Veritas: Decentralized Truth Discovery Through Emergent Consensus[*]

Josh van Cuyck
x.com/joshvancuyck

Last updated: August 17, 2025

"The price of apathy towards public affairs is to be ruled by evil men."
[5]
— Plato

Abstract

Veritas is a decentralized protocol for truth discovery that extends Bayesian Truth Serum from discrete choices to continuous probability distributions, operating in temporal, asynchronous settings. Agents stake capital and report both their beliefs and meta-predictions of others' beliefs. A leave-one-out "density surprise" score rewards positions where actual belief density exceeds predicted density, the continuous analogue of being "surprisingly popular." This design prevents self-influence and makes truthful reporting the natural strategy under common priors and commit-reveal, as genuine surprise can only arise from correlated private information. Veritas provides a principled foundation for collective knowledge formation without arbiters, even when no verifiable ground truth exists.

[*] Draft v0.1. This is a working paper. Please do not cite without permission.

Contents

1 Introduction: The Crisis of Truth **3**

2 Philosophical Foundations **3**

2.1 Epistemological Foundations 3

2.2 Information Theory and Entropy 4

2.3 Economic and Game-Theoretic Principles 5

2.4 Temporal Dynamics and Emergent Truth 6

3 Protocol Design **6**

3.1 System Overview 7

3.2 Epistemic Weight and Stake Economics 7

3.3 Commit-Reveal Process 8

3.4 Aggregation and Entropy 8

3.5 Certainty as Master Control 9

3.6 Mirror Descent Updates 10

3.7 Learning Assessment and Scoring 11

3.8 The Temporal Dissent Goldilocks Zone 14

3.9 Belief Creation as Epistemic Commitment 14

3.9.1 Structural Parameters and Market Integrity 15

3.9.2 Adaptive Governance via Meta-Belief Feedback 15

4 Delegation Mechanics **15**

5 Self-Referential Governance **15**

6 Conclusion **16**

1 Introduction: The Crisis of Truth

Information is humanity's most fundamental coordination technology. Throughout history, the processes for discovering and validating truth have often been shaped by centralized control. Those who controlled information flow also shaped collective beliefs, aligning them with their own interests.

Societies organize themselves around shared beliefs about reality. When these beliefs are grounded in reality, they enable cooperation and progress. When they drift from it, coordination breaks down, and resources are misallocated, leading to preventable crises such as famines, financial collapses, and failed responses to global risks.

The internet once promised to democratize information, bypassing traditional gatekeepers. Yet in practice, digital platforms have reproduced the same power dynamics at scale. Algorithms prioritize engagement over accuracy, echo chambers reinforce existing biases, and an overwhelming abundance of information makes it harder, not easier, to discern what is true. Entire segments of society now operate from fundamentally different understandings of reality.

This fragmentation undermines our ability to address collective challenges, from climate change to AI safety. The problem is not disagreement itself; diversity of thought is essential. Rather, it is the absence of neutral infrastructure where competing beliefs can be tested, refined, and converge. Current institutions, shaped by their own incentives and constraints, cannot fulfill this role.

What is needed is a new foundation for coordination: one that allows collective intelligence to emerge organically. Such a system could help surface hidden knowledge, support value-aligned decision-making, and enable societies to adapt shared norms and governance in step with changing realities.

This paper introduces VERITAS, a protocol designed to provide this foundation. By combining game theory, information theory, and distributed systems, VERITAS creates conditions where truth can emerge through aligned incentives rather than centralized control. It represents a shift from truth imposed by authority to truth discovered collectively.

2 Philosophical Foundations

2.1 Epistemological Foundations

The question of what constitutes knowledge and truth has occupied philosophy since ancient Greece. Traditional epistemology often frames the debate as a dichotomy: is truth an external reality that exists independently of observers (realism), or is it constructed through experience and interpretation (anti-realism)? For systems of coordination, however, neither extreme suffices. Pure objectivism neglects the distributed, partial knowledge of real-world agents, while pure relativism offers no stable ground for collective convergence.

Recent developments in social epistemology [1, 2, 4] highlight that knowledge often emerges through collective processes. In distributed systems, three epistemological categories are particularly relevant:

1. **Objective Reality**: Some truths exist independently of any observer: physical laws, mathematical theorems, causal relationships. Yet even these must be mediated through subjective experience and validated via social consensus. Scientific facts, for instance, are maintained by communities of inquiry.
2. **Subjective Experience**: Individual preferences, aesthetic judgments, and personal values are inherently first-person and non-transferable. While they cannot be directly shared, they can be expressed and aggregated to reveal meaningful patterns.
3. **Intersubjective Construction**: Many essential "truths" are sustained by collective belief. Concepts such as money, property rights, and social norms are not physical realities, but they exert real causal power because societies act as if they are real.

