`Attestia`: A Decentralized Protocol for Verifiable Media Authenticity

`Attestia`

April 1, 2026

, ,

Contents

Contents 2
Abstract 3
1 Problem Statement 3
1.1 The Crisis of Digital Authenticity 3
1.2 Limitations of Centralized Verification 3
1.3 The Limitations of Purely Algorithmic Detection 3
1.4 The Need for a Decentralized Verification Layer 4
1.5 Toward Trustless Media Verification 4
2 Core Concepts 4
2.1 Attestations 4
2.2 Participants 4
2.3 Verification as a Market 5
3 The `Attestia` Workflow 5
3.1 Submission and On-Chain Attestation 5
3.2 Verification Phase 5
3.3 Confidential Off-Chain Attestations 6
3.4 Aggregation and Verifiable Computation 6
3.5 On-Chain Settlement and Incentives 6
4 Cryptoeconomic Design 6
4.1 Staking Model 6
4.2 Reward Mechanism 7
4.3 Reputation System 7
4.4 Slashing Mechanism 7
4.5 Network Phases 7
4.6 Fee Distribution 8
5 Token Design 8
5.1 Token Utility 8
5.2 Reward Denomination Model 8
5.3 Staking Transition Model 9
5.4 Emission Model 9
6 Use Cases 9
6.1 Media & Journalism 9
6.2 Social Platforms 10
6.3 Financial / Legal Evidence 10
7 Conclusion 10
References 10

Abstract

The rapid advancement of artificial intelligence has transformed how digital content is created and distributed—but at the cost of trust. AI-generated media, including deepfakes, is becoming indistinguishable from reality, undermining confidence in journalism, social platforms, and online information.

`Attestia` is a decentralized protocol designed to restore trust in digital media through verifiable authenticity. Instead of relying on centralized authorities, it introduces an open verification network where independent participants assess content collaboratively. Submitters stake to request verification, while verifiers—running their own models—produce authenticity scores that are aggregated into a consensus and recorded as on-chain attestations. The protocol aligns incentives through cryptoeconomics: verifiers are rewarded when their evaluations align with consensus and penalized when they deviate. This creates a system where participation and accuracy are economically incentivized and low-quality or malicious behavior is discouraged. By combining off-chain computation with on-chain attestations, `Attestia` ensures scalability, privacy, and verifiability. By turning verification into a decentralized market, `Attestia` establishes a new trust layer for the internet—where authenticity is not assumed, but proven.

Keywords

Blockchain, Web3, Deepfake, AI

1 Problem Statement

1.1 The Crisis of Digital Authenticity

The rapid proliferation of artificial intelligence has fundamentally altered the nature of digital content creation. Advanced generative models can now produce highly realistic images, videos, and audio that are often indistinguishable from authentic media. While these technologies unlock new creative possibilities, they simultaneously lead people scrolling their feeds to always ask themselves: "*Am I looking at something real?*" [5].

Deepfakes and synthetic media are no longer confined to research environments or niche applications. They are increasingly accessible, scalable, and deployable at low cost [9]. As a result, misinformation can be generated and disseminated at unprecedented speed and scale, impacting journalism, public discourse, financial markets, and institutional trust [7]. In such an environment, the absence of reliable mechanisms to verify authenticity poses systemic risks to both digital ecosystems and society at large.

1.2 Limitations of Centralized Verification

Existing approaches to content verification rely heavily on centralized entities, such as fact-checking organizations, social media platforms, or specialized service providers (e.g., *Sensity AI*, *Hive AI*). In practice, these systems are predominantly based on proprietary automated models, with little to no human involvement in the verification loop. While these actors play an important role, their approaches exhibit several structural limitations [3],[6].

First, *verification processes are inherently opaque.* The underlying models, training data, and decision criteria are typically not disclosed, making it difficult for users to understand or independently validate how conclusions are reached.

Second, *algorithmic detection is not infallible.* Deepfake generation techniques evolve rapidly, and there exist multiple strategies to bypass detection models, including adversarial perturbations, compression artifacts, and distribution shifts. As a result, even state-of-the-art systems can produce unreliable or inconsistent outputs.

Third, *centralization introduces single points of failure.* A single provider ultimately determines the verification outcome, exposing the system to risks of bias, manipulation, censorship, or operational failure.

Finally, *verification outputs are not composable.* Results are typically confined within platform-specific infrastructures and cannot be easily reused, audited, or aggregated across different applications.

1.3 The Limitations of Purely Algorithmic Detection

Automated detection systems based on machine learning have emerged as a promising approach for identifying manipulated content at scale. These systems are capable of processing large volumes of media efficiently and detecting subtle statistical artifacts that may not be visible to human observers. As such, they represent a critical component in addressing the challenges posed by synthetic media [4].

