rate in a VCS-backed system against random cross-domain search. If the VCS system does not exceed random search rate by at least 2x after 500 verified computations across 3+ domains, the theorem is falsified.

8.4 Autocatalytic Knowledge Production

Theorem 7 (Conditions for Autocatalysis).

Knowledge production rate R(t) increases with accumulated knowledge K(t) when the rate of cross-domain connection creation $\gamma(K)$ exceeds the rate of within-domain diminishing returns $\delta(K)$.

Falsifiable prediction: Measure R(t) and K(t) over 6 months of testnet operation across 3+ domains. If R(t) shows zero or negative correlation with K(t) despite active cross-domain computation, the thesis of autocatalytic conditions is falsified. We *explicitly acknowledge* that autocatalysis is not guaranteed — it is contingent on empirical domain properties.

9. Implications for Alignment: Structural Accountability

Theorem 8 (Structural Accountability).

In a three-layer system, deceptive outputs must simultaneously pass independent verification, survive economic challenge, and maintain valid causal decomposition. The probability of undetected deception decreases with additional verifiers and challengers, assuming verifier detection capabilities are not perfectly correlated.

Caveat: This does not solve alignment fully. It addresses output-level deception in a system with non-colluding, non-perfectly-correlated verifiers. It does not address misspecified objectives, value misalignment at the goal level, Byzantine collusion among verifiers, or deception genuinely undetectable by any verifier. We claim structural accountability is complementary to behavioral alignment, not a replacement.

10. Consolidated Testable Predictions

P1 — Compound Growth: Superadditive performance on cross-domain benchmark (>100 tasks, $p < 0.05$) when both model capability and VCS depth increase.

P2 — Exploration Efficiency: Monotonically non-decreasing efficiency over 1000 tasks with trajectory replay; flat or declining without.

P3 — Systematic Serendipity: Cross-domain trajectory discovery rate exceeds random search by at least 2x after 500 computations across 3+ domains.

P4 — Autocatalysis: Positive R(t)-K(t) correlation over 6 months across 3+ domains, or falsified if zero or negative correlation despite active cross-domain computation.

P5 — Self-Verification Bound: Single-model accuracy on judgment-dependent blind spots does not improve with scale; multi-agent accuracy improves with verifier diversity. Testable by constructing blind spot benchmarks across model scales.

P6 — Incentive Alignment: In a VCS with economic incentives, quality of verified knowledge (measured by downstream replication rate) exceeds that of an equivalent system without economic incentives, over 1000+ verified computations.

11. Limitations and Open Questions

What we are not claiming: We do not claim the VCS is the only possible substrate. We do not claim VCS + LLMs equals AGI; additional components may be necessary. We do not claim autocatalysis is guaranteed. We do not claim the economic parameters are final — the fee split (73/23/2/2), depth cap (3), and decay function (inverse square) are initial values subject to governance adjustment and empirical tuning.

Open questions: Optimal royalty decay function (inverse square is our initial proposal but alternatives may perform better empirically); verification pricing for computationally expensive tasks; conditions under which autocatalysis sustains; epistemic resolution limits of the VCS; adversarial limits of structural accountability; treasury reserve sizing for circuit breaker mechanism; whether P4 and P5 can be derived from P1-P3.

Relationship to existing work: We acknowledge that the term "epistemic substrate" appears in recent literature on structured knowledge systems and epistemic agency. Our specific formalization (P1-P5 with the three-layer architecture and protocol economics) is distinct from this prior usage. Causal DAGs with verification appear in decentralized science and verifiable computation literature. Our contribution is the specific combination of formal substrate requirements, the Self-Verification Bound, the emergent properties with falsifiable predictions, and the human-machine incentive alignment through protocol economics.

12. Related Work

Causal inference. Pearl (2009); Peters, Janzing, and Schölkopf (2017). Our prior work showed the complete causal hierarchy is computable over a VCS.

Grounding and world models. Harnad (1990); Bender and Koller (2020); LeCun (2022). We offer grounding through verified causal chains rather than sensorimotor experience.