Veritas operates across these domains by focusing on functional truth: beliefs that enable effective coordination and guide action. Rather than resolving age-old philosophical disputes, the protocol provides a neutral substrate where diverse types of truth can emerge, interact, and stabilize. It supports the discovery of objective facts, the aggregation of subjective perspectives, and the creation of new intersubjective realities.

In Veritas, truth manifests as convergence. When independent agents, each with partial and localized information, arrive at similar beliefs, their overlap serves as a stronger signal than any individual assessment. Permissionless and censorship-resistant infrastructure further ensures that these convergences reflect authentic consensus, untainted by institutional filtering or manipulation. In Veritas, the process of belief aggregation reduces collective uncertainty, anchoring truth discovery in measurable entropy reduction.

2.2 Information Theory and Entropy

Information theory [7, 8] provides a mathematical framework for understanding truth discovery as the transformation of noise into signal. In this context, high entropy represents maximal uncertainty: a landscape of scattered, partial knowledge. As beliefs converge, entropy declines, and coherent collective understanding emerges.

Shannon [7] defined information as the reduction of uncertainty, formally expressed as information gain:

$$IG(T, a) = H(T) - H(T|a)$$

where learning something a about a topic T reduces our uncertainty from $H(T)$ to $H(T|a)$.

KL divergence quantifies the informational divergence between probability distributions:

$$D_{KL}(P\|Q) = \mathbb{E}_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right]$$

It measures how much information is lost when approximating P with Q. In epistemic contexts, KL divergence captures surprise: when actual outcomes differ from predictions, the divergence measures the magnitude of that surprise in information-theoretic units.

From an information geometry perspective, probability distributions form a manifold where KL divergence provides the natural divergence measure. Beliefs that are "far apart" in this space represent fundamentally different worldviews. When multiple agents independently position themselves in the same sparse region while underestimating how many others will do likewise, they create the continuous analogue of Bayesian Truth Serum's "surprisingly popular" mechanism: density that exceeds collective predictions.

While KL divergence measures directional information distance, collective disagreement requires a symmetric measure. Jensen-Shannon disagreement entropy provides this symmetry while respecting honest uncertainty about inherently random outcomes. Unlike traditional entropy measures that penalize expressing uncertainty, it distinguishes between reducible epistemic uncertainty (which decreases as evidence emerges) and irreducible aleatory uncertainty (which should be preserved). When agents converge while maintaining appropriate uncertainty about inherently uncertain outcomes, this signals authentic knowledge discovery rather than false confidence.

As agents submit beliefs, epistemic uncertainty diminishes asynchronously across the network, reflecting uneven evidence arrival. This temporal dynamic allows early knowledge holders to signal truth before it becomes common knowledge, transforming distributed asynchrony into an engine for rewarding genuine epistemic contribution.

2.3 Economic and Game-Theoretic Principles

Entropy reduction is value creation. When uncertainty declines through evidence rather than mere coordination, authentic knowledge emerges, enabling better decisions, efficient resource allocation, and coordinated action.

Markets have long been humanity's most effective mechanism for aggregating distributed information. As Hayek [3] observed, prices encode dispersed knowledge no planner could fully access. VERITAS extends this logic: instead of trading goods, agents stake capital on beliefs and earn returns for genuine epistemic contribution.

Epistemic weight combines two factors: stake and belief-specific trust. Stake provides skin in the game; trust reflects proven ability to deliver valuable signals under uncertainty. This prevents capital alone from determining influence: credibility must be earned.

Building on Bayesian Truth Serum [6], VERITAS addresses a fundamental problem in information elicitation: when agents are rewarded based on how well their reports match others', they have incentive to guess what others will say rather than report their true beliefs. This creates a self-referential loop where everyone tries to predict everyone else's predictions. BTS breaks this circularity by requiring agents to report both their own beliefs and their predictions of others' beliefs, making truthful revelation the dominant strategy regardless of what others do.

This design fosters specialization: agents focus where their expertise has highest impact. Over time, the system self-organizes into an efficient division of epistemic labor, leveraging the same distributed information aggregation principle that makes markets effective.

2.4 Temporal Dynamics and Emergent Truth

VERITAS supports two distinct temporal categories of truth, illustrating why dynamic mechanisms are essential:

Living Truth Realities that evolve over time. These cannot be captured through static snapshot elicitation, as the ground truth itself is in flux. VERITAS enables continuous reassessment, allowing beliefs to track shifting realities as new evidence emerges.