However, relying exclusively on algorithmic detection introduces fundamental limitations. Many detection models operate as black boxes, offering limited transparency into how conclusions are reached and making their outputs difficult to interpret or audit. At the same time, the rapid evolution of generative models creates a persistent adversarial dynamic, where improvements in detection are continuously matched by more sophisticated forms of content synthesis. In addition, purely algorithmic approaches often struggle to capture contextual and semantic aspects of authenticity, such as intent, narrative coherence, or domain-specific knowledge, which are essential in real-world verification scenarios. Finally, and most critically, the outputs of these systems are not inherently verifiable: users must ultimately trust the entity operating the model, rather than being able

to independently validate the result [8]. An additional, often underestimated limitation lies in the data demands required to train effective detection models. Machine learning systems only achieve strong generalization when they are trained on large and diverse datasets. To reliably differentiate between authentic and synthetic media across many contexts, a model must be exposed to a wide range of examples during training. This diversity is not simply beneficial, but it is essential for robustness. Without sufficient coverage of rare, unusual, or out-of-distribution cases, even advanced detectors may fail, generating highly confident yet incorrect predictions when faced with unfamiliar inputs. Building and maintaining such comprehensive datasets is both expensive and complex to manage, while also introducing challenges related to annotation quality and potential bias. As a result, no single model or organization can realistically guarantee comprehensive and fully reliable detection across all types of content and domains.

While algorithmic detection is essential for scalability, it is insufficient on its own to establish trust in high-stakes environments. A robust solution must integrate automated analysis within a broader framework that enables complementary forms of evaluation and verifiable outcomes.

1.4 The Need for a Decentralized Verification Layer

The challenges outlined above point to the need for a new paradigm: a system where verification is not controlled by a single authority, but emerges from the coordinated input of multiple independent participants. Such a system should satisfy the following properties:

- **Decentralization:** No single entity should control the verification process.
- **Transparency:** Verification outcomes should be publicly auditable.
- **Incentive alignment:** Participants should be economically motivated to provide accurate evaluations.
- **Verifiability:** Results should be cryptographically provable and independently checkable.
- **Scalability:** The system should support large and diverse volumes of content and participants.

1.5 Toward Trustless Media Verification

`Attestia` is motivated by the need to establish a trustless, programmable layer for media authenticity. Instead of asking users to trust platforms, institutions, or proprietary models, the goal is to enable a system where authenticity is derived from verifiable attestations produced by a decentralized network. By transforming verification into an open and incentive-driven process, `Attestia` addresses the fundamental limitations of existing approaches and lays the groundwork for a more resilient and trustworthy information ecosystem.

2 Core Concepts

2.1 Attestations

At the foundation of `Attestia` lies the concept of an *attestation*: a verifiable and auditable statement made by a participant about a specific piece of content. In the context of media authenticity, an attestation represents an evaluation of whether a given image, video, or other digital asset is authentic, manipulated, or synthetic with a certain confidence level. Attestations are cryptographically signed and linked to both the issuer and the subject of the evaluation. This ensures that each contribution is attributable, traceable, and cannot be altered after creation. Rather than producing a single authoritative verdict, the protocol aggregates multiple attestations, allowing authenticity to emerge as a consensus among independent evaluations.

To balance transparency, scalability, and privacy, `Attestia` adopts a hybrid approach to attestation storage. Individual evaluations are generated and stored off-chain, enabling efficient handling of large data and flexible access control. At the same time, aggregated results and their corresponding commitments are anchored on-chain, providing an immutable and publicly verifiable record. This design allows participants to independently verify outcomes without exposing sensitive intermediate data.

2.2 Participants

`Attestia` is built around a decentralized network of participants who contribute to the verification process through distinct but complementary roles.

Contributors are responsible for introducing content into the protocol. These participants, which may include media organizations, platforms, or individual users, stake tokens to participate to the network and submit media for verification. The staking requirement acts as a mechanism to discourage spam and ensure that submitted content has economic weight within the system.

Verifiers are the core evaluators of the network. Each verifier independently analyzes submitted content and produces an authenticity score or assessment, bringing to the network a diverse set of capabilities — ranging from advanced machine learning models to domain-specific human expertise and contextual reasoning. These different approaches are complementary and can collaborate to improve overall accuracy. Regardless of their nature, all verifiers are required to stake tokens as a guarantee of honest behavior, operate under the same protocol rules, and are evaluated based on the quality and consistency of their contributions. Verifiers whose assessments align with the network consensus are rewarded, while those who deviate significantly or behave maliciously are penalized through a slashing mechanism, losing a portion of their staked tokens.

