Evidence for Trust: Verifiable Claims for Private AI Governance via Cryptographic Proofs

Anonymous Author(s)
Affiliation
Address
email

Abstract

This article integrates technical illustrations with governance analysis to examine the role of cryptographic primitives, particularly zero-knowledge proofs, in enabling independent and tamper-proof verification of corporate AI claims without the need to disclose proprietary information. We distinguish between the potential use of such proofs in contexts of *public governance*, where regulators and state institutions require mechanisms of oversight and compliance, and *private governance*, where industry stakeholders and users develop trust in AI systems through voluntary mechanisms. Particular attention is given to the latter, exploring how cryptographic proofs may strengthen private governance by shifting trust away from self-reporting of unverifiable commitments toward verifiable, tamper-resistant guarantees. Building on this analysis, we develop a conceptual framework for the design of cryptographic proofs that is tailored to governance-relevant needs, encompassing the consideration of AI components, intended audiences, and the timing of proof generation. More broadly, the article demonstrates how cryptographic proofs can reconfigure the foundations of trust in AI governance by offering pathways to reconcile proprietary constraints with demands for accountability and transparency.

1 Introduction

The artificial intelligence (AI) ecosystem has increasingly underscored the necessity of governance frameworks that align technological development with ethical and societal imperatives [1]. Foundational principles such as transparency, fairness, and accountability have been advanced by regulators and policymakers as mechanisms for mitigating systemic risks and cultivating public trust in AI applications. Notably, the global regulatory environment is shifting away from voluntary guidelines toward legally binding frameworks, with the European Union's AI Act at the forefront, mandating the explicit incorporation of principles of transparency, security, and fairness into the design and deployment of AI systems [2]. Despite these regulatory advances, persistent obstacles undermine the visibility and viability of effective implementation of regulatory mandates, particularly in proprietary "black box" AI systems. As research studies have consistently noted, black box audits alone are insufficient to secure accountability and trustworthiness in AI systems [3–5]. Public scrutiny is further curtailed by structural barriers, including API-only access, trade secrecy protections, and the high costs of external validation [3, 4, 6].

Concurrently, private governance has emerged as a complementary mode of oversight, with AI companies pledging commitments to fairness, security, and privacy [1].Yet, a central challenge lies in cultivating trust among users and regulators [7]. Corporate assurances of compliance are typically derived from internal self-assessments, raising concerns about credibility and effectiveness. Voluntary disclosure mechanisms such as Model Cards and Datasheets for Datasets offer incremental improvements in transparency, but ultimately remain contingent upon the trustworthiness of their

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

authors [8, 9]. Moreover, firms frequently resist independent third-party audits, citing intellectual property concerns, while external evaluations remain vulnerable to cyber risks [6]. Consequently, many ethical commitments risk becoming performative, serving as compliance signals or public relations strategies rather than verifiable practices [10].

Within this context, the notion of the *'verifiable claim'* has emerged as a tool for strengthening transparency and compliance in proprietary AI systems [11]. In this article, we examine how the integration of cryptographic proofs into verifiable claims can enable independently verifiable and tamper-resistant attestations of compliance that preserve proprietary information, while enhancing credibility with regulators, users, other companies in the value chain and the general public. Arguably, such mechanisms can create avenues for private governance to move beyond merely symbolic commitments, facilitating a transition toward technically grounded and enforceable forms of transparency and accountability. In particular, this article foregrounds the potential of verifiable claims and the cryptographic afrtifacts that sustain them to recalibrate the relationship between voluntary industry self-regulation and formal oversight, advancing a new paradigm of trustworthy AI governance.