Emergent Collective Belief Even for static facts, humanity's understanding evolves through interaction. Participation in VERITAS shapes collective belief, surfacing patterns that no individual or institution could predefine. These truths emerge organically from decentralized expression and cannot be extracted; they must be cultivated.

VERITAS is designed for this dynamic reality. The protocol continuously updates beliefs as new information arrives, while simultaneously detecting emergent consensus through collective convergence patterns.

3 Protocol Design

*For discrete distributions with k outcomes. For continuous distributions, both entropy and disagreement entropy are unbounded but normalized using historical maximum.

Table 1: Core Protocol Variables

Variable	**Description**	**Range**
S_i	Total stake for agent i	$[0, \infty)$
n_i	Active belief count for agent i	$\mathbb{N}$
$T_{i,b}$	Trust multiplier for agent i in belief b	$[1.0, \infty)$
$W_{i,b}$	Epistemic weight of agent i in belief b	$[0, \infty)$
p_i	Agent i's belief (discrete: $p_i(x)$, continuous: μ_i, σ_i)	See note**
m_i	Agent i's meta-prediction	See note**
$P_{\text{pre}}(x)$	Pre-mirror descent aggregate	Distribution
$P_{\text{post}}(x)$	Post-mirror descent aggregate	Distribution
$H(P)$	Entropy of distribution P	$[0, \log k]$*
$\hat{H}(P)$	Normalized entropy	$[0, 1]$
D_{JS}	Jensen-Shannon disagreement entropy	$[0, \log k]$*
$\hat{D}_{JS}$	Normalized disagreement entropy	$[0, 1]$
η	Learning rate	$[0, 1]$
c	Certainty (1 - normalized disagreement entropy)	$[0, 1]$
s_i	BTS score for agent i	$(-\infty, \infty)$
$\alpha_{\max}$	Max stake allocation per belief	$[0.05, 0.5]$

**For discrete beliefs: $p_i(x) \in [0, 1]$ with $\sum_x p_i(x) = 1$. For continuous beliefs: $\mu_i \in \mathbb{R}$, $\sigma_i > 0$.

3.1 System Overview

Agents stake capital and submit beliefs as probability distributions, not point estimates. This preserves honest uncertainty about inherently random outcomes.

The protocol aggregates these weighted beliefs into a collective distribution. When agents independently cluster in unexpected regions while underestimating this clustering, they earn positive information scores: the continuous extension of BTS's "surprisingly popular" mechanism.

The system maintains temporal continuity by preserving belief state between epochs. When the network learns, passive beliefs update toward the new aggregate through mirror descent, scaled by collective certainty. This ensures dormant agents benefit from revealed information without active participation.

Rewards flow only when the network learns: when entropy decreases between epochs. Winners are those whose positions contributed signal; losers added noise. This zero-sum redistribution occurs in proportion to information scores.

Epistemic weight combines stake and trust. Trust grows when agents outperform their weight-proportional expectation, creating meritocracy beyond pure capital. High trust raises expectations, making influence self-regulating.

The aggregate belief represents current collective knowledge. Through continuous revelation of private information and rational belief updates, truth emerges without central authority: an efficient market for epistemic discovery.

3.2 Epistemic Weight and Stake Economics

Agents maintain a unified stake pool that dynamically allocates across all active beliefs. Similar to Proof-of-Stake networks, where a validator's stake secures multiple operations, Veritas allows a single agent's stake to be distributed simultaneously across any number of beliefs they choose to engage with. This design addresses the capital inefficiency of traditional prediction markets, which require separate, fragmented liquidity for each question.

The effective stake per belief equals the agent's total stake divided by their active belief count, subject to a governance-defined maximum allocation per belief. This cap limits the influence of wealthy actors on individual propositions while preserving their ability to participate broadly.

Thus epistemic weight resolves to the following formula, where both trust and stake scale dynamically with performance, but at varying rates to achieve different ends:

$$W_{i,b} = \min\left(\frac{S_i}{n_i}, \sigma_{\max} \times S_i\right) \times T_{i,b}$$

where S_i represents total stake, n_i is the agent's active belief count, $\sigma_{\max}$ is the governance-determined maximum stake allocation per belief, and $T_{i,b}$ is belief-specific trust starting at 1.0.

For aggregation purposes, these weights are normalized to ensure proper probability distributions:

$$w_i = \frac{W_{i,b}}{\sum_j W_{j,b}}$$

where w_i represents the normalized weight of agent i in the current belief.