In addition to contributors and verifiers, the protocol can be consumed by external entities. Applications, platforms, and third-party services may query `Attestia` to retrieve

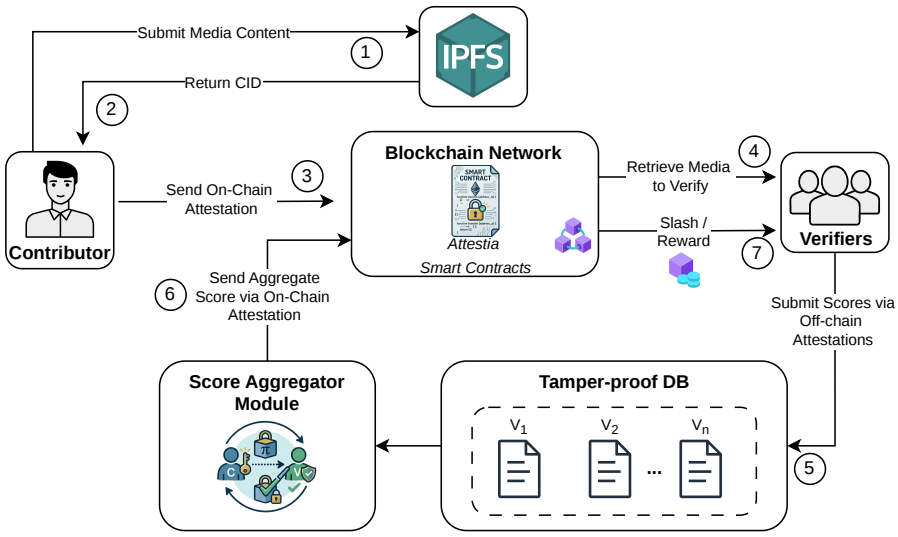

Figure 1: High-level architecture of `Attestia`

authenticity attestations, integrating verifiable trust signals directly into their products and workflows.

2.3 Verification as a Market

`Attestia` models verification as an open and competitive market, where participants are incentivized to provide accurate assessments of content authenticity. Instead of relying on a predefined authority or a single detection system, the protocol enables multiple independent verifiers to contribute their evaluations in parallel.

In this setting, different approaches to verification—ranging from advanced machine learning models to domain-specific human expertise—can coexist and compete. No single participant defines the ground truth; rather, authenticity emerges from the aggregation of diverse perspectives. This creates a dynamic environment in which better-performing verifiers are naturally rewarded, while less reliable participants are penalized.

The market-based design fosters continuous improvement. As new detection techniques are developed and new participants enter the network, the overall quality of verification evolves over time. By aligning economic incentives with accuracy, `Attestia` transforms verification from a static process into a self-improving system, capable of adapting to the rapidly changing landscape of synthetic media.

3 The `Attestia` Workflow

The Attestia protocol structures media verification as a time-bounded, multi-party process composed of sequential phases: submission, evaluation, aggregation, and settlement. This design enables open participation while preserving both confidentiality and verifiability. The protocol is built upon the VeriNet framework, a blockchain-based architecture for decentralized content verification originally proposed and peer-reviewed in [2], which demonstrates the viability of distributed verification systems across multiple domains, including deepfake detection. Figure 1 illustrates the overall architecture and the flow between these layers.

3.1 Submission and On-Chain Attestation

The process is initiated by a *Contributor*, who submits a media item for verification. The media itself is not stored on-chain; instead, it is uploaded to a content-addressable storage system such as *IPFS*, ensuring both integrity and efficient distribution ①–②. An on-chain attestation is then created via the *Ethereum Attestation Service* (EAS) [1] ③. This attestation encapsulates three elements: a cryptographic reference to the media (i.e., its IPFS hash), a minimal set of contextual metadata, and an expiration time defining the temporal boundary within which all verifier evaluations must be submitted. By default, this expiration time is set to 12 hours. The inclusion of an explicit expiration time is not merely operational—it defines the coordination window within which all evaluations must occur, effectively synchronizing participants under a shared constraint.

Once published, this attestation becomes the canonical entry point for the verification task. It is publicly discoverable, immutable, and uniquely identifiable, allowing downstream attestations to reference it unambiguously.

3.2 Verification Phase

Verifiers continuously monitor the set of active submissions, either through protocol-native interfaces or external integrations ④. Each submission remains open until its

expiration time, during which any eligible Verifier may participate.