2 Verifiable Claims, Research Gaps and Opportunity for Private Governance

From a policy standpoint, verifiable claims are understood as statements formulated with sufficient precision to be evaluated, challenged, and trusted by external stakeholders. As noted in [11], they are assertions for which supporting evidence and reasoning can be marshalled, whose falsifiability allows varying degrees of certainty depending on context. On the other hand, from a technical standpoint, verifiable claims denote structured, evidence-based assertions underpinned by formal mathematical or computational methods that can be independently tested, reproduced, or validated [12]. In this article, we adopt the latter interpretation, as verifiable claims operationalized through formal cryptographic proofs.

Such proofs aim to deliver concrete guarantees regarding critical aspects of AI systems, including fairness, privacy, provenance, security, and compliance [11, 13–16]. The growing field of cryptographic verification is therefore oriented around practical questions such as the ones described below [12, 17]:

- How can it be verified that a model was (or was not) trained on a specific dataset?
- How can it be verified that a dataset possesses certain properties, or excludes certain sensitive information?
- How can the implementation of specific safety or security measures be reliably verified?
- How can we prove that an AI model satisfies fairness constraints without revealing proprietary information?

Nevertheless, the ability to design cryptographic proofs for verifiable claims faces several well-recognized challenges. One is the translation gap: company declarations such as "our AI model is fair" are rarely expressed in forms that can be directly verified, particularly in proprietary black-box systems where critical information must remain confidential. Bridging this gap requires the creation of shared benchmarks, taxonomies, and formal specifications that map high-level ethical and policy concepts (e.g., fairness) onto concrete mathematical definitions with verifiable properties [18, 19]. A second challenge, adresssed in current research, is the enforcement gap, which emerges when claims of fairness, security, or privacy are made without corresponding evidentiary mechanisms capable of carrying legal weight [17].

As we show in this article, despite existing challenges, applying cryptographic proofs to verifiable claims offers considerable potential for strengthening private governance. By embedding formal guarantees into corporate disclosures, AI companies can provide evidence that is independently checkable and tamper-resistant, while at the same time safeguarding proprietary information. This transformation enhances the credibility of private governance commitments by grounding them in verifiable technical evidence that can be trusted by regulators, civil society, industry stakeholders, and end users. Ultimately, a significant impact of these mechanisms lies in reinforcing the trustworthiness of private governance itself, enabling it to function as a credible complement to formal oversight.

3 Trust, Transparency and Private Governance

Trust in technology has become one of the defining questions of the digital age. Unlike social trust, which rests on shared norms, institutions, and interpersonal relations, trust in technology is increasingly mediated by systems whose rules and logic are encoded in software. As Lessig (1999) famously argued, "code is law", the architecture of digital systems regulates behavior as effectively as legal frameworks, shaping what can and cannot be done [20]. In the case of AI, this means that questions of trust extend beyond human actors to the technical infrastructures themselves, where lines of code and learning algorithms effectively exercise governance. However, trust in technology cannot simply be equated with blind reliance. As Taddeo (2010) and Botsman (2017) note, technological trust is fragile, often arising from dependency rather than assurance, and may quickly erode in the absence of credible guarantees [21, 22]. One established way of generating trust is transparency—the ability of stakeholders to access, interpret, and evaluate information about how systems function. Transparency can foster accountability and help mitigate misplaced or blind trust, but on its own it is often insufficient, especially when dealing with complex models that remain opaque even when certain technically artifacts and documentations are disclosed [23]. This challenge is particularly pressing in the realm of private AI governance, where companies and developers increasingly act as de facto regulators by embedding their own values, norms, and enforcement mechanisms into AI systems [1]. In such contexts, governance is exercised not only through public institutions but also privately, shaped by the design choices and protocols of private actors, whose technical architectures condition how transparency, compliance, and trust can be realized. Building trust in AI therefore requires moving beyond surface-level transparency toward more granular forms of trust that correspond to specific components of the AI system and to the various entities involved, as discussed in Section 6. As such, establishing trust in AI requires building confidence across its foundational components, including the provenance and integrity of data, the fairness of its model, the transparency of algorithmic decision-making within its operational context, etc. [7]. Before exploring how cryptographic primitives can be applied to strengthen trust across different components of AI systems and among diverse stakeholders, we first outline the foundational principles of cryptography, with particular emphasis on zero-knowledge (ZK) proofs as a promising means of achieving verifiable transparency in AI without disclosing proprietary information.