3.3 Commit-Reveal Process

Veritas uses a two-phase commit-reveal process to preserve the privacy required for Bayesian Truth Serum's incentive compatibility. This design prevents strategic coordination by ensuring agents cannot adjust submissions based on observed beliefs. In the commit phase, agents submit encrypted payloads containing their belief distributions and may target any future epoch. Commitments remain updatable until revelation, giving agents temporal flexibility to manage timing risk when information propagates unevenly through the network. At the close of each epoch, the protocol deterministically reveals all committed beliefs, enabling aggregation and scoring.

3.4 Aggregation and Entropy

Aggregation proceeds through weighted averaging using the normalized epistemic weights w_i defined in Section 3.2. The pre-mirror descent aggregate P_{pre} begins with the previous epoch's post-mirror descent state, then incorporates all changes: existing agents whose committed beliefs are revealed in this epoch have their positions updated, newly participating agents are added with their initial submissions, and agents who withdrew are removed. This creates a complete snapshot of current collective belief before any convergence mechanism is applied.

For scoring purposes, the protocol also computes leave-one-out aggregates that exclude the agent being scored, preventing self-influence in the BTS mechanism.

For discrete beliefs, the aggregate distribution is:

$$P_{\text{pre}}(x) = \sum_i w_i \cdot p_i(x)$$

For continuous Gaussian beliefs where each agent's belief is $p_i = \mathcal{N}(\mu_i, \sigma_i^2)$, the aggregate is $P_{\text{pre}} = \mathcal{N}(\mu_{\text{pre}}, \sigma_{\text{pre}}^2)$ with parameters:

$$\mu_{\text{pre}} = \sum_i w_i \cdot \mu_i, \quad \sigma_{\text{pre}}^2 = \sum_i w_i \cdot \sigma_i^2$$

This approach preserves uncertainty in the aggregate, aligning with the use of Jensen-Shannon (JS) disagreement entropy, which respects honest uncertainty rather than penalizing it. The resulting aggregate reflects the immediate consensus among active participants.

JS disagreement entropy measures the informational content of collective disagreement: the extent to which individual beliefs diverge from their aggregate. Crucially, it distinguishes disagreement from uncertainty: agents can reach perfect consensus (zero JS disagreement) while collectively acknowledging high uncertainty about an outcome. This distinction respects the difference between reducible epistemic uncertainty and irreducible aleatory uncertainty. Mathematically, JS disagreement is defined as the difference between the entropy of the mixed aggregate distribution and the weighted average of individual entropies. High values indicate significant epistemic diversity; low values indicate convergence, whether certain or uncertain. The protocol computes this disagreement measure as:

$$D_{JS} = H\left(\sum_i w_i p_i(x)\right) - \sum_i w_i H(p_i)$$

where $H(\cdot)$ denotes Shannon entropy. The specific form depends on the belief type:
For discrete distributions:

$$H(P) = -\sum_x P(x) \log_2 P(x)$$

For continuous Gaussian distributions:

$$H(P) = \frac{1}{2} \log_2(2\pi e \sigma^2_{\text{pre}})$$

This captures the uncertainty inherent in the variance.

Disagreement Entropy Normalization

To derive certainty values in $[0, 1]$, disagreement entropy is normalized:

$$\hat{D}_{JS} = \frac{D_{JS}}{D_{JS,\max}}$$

where:

- For discrete beliefs with k outcomes: $D_{JS,\max} = \log_2(k)$
- For continuous beliefs: $D_{JS,\max} = \max_{t' \leq t} D_{JS,t'}$ (historical maximum)

3.5 Certainty as Master Control

The protocol derives a certainty measure from normalized disagreement entropy:

$$c = 1 - \hat{D}_{JS}$$

This single parameter represents the network's collective confidence, ranging from 0 (maximum disagreement) to 1 (complete consensus). Certainty serves as a master control variable used throughout the protocol to modulate behavior. When uncertainty is high, the system protects epistemic diversity and encourages exploration; when certainty is high, it accelerates convergence and locks in consensus. This adaptive response ensures VERITAS remains responsive to its epistemic state without requiring external governance or manual tuning.

3.6 Mirror Descent Updates

Veritas assumes each agent holds private information not directly observable by others.[1] Each manual belief update is treated as a revelation of that private information, triggering a system-wide adjustment. The protocol updates all passive agents' beliefs rationally, scaling adjustments by the network's collective certainty. In effect, every new signal reshapes the global belief landscape as non-participating agents gradually incorporate the revealed information.