Evaluation is intentionally left unconstrained at the methodological level. Verifiers may rely on machine learning models, heuristic analysis, or domain-specific expertise. What the protocol enforces is not *how* the evaluation is performed, but *when* and *under which information conditions* it is submitted. By requiring all scores to be provided within the predefined time window, `Attestia` reduces the risk of reactive or imitative behavior, preserving a degree of independence across contributions.

3.3 Confidential Off-Chain Attestations

Each evaluation is encoded as an off-chain attestation that references the original submission. These attestations are stored in a tamper-resistant data layer, where every record is bound to a cryptographic commitment ⑤. The decision to keep verifier outputs off-chain is central to the protocol's design. Publicly revealing intermediate scores would introduce undesirable dynamics—most notably herding effects, where later participants condition their outputs on earlier ones. By maintaining confidentiality until the aggregation phase, the system encourages genuine, independent assessments. At the same time, integrity is not sacrificed: each off-chain attestation is anchored on-chain via a timestamp, ensuring that once a score is submitted, its existence and timing are permanently recorded on the blockchain and cannot be altered without detection. This design also accommodates a heterogeneous set of verifiers, including those operating proprietary models or handling sensitive signals, who may otherwise be unwilling to participate in a fully transparent environment.

3.4 Aggregation and Verifiable Computation

After the expiration time has elapsed, the system transitions from data collection to aggregation. An `Attestia` *Aggregator* module retrieves all valid verifier attestations and computes a consensus score according to a predefined rule, such as a reputation-weighted average:

$$\bar{x} = \frac{\sum_{m=1}^{N} \rho_m x_m}{\sum_{m=1}^{N} \rho_m} \tag{1}$$

where $x_m \in [0, 1]$ is the authenticity score submitted by verifier m, N is the total number of participating verifiers, and ρ_m is a reputation score reflecting the verifier's historical reliability within the network — its computation is discussed in detail in Section 4.3.

While the aggregation function itself is straightforward, the challenge lies in making its execution trustless. To this end, the computation is accompanied by a *Zero-Knowledge* (ZK) *proof* attesting that the result was derived from the committed dataset, that all eligible inputs were considered, and that the specified function was applied correctly. Crucially, this proof reveals nothing about individual contributions, exposing only the final aggregate.

The outcome is therefore both *private* at the micro level and *verifiable* at the macro level—a property that is difficult to achieve in traditional verification systems.

3.5 On-Chain Settlement and Incentives

The final step consists in anchoring the result on-chain. The aggregated score, together with a commitment to the underlying proof, is recorded as an EAS attestation following a predefined schema ⑥. This attestation represents the protocol's public output: a compact, tamper-resistant statement about the authenticity of the submitted media.

Once the result is finalized, the protocol evaluates the behavior of each Verifier ⑦. Rather than rewarding absolute correctness—which is generally unknowable—the system measures alignment with the collective outcome. Verifiers whose scores fall within an acceptable range of the aggregate are rewarded, while those that deviate significantly may incur penalties through a slashing mechanism. Over time, this feedback loop induces a form of emergent calibration: participants are incentivized to produce estimates that are both independent and statistically consistent with the broader set of evaluations. In this sense, consensus is not imposed ex ante, but arises ex post as the equilibrium of a cryptoeconomic process.

4 Cryptoeconomic Design

The security of `Attestia` emerges from a cryptoeconomic system that aligns incentives between contributors and verifiers while adapting to the maturity of the network. The protocol is designed to balance three competing objectives: preventing spam and manipulation, incentivizing accurate verification, and enabling early-stage growth without penalizing honest participants. To achieve this, `Attestia` introduces differentiated staking requirements, phase-dependent rewards, confidence-aware slashing, and a reputation system that reinforces long-term reliability. In addition, the design ensures that the expected loss from dishonest behavior significantly exceeds the expected gain from a single verification task, making manipulation economically irrational.

4.1 Staking Model

`Attestia` distinguishes between two economically distinct roles: contributors and verifiers.

Contributors stake tokens when requesting verification of a media item. This stake serves primarily as an anti-spam mechanism, ensuring that each submission carries a non-negligible cost. At initialization, the protocol adopts a default contributor stake of

$$s_c = 0.01 \text{ ETH}. \tag{2}$$

The protocol fee retained from this stake varies according to the type of submitted content, reflecting the different computational demands imposed on the verifier network. For images and audio, the retained fee is 15% of s_c, while for video content, which requires significantly greater processing

resources from verifiers, it is set at 20%. The remaining portion — 85% for images and audio, 80% for video — is returned to the contributor after the verification process completes. This design discourages low-quality or adversarial submissions without excessively penalizing legitimate users, while ensuring that the reward pool scales proportionally with the computational cost of each verification task.