4 Cryptography, Zero-Knowledge Proofs and Private AI Governance

Cryptographic techniques offer promising solutions to verification challenges in AI without requiring the disclosure of proprietary information. Core techniques include zero-knowledge proofs (zk-SNARKs, zk-STARKs), succinct non-interactive arguments, homomorphic encryption, secure multi-party computation (MPC), authenticated data structures such as Merkle trees and polynomial commitments, digital signatures, and tamper-evident logs. Recent surveys of ZK machine learning outline how these tools apply across the AI lifecycle [14, 24]. Furthermore, research advances show how cryptographic primitives can enable privacy-preserving attestation of key AI processes—including training, inference, and evaluation [14, 16, 25–27]. For fairness, zero-knowledge formulations show how to certify constraints on given datasets[12, 13].

Implementing such capabilities in practice requires careful selection among cryptographic proof systems, each with different performance and trust trade-offs. PLONK-style arithmetizations with KZG commitments provide compact proofs and fast verification under a structured reference string. AIR/FRI-based STARK systems avoid trusted setup and are better positioned for post-quantum resilience at the cost of larger proofs. Reported benchmarks indicate feasible proving and verification times, on the order of seconds to minutes, even for complex use cases [13, 27].

Beyond proof construction, regulatory scrutiny also demands reproducibility and transparency. To ensure reproducibility and evidentiary weight, implementers ought to report, for representative small, medium, and large cases, the constraint or trace size, proof size, prover and verifier runtime, peak memory, and exact hardware and software configuration. Where relevant, batching, recursion, or accumulation ought to be described together with their asymptotic and empirical effects.

Foundational lines in verifiable computing complement these approaches, including secure multi-party computation, zero-knowledge proofs, and trusted execution environments (TEE), and motivate two-step audit designs in which auditors certify a pipeline inside secure enclaves and users receive cryptographic certificates per output [12, 28–30]. To concretize some of these technical concepts in

a verification setting and to provide an illustrative governance example throughout the article, we now present a representative protocol scenario. It shows how cryptographic methods can verify claim compliance (e.g., regarding personal data) without disclosing training data, thereby establishing the principle of verifiable claims.

Consider a provider that has trained an AI system on a proprietary dataset X. An ethics organization suspect that X may contain prohibited personal identifiers. To demonstrate compliance while preserving confidentiality, the protocol proceeds as follows:

1. **Canonical preprocessing.** The parties define a preprocessing function f that normalizes the schema and the fields to be checked, and publish hashes of f and of the schema for integrity.
2. **Commitment generation.** The provider computes a commitment

$$D = \mathrm{Com}_{pp}(f(X); r)$$

   using public parameters pp and randomness r, ensuring hiding and binding.
3. **Predicate formalization.** Compliance is encoded as a machine-readable predicate C. Two common instantiations are:
   (a) Format-based screening: compiling a catalogue of prohibited patterns into deterministic finite automata and checking all tokens in $f(X)$ inside the circuit.
   (b) Register-based non-inclusion: proving $f(X) \cap S = \emptyset$ with zero-knowledge private set-intersection or accumulator-based non-membership proofs against an authoritative register S.
4. **Statement identifier.** The parties derive

$$sid = H(C \,\|\, H(f) \,\|\, H(\text{schema}) \,\|\, H(\text{columns}) \,\|\, catalog_version \,\|\, params),$$

   which binds the claim to the predicate, preprocessing, schema and column selection, the pattern-catalogue version, and protocol parameters, preventing reinterpretation or replay.
5. **Proof and verification.** The provider (e.g., an AI company) generates proof

$$\pi \leftarrow \mathrm{Prove}_{pp}(D, sid; X, r).$$

   The ethics organization (the verifier) can now verify

$$\mathrm{Verify}_{pp}(D, sid, \pi) \in \{\text{ok}, \text{fail}\}.$$

   The proof jointly establishes dataset linking, $\mathrm{Com}_{pp}(f(X); r) = D$, and predicate satisfaction, $C(f(X)) = \text{ok}$.