This design is informed by Aumann's Agreement Theorem, which shows that rational agents with common priors and common knowledge cannot agree to disagree. While Veritas does not assume shared priors or full transparency, it does model a rational convergence process over time. As private information surfaces through belief updates, collective entropy naturally decreases. Mirror descent provides a mathematically grounded mechanism for realizing this convergence.

Mirror descent extends BTS to temporal settings by maintaining a living epistemic system where collective understanding accumulates over time. This adaptation is necessary for modeling changing truths and asynchronous information propagation. Passive agents' beliefs evolve with emerging consensus even when not actively participating, ensuring the aggregate reflects both current signals and accumulated knowledge.

The learning rate for these updates adapts to network certainty:

$$\eta = c = 1 - \hat{D}_{JS}$$

Discrete Beliefs

Updates follow a multiplicative (geometric) interpolation that preserves the probabilistic structure:

$$p_i^{(t+1)}(x) = \frac{p_i^{(t)}(x)^{1-\eta} \cdot P_{\text{pre}}(x)^{\eta}}{\sum_{x'} p_i^{(t)}(x')^{1-\eta} \cdot P_{\text{pre}}(x')^{\eta}}$$

This formula geometrically averages the agent's belief with consensus: when $\eta = 0$ (maximum uncertainty) beliefs remain unchanged, and when $\eta = 1$ (complete certainty) agents fully adopt the consensus.

Continuous Gaussian Beliefs

For beliefs represented as $\mathcal{N}(\mu, \sigma^2)$, the protocol applies linear interpolation to both parameters:

$$\mu_i^{(t+1)} = (1-\eta)\mu_i^{(t)} + \eta\mu_{\text{pre}}$$

$$\sigma_i^{(t+1),2} = (1-\eta)\sigma_i^{(t),2} + \eta\sigma_{\text{pre}}^2$$

This deviation from traditional mirror descent (which would use KL divergence and precision weighting) reflects Veritas's philosophy of respecting uncertainty:

1. **Symmetric treatment** – Uncertain and confident beliefs have equal influence, avoiding overconfidence bias.
2. **Preserves consensus** – When agents agree on uncertainty levels, that uncertainty is maintained rather than artificially reduced.

[1] We use the term "mirror descent" to maintain consistency with optimization literature, though our approach deviates from the traditional definition in important ways detailed below.

3. **Simplicity and coherence** – Variance is treated as legitimate information to be aggregated, not error to be minimized.

Traditional approaches like precision-weighted updates or geometric variance interpolation systematically favor low-variance (confident) beliefs. Veritas avoids this by applying linear interpolation in continuous beliefs, ensuring honest uncertainty is neither amplified nor suppressed. This reflects a core principle: uncertainty is valuable information about the limits of collective knowledge.

Notably, discrete and continuous beliefs differ in their update methods. Discrete beliefs use geometric interpolation to preserve probability ratios, while continuous beliefs rely on linear interpolation to mitigate confidence bias. This mathematical asymmetry achieves philosophical consistency: both approaches prevent confident beliefs from overwhelming uncertain ones in their respective domains. Agents gradually align with emerging consensus at a rate proportional to collective certainty, preserving epistemic diversity when it matters most.

3.7 Learning Assessment and Scoring

The protocol rewards agents only when collective learning occurs, measured by disagreement entropy reduction between epochs:

$$\Delta D_{JS} = D^{(t-1)}_{JS,\text{post}} - D^{(t)}_{JS,\text{post}}$$

When $\Delta D_{JS} > 0$, the network has learned, triggering scoring and economic redistribution. This simple rule provides a powerful protection for rational dissent: agents who increase entropy by challenging consensus are penalized only if subsequent information validates the original belief. In the absence of new information, no learning occurs, and dissenters face no economic loss.

This asymmetry transforms the economics of contrarian positions. Early dissenters who correctly identify flaws in collective belief can sustain their positions through high-entropy phases without penalty. They incur economic risk only when competing evidence emerges: exactly when the system should adjudicate between conflicting views. By design, Veritas avoids punishing legitimate skepticism while still rewarding contributions that drive genuine epistemic progress.

Agent Information Score

Each active agent's information score combines their epistemic weight with their BTS signal quality:

$$g_i = W_{i,b} \times s_i$$

where $W_{i,b}$ is the epistemic weight (stake × trust) and s_i is the BTS signal quality score.