Verifiers stake tokens to participate in the evaluation process. Unlike submitters, their stake acts as active collateral: it grants access to rewards but is also subject to slashing. Since verifiers directly determine the credibility of the protocol, their economic exposure is significantly higher. At initialization, the protocol adopts a default verifier stake of:

$$s_v = 0.10 \text{ ETH}. \tag{3}$$

To maintain a consistent level of economic security, both stakes are defined in ETH-equivalent terms and dynamically adjusted using price oracles. This prevents fluctuations in the native token price from weakening the protocol's security guarantees and ensures that the economic exposure of each participant remains meaningful regardless of market conditions.

4.2 Reward Mechanism

Rewards are composed of a stable component denominated in ETH and a variable component denominated in the native token (introduced in Section 5). Each verifier m submits a score x_m for a given content item. The deviation from consensus is defined as:

$$d_m = |x_m - \bar{x}|. \tag{4}$$

Rewards are composed of two components: alignment with consensus and contribution to the informativeness of the aggregate outcome. The reward function is defined as:

$$r_m = R_\phi \cdot \left(w_1 e^{-\alpha d_m} + w_2 I_m \right) \cdot \rho_m, \tag{5}$$

where R_ϕ is the phase-dependent reward pool, α controls sensitivity to deviation, ρ_m is the verifier's reputation multiplier, and I_m is an influence factor measuring the marginal contribution of verifier m to the aggregate result (e.g., the change in $\bar{x}$ when excluding m). The weights satisfy $w_1 + w_2 = 1$.

This formulation ensures that verifiers are incentivized not only to align with consensus, but also to provide independent and informative evaluations. Small deviations result in marginal reward reductions, while large deviations are penalized exponentially. At the same time, verifiers whose contributions meaningfully affect the outcome are rewarded, discouraging low-effort or purely imitative strategies.

4.3 Reputation System

Each verifier maintains a reputation score ρ_m reflecting their historical performance. Reputation evolves over time as an exponentially weighted average of past alignment scores:

$$\rho_m^{(t)} = \lambda \rho_m^{(t-1)} + (1 - \lambda) e^{-\alpha d_m}, \tag{6}$$

with $\lambda = 0.8$.

This formulation gives more weight to recent performance while still retaining memory of past behavior. Reputation is bounded within:

$$\rho_m \in [0.5, 1.5]. \tag{7}$$

Reputation acts as a soft multiplier on rewards without affecting the aggregation mechanism, ensuring that high-performing verifiers are incentivized while preserving openness and preventing centralization of influence.

4.4 Slashing Mechanism

Slashing is designed to penalize clear outliers while preserving tolerance for honest disagreement. The key quantity governing slashing is the deviation value d_m, which measures how far a verifier's score departs from the aggregate defined in Section 4.2.

However, a large deviation alone is not sufficient to trigger a penalty. A verifier may legitimately disagree with the majority in situations where the network is uncertain—for example, when submitted scores are widely dispersed. In such cases, high variance among verifier scores indicates that the content may be ambiguous or that consensus has not yet emerged. Penalizing deviations under these conditions would be economically unjustified and could discourage participation on genuinely difficult evaluations.

For this reason, slashing is additionally conditioned on the variance of submitted scores, denoted σ^2. Low variance signals that the verifier set has reached a strong consensus, making a large individual deviation less defensible as honest disagreement. Accordingly, slashing is triggered only when both of the following conditions are satisfied:

$$d_m > \delta \quad \text{and} \quad \sigma^2 < \tau \tag{8}$$

where δ is the deviation threshold and τ is the variance ceiling above which slashing is suppressed. When both conditions are met, a fraction of the verifier's stake is forfeited.

The precise values of δ, τ, and the penalty fraction are phase-dependent and specified in Section 4.5.

4.5 Network Phases

The incentive structure of `Attestia` evolves as the verifier set grows. In early stages, consensus is fragile and penalties must be conservative. As the network matures, stronger economic guarantees can be enforced. Let N be the number of active verifiers. The protocol operates under three phases.

Phase 0: Bootstrapping ($N < 5$). At this stage, the verifier set is too small to produce a reliable consensus. Disagreement between verifiers is expected and does not necessarily indicate malicious behavior. Slashing is disabled in this phase. Verifiers receive rewards at 50% of the base reward, reflecting the lower confidence in the aggregate outcome. Reputation updates remain active, allowing the protocol to start building a performance history. The primary objective of this phase is to maximize participation and bootstrap the verifier set.