While the scenario presented is intentionally simplified, protocols of this kind are of particular significance for private governance, as they enable verifiable assurances without requiring disclosure of sensitive or proprietary information as we discuss in detail in the following section.

5 Zero-Knowledge Proofs: Trust without Proprietary Information Disclosure

One of the central dilemmas in AI governance lies in the tension between transparency requirements and the protection of proprietary information [23]. Regulators, auditors, and the public increasingly demand evidence that AI systems meet standards of fairness, security, privacy, safety, etc. Yet, companies often resist such scrutiny on the grounds that full disclosure could compromise trade secrets, intellectual property, or privacy of datasets. This transparency challenge is especially acute in less regulated countries, where compliance is typically voluntary, and incentives to disclose may be weak. Zero-Knowledge Proofs (ZKPs) offer a powerful cryptographic technique to mitigate this tension. As described in the example above, a ZKP can enable a party (the prover) to demonstrate to another entity (the verifier) that a certain statement is true, without revealing any additional information beyond the validity of the statement itself. In the context of AI, this allows a company to prove compliance with specific requirements (e.g., that a model satisfies fairness criteria, was trained on publicly available data, does not contain personally identifiable information, etc.) without revealing the underlying information.

5.1 Advantages

By providing a mechanism to "trust without disclosure," ZKP create new pathways for companies to substantiate claims, for industries to enforce evidentary mechanism, and for regulators to evaluate compliance without undermining proprietary interests. In this sense, ZKPs hold the potential to bridge the gap between corporate needs and societal demands for transparency, positioning them as a cornerstone of verifiable AI governance. As such, from the standpoint of *Private Governance*, ZKPs present several advantages:

- **Trust without Disclosure** — ZKPs can allow AI companies to substantiate claims of compliance without exposing proprietary details. This feature is crucial in an industry where intellectual property is central to competitive advantage.
- **Independent Verifiability** — Unlike self-assessments or voluntary pledges, proofs generated through ZKPs can be checked by independent auditors, experts or regulators, fostering transparency, trust and accountability.
- **Tamper-Resistance** — Because ZKPs are grounded in cryptographic formalisms, the resulting claims are resistant to manipulation, thereby reducing risks of performative or misleading disclosures.
- **Scalability in Private Governance** — ZKPs can be embedded in certification frameworks, industry standards, or compliance audits, enabling governance bodies to scale trust across large numbers of AI systems without demanding full access to proprietary assets.

5.2 Limitations

Despite their promise, the application of ZKPs in AI governance faces several limitations.

- **Translation Gap** — High-level normative commitments (e.g., "the model is fair") are not trivially translatable into mathematical statements suitable for cryptographic proofs. Developing shared benchmarks and taxonomies that link policy concepts such as fairness or privacy to formal definitions remains an open challenge [18, 19].
- **Computational Costs** — Generating ZKPs for large and complex AI models, particularly large neural networks, can be computationally expensive. While advances in zkML (zero-knowledge machine learning) are beginning to mitigate this, practical efficiency remains a barrier to widespread adoption [14].
- **Enforcement Gap** — Even if ZKPs provide technically rigorous guarantees, their legal and regulatory recognition is not yet assured [17, 24]. Assertions of fairness, data privacy, or security may lack binding force if corresponding evidentiary mechanisms are not formally acknowledged by courts or regulators.
- **Standardization and Interoperability** — For ZKPs to become wide spread governance tools that can be also sustained before court as valid evidence, common standards and protocols are necessary [17]. Although several standardization initiatives are close to establishing standards for ZKP [31], the lack of widely accepted verification frameworks may keep proofs siloed, limiting their broader application.