Economic Learning Rate

The economic learning rate scales the reward pool based on the reduction in post-mirror descent disagreement entropy between consecutive epochs:

$$\eta_{\text{econ}} = \frac{\max(0, \Delta D_{JS})}{D^{(t-1)}_{JS,\text{post}}} \times 100\%$$

Score Normalization and Redistribution

Economic redistribution follows a zero-sum structure. Agents partition into winners (W) with $g_i > 0$ and losers (L) with $g_i < 0$, based on their information scores.

The total amount of stake to be redistributed is determined by multiplying the economic learning rate by the sum of all losing agents' effective stakes:

$$\text{Pool}_{\text{slash}} = \eta_{\text{econ}} \times \sum_{j \in L} S_{j,\text{eff}}$$

where η_{econ} is the economic learning rate (the percentage the network learned), L is the set of losing agents, and $S_{j,\text{eff}}$ is agent j's effective stake for this belief. This ensures that when the network learns more (higher η_{econ}), a larger portion of losers' stakes gets redistributed to winners.

The reward pool is set equal to the slashing pool, creating a zero-sum redistribution. Each loser is slashed proportionally to the noise they added to the system, while each winner is rewarded proportionally to the signal they contributed. The individual transfers are determined by:

$$\Delta S_j = \underbrace{\frac{|g_j|}{\sum_{k \in L} |g_k|}}_{\text{noise share}} \times \text{Pool}_{\text{slash}} \quad \text{and} \quad \Delta R_i = \underbrace{\frac{g_i}{\sum_{k \in W} g_k}}_{\text{signal share}} \times \text{Pool}_{\text{slash}}$$

This ensures exact conservation: $\sum_j \Delta S_j = \sum_i \Delta R_i$.

We hypothesize that if the network achieves collective learning without noise or malicious behavior, there would be no losers to slash and no rewards to distribute. This suggests Veritas may function more effectively as a positive-sum system. Zero-sum redistribution alone may be insufficient to sustain consistent information provision over time. Unlike prediction markets with directional bets and defined risk-reward profiles, participation in Veritas requires continuous effort: updating beliefs, monitoring information, and maintaining positions. To support this, the protocol incorporates market-like scoring for quality assessment but may ultimately rely on positive-sum revenue generation to incentivize sustained participation.

BTS Signal Quality

The signal quality score extends Bayesian Truth Serum from discrete to continuous outcome spaces, enabling truth discovery over continuous domains like prices or weather while preserving uncertainty tracking as beliefs evolve over time. Agents observing the same reality share correlated signals but underestimate how many others share them. When their mass lands in regions denser than predicted, they earn information gain. Fabricated or noisy reports lack this hidden correlation and do not generate surprise.

Agents report a belief distribution p_i and a meta-prediction m_i over others' beliefs. Let $\bar{p}_{-i}$ and $\bar{m}_{-i}$ denote leave-one-out weighted aggregates of others' beliefs and meta-predictions (excluding agent i), preventing self-influence. The score is:

$$s_i = \underbrace{D_{KL}(p_i \| \bar{m}_{-i}) - D_{KL}(p_i \| \bar{p}_{-i})}_{\text{information gain}} - \underbrace{D_{KL}(\bar{p}_{-i} \| m_i)}_{\text{prediction accuracy}}$$

The first two terms form the information gain (continuous "surprisingly popular"); the last is prediction accuracy (proper scoring on the realized distribution of others' reports). In the discrete setting, these first two terms collapse exactly to the first term of the original BTS score, showing that our formulation is a direct continuous generalization.

Continuous extension. In discrete BTS, the information term is a log frequency ratio evaluated at the reported bin. When agents can spread mass across many bins,

$$\sum_{A\in\Pi_h} p_i(A)\log\frac{\bar{p}_{-i}(A)}{\bar{m}_{-i}(A)}$$

is the natural extension. As the partition Π_h is refined ($h \to 0$), this Riemann sum becomes an expected log density ratio:

$$\int p_i(x)\log\frac{\bar{p}_{-i}(x)}{\bar{m}_{-i}(x)}dx = D_{KL}(p_i\|\bar{m}_{-i}) - D_{KL}(p_i\|\bar{p}_{-i})$$

i.e., a difference of KLs. Thus, KL is the continuous generalization of BTS's frequency-based surprise.

Truthfulness preservation. Under standard measurability and boundedness, the dominated convergence theorem justifies:

$$\sum_{A\in\Pi_h} p_i(A)\log\frac{\bar{p}_{-i}(A)}{\bar{m}_{-i}(A)} \xrightarrow{h\to 0} \int p_i(x)\log\frac{\bar{p}_{-i}(x)}{\bar{m}_{-i}(x)}dx$$

so the continuous information term is precisely the limit of the discrete one. The leave-one-out construction preserves the key BTS independence condition: $(\bar{p}_{-i}, \bar{m}_{-i})$ do not depend on i's report. The meta term $-D_{KL}(\bar{p}_{-i}\|m_i)$ remains strictly proper because the "outcome" is the realized distribution of others' reports. Under the usual BTS assumptions (common prior, conditionally independent signals, commit-reveal), the discrete proof carries over to the limit; truthful reporting is therefore a Bayes-Nash equilibrium by inheritance.