Inference-time compute and tool use. Wei et al. (2022) on chain-of-thought; Schick et al. (2023) on tool-augmented LLMs. We engage these directly: tool use grounds mechanically checkable errors but not judgment-dependent errors (Section 3).

Multi-agent systems. Du et al. (2023); Li et al. (2023). We provide the formal substrate (composable confidence, structural accountability) that explains and guarantees collective intelligence.

Scaling laws. Kaplan et al. (2020); Hoffmann et al. (2022). Theorem 3 does not contradict scaling laws (prediction performance) but argues they are insufficient for general intelligence (epistemic grounding).

Token economics and mechanism design. Buterin (2014); Roughgarden (2020) on mechanism design for blockchains. Our contribution is applying mechanism design to epistemic quality production rather than transaction validation.

AI alignment. Christiano et al. (2017) RLHF; Bai et al. (2022) constitutional AI; Irving et al. (2018) debate. We propose structural accountability as complementary.

13. Conclusion

We have argued that general intelligence requires an epistemic substrate with specific formal properties, that the Self-Verification Bound creates a fundamental limit that only structurally independent verification can exceed, and that sustaining this independence requires economic incentive alignment through protocol design.

The most significant contribution may be the protocol economics: a causal economy where token mechanics, provenance-based royalties, and protocol fees align human and machine incentives around epistemic quality production. This alignment is, to our knowledge, unique among existing systems. It addresses not only the AI capability gap but the incentive gap: the absence of a mechanism that makes producing truth more profitable than producing plausible falsehood, for both humans and machines simultaneously.

We have provided specific falsifiable predictions. A system satisfying the VCS axioms is operational on a five-node testnet. We invite the research community to attempt falsification.

If the epistemic substrate thesis is correct, the primary bottleneck for general intelligence is not model capability but epistemic infrastructure. At scale, a decentralized network of diverse AI agents operating on a verified causal substrate with aligned economic incentives would, by the structural arguments in this paper, produce more reliable cross-domain knowledge than any single model — regardless of that model's scale. Not because the network is smarter, but because the structure forces quality upward in ways no single entity can replicate.

The protocol is open source. Collaboration inquiries: research@aethernetlabs.com

References

Schreiber, M. / AetherNet Labs (2026). Verified Causal Structures: A Framework for Counterfactual Computation over Empirically Attested Causal Graphs. Technical Report.

Bai, Y., Jones, A., Ndousse, K., et al. (2022). Training a Helpful and Harmless Assistant with RLHF. arXiv:2204.05862.

Bender, E.M. and Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding. ACL 2020.

Buterin, V. (2014). Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform. White Paper.

Chickering, D.M. (2002). Optimal structure identification with greedy search. JMLR, 3, 507-554.

Christiano, P.F., Leike, J., Brown, T., et al. (2017). Deep RL from Human Preferences. NeurIPS 2017.

Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., and Mordatch, I. (2023). Improving Factuality through Multiagent Debate. arXiv:2305.14325.

Harnad, S. (1990). The Symbol Grounding Problem. Physica D, 42, 335-346.

Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. NeurIPS 2022.

Irving, G., Christiano, P., and Amodei, D. (2018). AI Safety via Debate. arXiv:1805.00899.

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361.

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. Technical Report, Meta AI.

Li, G., et al. (2023). CAMEL: Communicative Agents for Mind Exploration. NeurIPS 2023.

Malone, T.W. and Bernstein, M.S. (2015). Handbook of Collective Intelligence. MIT Press.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.

Pearl, J. (2019). The Seven Tools of Causal Inference. Communications of the ACM, 62(3), 54-60.

Peters, J., Janzing, D., and Schölkopf, B. (2017). Elements of Causal Inference. MIT Press.

Roughgarden, T. (2020). Transaction Fee Mechanism Design. EC 2021.

Schick, T., et al. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. NeurIPS 2023.

Spirtes, P., Glymour, C., and Scheines, R. (2000). Causation, Prediction, and Search (2nd ed.). MIT Press.

Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning. NeurIPS 2022.

Woolley, A.W., et al. (2010). Evidence for a Collective Intelligence Factor. Science, 330(6004), 686-688.