Phase 1: Weak Consensus ($5 \leq N \leq 20$). The network begins to produce meaningful signals, but uncertainty remains. Slashing is enabled selectively: a penalty equal to 10% of the verifier stake is applied only when $d_m > 0.25$ and $\sigma^2 < 0.015$. This joint condition ensures that penalties are triggered only when the network has converged with sufficient confidence. Rewards are distributed at 80% of the base rate.

Phase 2: Mature Network ($N > 20$). The aggregated outcome is robust, and the variance condition is lifted. Slashing is applied whenever $d_m > 0.20$, with a penalty equal to 25% of the verifier stake. Rewards are distributed at full rate.

4.6 Fee Distribution

The protocol incorporates a structured fee distribution mechanism to ensure sustainable verifier incentives and long-term protocol development.

For each submission, the protocol retains a fee of 15% of the contributor stake s_c for images and audio, and 20% for video content. This fee is distributed across two components:

- **Verifier Rewards ($\approx$ 85%):** The large majority of the fee is allocated to the reward pool for verifiers participating in the task. This ensures that verification activity is economically incentivized even in low-demand conditions and complements token emissions during early network phases.
- **Protocol Treasury ($\approx$ 15%):** A portion of the fee is directed to a treasury controlled by the protocol. These funds support ongoing development, infrastructure, research, and ecosystem expansion. Over time, this treasury may also be governed by token holders.

Let f_v and f_t denote the fractions allocated to verifiers and treasury respectively, such that:

$$f_v + f_t = 1. \tag{9}$$

At initialization, the protocol adopts:

$$f_v = 0.85, \quad f_t = 0.15. \tag{10}$$

This distribution ensures that the vast majority of economic value flows directly to active verification, while simultaneously funding protocol sustainability. As the network matures, these parameters may be adjusted through governance to reflect changing economic conditions and usage patterns.

5 Token Design

5.1 Token Utility

The native token of the protocol, **ATTA**, serves as the economic coordination layer of Attestia. Its role is not to act as a general-purpose currency, but to secure the behavior of verifiers and enable trustless coordination within the network. The protocol distinguishes between two categories of participants with different relationships to ATTA.

Contributors and consumers of attestations — media organizations, platforms, and downstream applications — interact with the protocol exclusively through standard payment rails, using ETH. They are not required to acquire or manage ATTA at any point. From their perspective, the token is entirely abstracted from the user experience.

Verifiers, by contrast, are the primary economic actors within the ATTA ecosystem. Their relationship with the token evolves across two distinct stages of the protocol, as described in Section 5.3. In the mature network, verifiers are required to lock ATTA as collateral in order to participate in the evaluation process. This requirement creates structural demand for the token and aligns the economic interests of verifiers with the long-term health of the network: a verifier who holds staked ATTA has a direct financial stake in the credibility of the protocol.

At the protocol level, ATTA fulfills two concrete functions. First, it enables stake-backed participation in the mature network: once the token has established a stable market price, verifiers are required to lock ATTA as collateral. Second, it serves as a long-term incentive mechanism: verifiers accumulate ATTA as the variable component of their reward from the earliest stages of the network, aligning their interests with the growth of the protocol before the staking transition occurs.

5.2 Reward Denomination Model

A critical design decision in any token-based incentive system concerns the denomination of rewards: whether participants are compensated in the native token, in stable assets, or in a combination of both.

Compensating verifiers exclusively in ATTA would tie their real income entirely to the market price of the token. In practice, an exclusively token-denominated reward structure would deter technically capable verifiers and attract primarily speculative participants, undermining the credibility of the consensus mechanism.

For this reason, Attestia adopts a mixed denomination model in which verifier rewards are composed of two components. The first is a stable component, denominated in ETH, derived directly from the fee paid by contributors. This component guarantees a minimum level of real economic return for verifiers regardless of token price conditions, making participation viable for operators with concrete infrastructure costs. The second is a variable component, denominated in ATTA and funded through protocol emissions. This component aligns the long-term interests of verifiers with the success of the protocol: as ATTA appreciates with adoption, so does the value of their accumulated rewards.

The proportion between the two components is not fixed and evolves in a way that reflects the maturity of both the protocol and the token. In the initial phase, the stable component is dominant: verifiers must be able to cover real operational costs today, and cannot rely on an asset that has not yet established a market price. The variable ATTA component is present but in a minority, serving as a long-term incentive for verifiers who believe in the protocol and are willing to accumulate the token ahead of its market formation. As the protocol enters the

mature phase, emissions decrease and ATTA has established a stable market price. The stable component remains significant, sustained by the growth of fee income, while the ATTA component becomes smaller in quantity but not necessarily in value — since each unit of ATTA is now priced by an active market.