Trust Updates

Trust counteracts plutocracy by rewarding epistemic performance over capital. Without it, influence derives solely from stake size, enabling wealthy agents to dominate regardless of their truth-telling ability. Trust weights reliable agents more heavily, fostering a meritocracy of demonstrated competence rather than simple wealth accumulation.

The mechanism achieves this through emergent properties rather than arbitrary parameters. As agents gain trust and influence, expectations rise proportionally, making additional trust progressively harder to earn and easier to lose. This creates natural churn: consistent performers maintain their position, while underperformers lose influence to newcomers who exceed modest initial expectations. During high-uncertainty periods, the system amplifies the impact of performance differences, rewarding epistemic courage when truth-telling carries the greatest risk.

Trust evolves based on an agent's contribution relative to their weight-proportional expectation:

$$E_{i,b} = \frac{W_{i,b}}{\sum_j W_{j,b}}$$

Performance ratios ($P_{i,b}$) separate signal from noise, where $g_{i,b}$ represents agent i's information score for belief b (their epistemic weight multiplied by their BTS signal quality):

For signal providers ($g_{i,b} > 0$):

$$P_{i,b} = \frac{g_{i,b}/\sum_{g_j>0} g_j}{W_{i,b}/\sum_j W_{j,b}}$$

For noise contributors ($g_{i,b} < 0$):

$$P_{i,b} = \frac{g_{i,b}/\sum_{g_j<0} |g_j|}{W_{i,b}/\sum_j W_{j,b}}$$

This separation ensures signal providers compete only against other signal providers, while noise is penalized relative to other noise contributors. $P > 1$ indicates outperformance; $P < 1$ reflects underperformance.

Cumulative memory ($M_{i,b}$) tracks performance:

$$M_{i,b}^{\text{new}} = M_{i,b}^{\text{old}} + (1 - c_{t,b}) \cdot (P_{i,b} - 1)$$

Here, $(1 - c_{t,b})$ modulates sensitivity to performance based on network certainty ($c_{t,b}$). During uncertain periods ($c \approx 0$), performance differences have larger effects, encouraging epistemic courage.

Trust is calculated as:

$$T_{i,b} = 1.0 + \max(0, M_{i,b})$$

It has no upper bound but a floor of 1.0. Combined with square root scaling ($W = S \times \sqrt{T}$), this creates diminishing returns: $T = 4$ yields 2× influence; $T = 16$ yields 4×. As trust grows, so do expectations, forming a self-regulating equilibrium where sustained excellence is required to maintain high influence.

Finally, the commit-reveal process preserves BTS's truth-telling guarantees. Agents cannot observe others' beliefs during submission, preventing herding and ensuring even highly trusted agents cannot manipulate outcomes through strategic reporting.

3.8 The Temporal Dissent Goldilocks Zone

Introducing time into Bayesian Truth Serum creates a nuanced emergent phenomenon: the Temporal Dissent Goldilocks Zone. This represents the optimal timing for rational dissent in an asynchronous information environment, maximizing long-term returns by balancing early signaling with network recognition.

Information propagates unevenly across the network. Agents who correctly identify emerging truths face a critical timing tradeoff:

- **Too Early**: Dissent is penalized initially because the information hasn't yet propagated. The agent appears to contribute noise until the network catches up.
- **Too Late**: The agent becomes a follower, capturing diminishing returns as consensus has already formed.
- **Just Right**: The agent's dissent is timed optimally: early enough to provide valuable signal, but late enough for the network to recognize and reward it.

This dynamic gives rise to a natural market for temporal arbitrage. The protocol prices timing implicitly: as information diffuses, passive agents converge toward emerging consensus, retroactively validating early dissent. Correct early signals are rewarded not only by recognition, but through compounding trust and stake as the network's epistemic state matures.

3.9 Belief Creation as Epistemic Commitment

Proposing a new belief requires creators to commit capital through two mechanisms: a creation fee and initial liquidity provision. The fee acts as Sybil resistance, deterring spam by making low-quality propositions costly. The liquidity requirement ensures creators signal genuine confidence in the belief's value.