6 Framework for Proof Design: Cryptographic Proofs for AI Governance

Designing cryptographic proofs with governance relevance is inherently complex, since their utility extends beyond mere computational soundness and verifiability. To be effective, such proofs must be analyzed and constructed in relation to multiple dimensions. In our framework *Cryptographic Proofs for AI Governance*, we begin by distinguishing proofs for *Public* versus *Private governance*. Each of these categories can then be further examined along three additional dimensions. The first is the *object of the proof*—the element and property of the AI system being formally verified. The second is the *audience of the proof*—the stakeholders to whom the proof is directed, which may include users, regulators, industry stakeholders and others. The third is the *timing of the proof*—whether attestations are produced *ex ante* or *ex post*. Taken together, these dimensions provide a systematic foundation for comparing proofs across governance contexts and for assessing how cryptographic primitives can be adapted to diverse needs.

6.1 Proofs for Public vs. Private Governance

Proofs for Public Governance As discussed in previous work [17], the design of cryptographic proofs for public governance must be law-anchored and regulatorily relevant. In this context, the primary object of the proof is to demonstrate compliance with statutory requirements (e.g., the GDPR's and the AI Act's data protection obligations). The intended audience for such proofs may include regulators and supervisory authorities, as well as users, affected individuals, and, ultimately, courts. AI-related cryptographic proofs for public governance in the EU should be closely aligned with the regulatory concept of *meaningful information*, which constitutes a cornerstone of the transparency obligations embedded across the EU's digital regulatory frameworks [2, 17, 32, 33]. For example, the GDPR anchors this concept in Articles 13, 14, 15 and 22 which obligate data controllers to provide individuals with relevant information about personal data in automated decision-making. These provisions establish a layered structure: ex-ante obligations (Articles 13–14) to inform individuals at the point of data collection about how their data will be processed, and ex-post rights (Article 15) to access information and verify the lawfulness of decisions after data processing has taken place. Cryptographic proofs can operationalize these obligations. For example, a verifiable proof of dataset provenance may satisfy *ex-ante* requirements, while a proof of non-discrimination in outcome (e.g., demonstrating whether sensitive information was used) can support individuals in contesting unfair decisions *ex-post*, thereby strengthening individual rights within the emerging frameworks of digital constitutionalism [34]. The AI Act extends these obligations beyond personal data creating a new set of obligations for which relevant proofs can be developed [17]. In this context, verifiable proofs for public governance can reinforce institutional trust, anchoring technical attestations in binding legal frameworks.

Nevertheless, providing proofs based on mathematical soundness alone is not sufficient. For cryptographic proofs to function as instruments of public governance, they must also be communicated in ways intelligible to their intended audiences. Regulators and supervisory authorities require documentation that is both comprehensive and legally valid, while affected individuals require explanations that are accessible and understandable in their natural language [35]. As emphacized in a recent judgment of the Court of Justice of the European Union (CJEU), technical complexity does not absolve data controllers of their duty to provide explanations in terms that users can understand [36]. Cryptographic proofs therefore cannot remain abstract or purely mathematical artifacts. They must be embedded within explanatory frameworks that link formal guarantees to legally relevant rights in an acessible and intellegent way. Otherwise, transparency through cryptographic proofs risks devolving into disclosure in form rather than in substance.

Proofs for Private Governance Proofs for private governance operate within the sphere of industry self-regulation and voluntary commitments. Unlike public governance, where legitimacy is anchored in regulatory frameworks and legal mechanisms, private governance relies on market trust, reputational credibility, and relational accountability [7]. In this context, cryptographic proofs can be particularly valuable, as they can enable AI companies to substantiate voluntary claims through mechanisms that are both independently verifiable and tamper-resistant, yet protective of proprietary information. For example, an AI company might use a zero-knowledge proof to demonstrate compliance with a fairness threshold in hiring algorithms or to show that sensitive data categories were excluded from training sets—without exposing model architectures or datasets themselves.