5.3 Staking Transition Model

A fundamental challenge in launching a token-based protocol is the cold-start problem for the token itself: verifiers cannot be expected to stake an asset that has no established market value, yet the token cannot acquire market value without an active verifier set that creates demand for it. Requiring ATTA staking from day one would make it impossible to attract the first participants, regardless of the protocol's technical merits. To address this, `Attestia` adopts a two-phase staking model that separates the bootstrapping of the verifier network from the bootstrapping of the token.

In the initial phase, verifiers stake ETH as collateral. This allows participation to begin immediately, without any dependency on ATTA's market price. ETH is a liquid, widely held asset with well-understood value, removing the barrier to entry for technically capable verifiers who would otherwise be unwilling to acquire and lock an illiquid or unpriced token. During this phase, verifiers are compensated through the mixed denomination model described in Section 5.2: a stable component in ETH derived from contributor fees, and a variable component in ATTA distributed through protocol emissions. Over time, verifiers accumulate ATTA as part of their rewards without being required to purchase it on the open market.

The transition to the mature phase occurs once ATTA has established a stable and liquid market price. At this point, the staking requirement migrates from ETH to ATTA. Verifiers who have accumulated sufficient ATTA through participation can transition their stake without any additional capital outlay, while new entrants must acquire ATTA on the open market. This migration creates the structural demand for the token that underpins its long-term value: every new verifier entering the network represents incremental buy pressure on ATTA, aligning token appreciation with protocol growth.

5.4 Emission Model

The total supply of ATTA is fixed at launch, establishing a hard cap that provides resistance to long-term inflation and gives participants a well-defined dilution horizon. Within this fixed supply, emissions follow an *exponentially decaying schedule*, rather than a constant issuance rate, ensuring that inflation is front-loaded and progressively reduced over time.

Formally, the emission rate is defined as a function of network progression:

$$E(N) = E_0 \cdot e^{-\lambda N}$$

where E_0 is the initial emission rate, $\lambda > 0$ is the decay constant, and N denotes the network phase as defined in Section 4.5. This formulation ensures a smooth and continuous reduction in emissions as the protocol matures.

During the bootstrapping phase ($N < 5$), the emission rate is at or near its maximum. Fee income is negligible and the verifier set is still forming, so the protocol relies primarily on token issuance to incentivize participation. Emissions in this phase are directed exclusively to verifiers, rewarding early involvement and enabling the accumulation of reputation history before organic demand emerges.

As the network enters the expansion phase ($5 \leq N \leq 20$), emissions decline according to the exponential schedule. At the same time, fee income begins to contribute meaningfully to the reward pool, reducing the protocol's reliance on inflation. Token emissions continue to supplement rewards, but their relative importance decreases as network usage grows.

In the maturity phase ($N > 20$), the emission rate approaches a low asymptotic level. At this stage, the protocol is sustained predominantly by fee-driven rewards, while residual emissions provide a baseline incentive for verifiers and help maintain liquidity during periods of low activity, without introducing significant dilution.

To counterbalance new issuance across all phases, the protocol incorporates a burn mechanism: a fixed fraction of the fee paid by each contributor is permanently removed from circulation. This establishes a direct link between network activity and token scarcity—higher usage leads to higher burn rates. When demand is low, emissions dominate and net supply expands. As adoption increases, the burn rate progressively offsets issuance. Under sufficiently high demand, the burn rate may exceed emissions, rendering ATTA net deflationary. This dynamic anchors token value to actual protocol usage rather than purely speculative demand.

6 Use Cases

`Attestia` is designed as a general-purpose verification layer for digital content, enabling trust to emerge from decentralized evaluation rather than centralized authority. Its applicability spans multiple domains where authenticity, provenance, and credibility are critical. The following use cases illustrate how the protocol can be integrated across different contexts.

6.1 Media & Journalism

In journalism, the erosion of trust is no longer driven solely by misinformation, but by the rapid proliferation of AI-generated and AI-manipulated content. Advances in generative models have made it possible to produce highly realistic images, videos, and audio at near-zero cost, often indistinguishable from authentic media even under expert inspection. This introduces systemic risks: fabricated evidence can be used to manipulate public opinion, synthetic interviews can be presented as real, and authentic content can be dismissed as fake, giving rise to the so-called "liar's dividend." Traditional verification workflows rely on editorial processes that are opaque, slow, and difficult to scale, making them ill-suited for a landscape where content can be generated and disseminated globally within seconds.