This structure frames belief creation as an epistemic act: creators must assess whether their proposition merits collective attention. Veritas thus operates on two interconnected markets: the primary market, where agents discover truth; and a meta-market, where creators compete to pose questions worthy of inquiry.

3.9.1 Structural Parameters and Market Integrity

Beliefs are minimal on-chain constructs containing only essential market data: agent positions, stakes, trust scores, and aggregate belief distributions. All contextual content remains off-chain.

At creation, belief creators define immutable structural parameters that shape the belief's operational form:

- **Outcome Space**: Discrete categories or continuous domains
- **Economic Model**: Fixed bounty, continuous market, or hybrid systems, where bounty markets also require duration
- **Quality Incentives**: Creator's share of value generated, aligning curation incentives with protocol goals

These parameters are fixed upon instantiation, ensuring participants can trust the underlying market rules when allocating stake.

3.9.2 Adaptive Governance via Meta-Belief Feedback

Each belief spawns associated meta-beliefs that govern its operational parameters. These meta-beliefs use the same truth-discovery process as primary propositions, enabling participants to calibrate market settings such as epoch duration and belief revenue structure through informed consensus.

Meta-beliefs allow local adaptation without imposing rigid global standards. Fast-moving topics can shorten cycles; complex ones can extend deliberation.

4 Delegation Mechanics

Delegation enhances efficiency by separating capital from epistemic labor, allowing agents to specialize in knowledge production while enabling capital holders to participate without domain expertise. Stakeholders can delegate their epistemic weight to specialized truth discoverers, retaining full economic exposure: the delegator absorbs both rewards and penalties tied to their stake.

This creates a dynamic market where experts compete for delegated stake by demonstrating consistent informational value, and capital flows adaptively toward those driving convergence. Delegation is partial, time-bounded, and revocable, ensuring flexibility and preventing permanent power concentration. The key dynamic is that delegates gain influence through epistemic judgments, while delegators retain economic risk, aligning incentives to reward both knowledge and risk-taking.

Delegates may charge a percentage fee on any rewards earned through their decisions, providing direct compensation for expertise without exposing themselves to downside risk. This arrangement transforms Veritas into a two-sided marketplace: capital seeks returns through backing skilled truth discoverers, while experts compete on performance and fee structure.

5 Self-Referential Governance

Veritas applies truth discovery to its own governance. Rather than fixing parameters or relying on external voting, the protocol uses meta-beliefs: propositions about its own design (e.g., "optimal creation fee is X") to determine its configuration.

Meta-beliefs function like any other belief, using stake, mirror descent, and BTS scoring to converge on values that best serve the protocol.

Protocol-level meta-beliefs include:

- Belief creation fee
- DAO revenue percentage
- Revenue distribution frequency
- Maximum stake allocation per belief

Participation requires VER tokens, aligning governance power with long-term stake.

VERITAS thus becomes a self-organizing epistemic system, discovering not only external truths, but its own optimal form through the same process.

6 Conclusion

In an era where misinformation spreads faster than truth, Veritas proposes a fundamentally different approach: harnessing human disagreement rather than suppressing it. By aligning economic incentives with epistemic accuracy and enabling beliefs to evolve through competitive revelation, the protocol transforms the complexity of diverse opinions into a structured process of knowledge discovery.

Truth emerges through the natural selection of ideas under economic pressure. Formal convergence proofs remain future work.

Acknowledgments

Thanks to Adam Shen and Aryan Sheikhalian for valuable contributions to the development of this protocol.

References

[1] Alvin I. Goldman. Foundations of social epistemics. *Synthese*, 73(1):109–144, 1987. Early foundational paper on social epistemology.

[2] Alvin I. Goldman. *Knowledge in a Social World*. Oxford University Press, Oxford, 1999. Foundational work on social epistemology and veritistic approach to collective knowledge.

[3] Friedrich A. Hayek. The use of knowledge in society. *The American Economic Review*, 35(4):519–530, 1945. Seminal paper on information aggregation in markets and decentralized knowledge.

[4] Philip Kitcher. The division of cognitive labor. *The Journal of Philosophy*, 87(1):5–22, 1990. Seminal work on cognitive labor division in scientific communities.

[5] Plato. The apology of socrates, c. 399 BCE. Classic philosophical work containing the famous quote about the price of apathy.

[6] Dražen Prelec. A bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004. Original paper introducing Bayesian Truth Serum methodology for eliciting truthful subjective opinions.

[7] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. Foundational paper establishing information theory and introducing concepts of entropy and information.

[8] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, 1948. Second part of Shannon's foundational information theory paper.