Attestia introduces a verifiable authenticity layer for media organizations, enabling newsrooms to attach cryptographic attestations to images, videos, and reports. Independent verifiers — ranging from AI systems to specialized analysts — evaluate content and contribute authenticity scores that are aggregated into a public, tamper-resistant signal.

This allows publishers to embed verifiable trust directly into their content distribution pipelines, while readers and downstream platforms can independently validate the integrity of the information they consume. Rather than trusting a single newsroom, users rely on a transparent, multi-party verification process.

6.2 Social Platforms

Social platforms operate at a scale where manual moderation is fundamentally insufficient, particularly in the presence of AI-generated content that can be produced in massive volumes and tailored for virality. Synthetic media can be deployed to amplify disinformation campaigns, create coordinated narratives, or bypass traditional detection systems through rapid iteration. At the same time, centralized moderation systems introduce concerns around bias, censorship, and lack of accountability, especially when opaque algorithms determine the visibility or suppression of content.

Attestia enables platforms to externalize content verification into an open market of verifiers. Content can be programmatically routed through the protocol, generating authenticity attestations that inform ranking algorithms, visibility decisions, or warning labels. This architecture decouples content distribution from content verification. Platforms retain control over user experience, while verification becomes a composable, interoperable primitive. Importantly, the system introduces economic accountability: verifiers are incentivized to provide accurate assessments, and penalized for systematic misalignment.

As a result, moderation evolves from a purely reactive and policy-driven process into a hybrid system combining automated detection, decentralized evaluation, and market-based incentives, better suited to counter the scale and adaptability of AI-generated content.

6.3 Financial / Legal Evidence

In financial and legal contexts, the integrity of digital evidence is paramount. Documents, records, and multimedia artifacts often require verification processes that are both auditable and resistant to tampering.

Attestia provides a decentralized mechanism to certify the authenticity and integrity of such evidence. Multiple independent verifiers can assess documents or datasets, producing attestations that are aggregated into a verifiable outcome anchored on-chain. This approach enhances traditional verification workflows by introducing transparency and redundancy. Rather than relying on a single certifying authority, trust is distributed across a network of evaluators, each economically incentivized to act honestly.

Potential applications include due diligence processes, forensic analysis of digital evidence, verification of financial disclosures, and certification of real-world assets. In all cases, Attestia enables a shift from trust-by-authority to trust-by-verification.

7 Conclusion

The rapid advancement of generative AI has fundamentally challenged the reliability of digital information, creating an urgent need for systems that can restore trust in online content. Attestia addresses this challenge by introducing a decentralized, verifiable, and incentive-aligned protocol for media authenticity. By combining independent evaluations, cryptographic attestations, and a carefully designed cryptoeconomic model, the protocol enables authenticity to emerge as a consensus rather than relying on centralized authorities or opaque algorithms. The integration of staking, reputation, and reward mechanisms ensures that accurate behavior is economically incentivized, while malicious or low-quality contributions are penalized.

As synthetic media continues to evolve, Attestia provides a flexible and extensible foundation for trustless verification, enabling applications across journalism, social platforms, and high-stakes digital evidence. In doing so, it contributes to the development of a more transparent, resilient, and trustworthy information ecosystem.

References

[1] [n. d.]. *Ethereum Attestation Service.* https://attest.org/

[2] Luigi Coppolino, Giovanni Maria Cristiano, Salvatore D'Antonio, Jonah Giglio, Giovanni Mazzeo, and Luigi Romano. 2025. A Blockchain Solution for Decentralized Content Verification and its Application to Deepfake Detection and Fintech Credit Scoring. *Blockchain: Research and Applications* (2025), 100406.

[3] Álvaro Figueira and Luciana Oliveira. 2017. The current state of fake news: challenges and opportunities. *Procedia computer science* 121 (2017), 817–825.

[4] Arash Heidari, Nima Jafari Navimipour, Hasan Dag, and Mehmet Unal. 2024. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 14, 2 (2024), e1520.

[5] Bart Jacobs. 2024. The authenticity crisis. *Computer Law & Security Review* 53 (2024), 105962.

[6] Abdullah Ayub Khan, Asif Ali Laghari, Syed Azeem Inam, Sajid Ullah, Muhammad Shahzad, and Darakhshan Syed. 2025. A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. *Discover Computing* 28, 1 (2025), 48.

[7] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[8] Hanwei Qian, Lingling Xia, Ruihao Ge, Yiming Fan, Qun Wang, and Zhengjun Jing. 2025. From Black Boxes to Glass Boxes: Explainable AI for Trustworthy Deepfake Forensics. *Cryptography* 9, 4 (2025), 61.

[9] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).