Furthermore, proofs of this type can help reinforce trust among so-called deployers, that are often organizations or developers that place AI systems on the market or integrate them into services and therefore may inherit certain liabilities and risks originating from the system itself. As noted in [23], deployers often face uncertainty when relying on upstream systems of 'partialy' open-source models (where key AI components such as training data are hidden) [37], as they need to ensure compliance without always having full visibility into information that is essential for effective inspection and auditing. Cryptographic proofs can help mitigate this uncertainty by providing verifiable assurances about critical elements and characteristics of the AI system.

Moreover, proofs deployed in private governance must be intelligible to a wide range of stakeholders, including customers, business partners, investors, industry consortia, and others. While the threshold for intelligibility may justifiably be lower when addressing expert bodies or technical entities, these proofs cannot rely solely on technical precision. If stakeholders are unable to understand what a proof guarantees or how it connects to ethical, performance or policy related assurances, the proof risks being dismissed as a symbolic gesture rather than a credible signal. For example, an "AI fairness

seal" backed by a cryptographic proof will only foster trust if its meaning is conveyed in terms that non-technical actors can grasp: which dimension of fairness was evaluated, what benchmark was satisfied, and how the assurance relates to user rights or expectations. Accordingly, proofs in private governance (much like those in public governance) must operate as dual instruments. On one level, they must be computationally rigorous, safeguarding proprietary interests while evidencing technical conformity with voluntary standards. On another level, they must be communicatively transparent, translating formal guarantees into clear and actionable assurances that can be trusted by non-technical stakeholders.

Dimension	**Public Governance Proofs** (law-anchored, regulatorily relevant)	**Private Governance Proofs** (industry/self-regulation, voluntary)	**Impact on Trust**
Object (What?)	Proofs for compliance with regulatory mandates e.g., on data protection (GDPR, AI Act, etc.)	Proofs for voluntary claims, ethical AI labels, quality seals, etc. for AI elements and related specifications (e.g., model and fairness)	Public governance strengthens *institutional trust* by confirming effectiveness of mandates implementation; Private governance supports *market and reputational trust*.
Audience (To whom?)	Regulators, courts, supervisory authorities; proofs serve as enforceable evidence.	Users, deployers, industry consortia, civil society, etc.; proofs signal good practice but lack legal enforceability.	Public governance underpins *formal trust*; Private governance builds *relational and societal trust*.
Timing (When?)	Ex-ante: regulatory gatekeeping before deployment; Ex-post: enforceable accountability and redress.	Ex-ante: industry trust marks or voluntary reviews before release; Ex-post: reputational disclosure or post-incident statements.	Public governance generates enforceable trust through accountability and redress; Private governance generates anticipatory trust through voluntary assurances and ex-ante guarantees.

Table 1: Proofs in Public vs. Private Governance and their Impact on Trust

6.2 Other Dimensions of Proofs in AI Governance: what, for whom and when

From a perspective of algorithmic transparency, the notion of a meaningful explanation is often conceptualized as comprising two distinct but complementary categories: system-level transparency and outcome-level (or case-specific) transparency [38]. System-level transparency refers to information about the general functioning, design, and governance of an AI system, while outcome-level transparency concerns explanations of particular decisions or outputs. Cryptographic proofs, however, do not usually fit neatly into the category of system-level transparency. Because they are designed to attest to specific properties or constraints (for example, that a dataset excludes personal identifiers) proofs often target narrowly defined specifications rather than broad system-wide explanations. As a result, system-level transparency will typically require a combination or bundle of proofs that together provide a system-level information. Nevertheless, we argue that this binary categorization (system-level vs. outcome-level transparency) do not sufice for the effective design of proofs.

Another important aspect is the mapping of relevant governance considerations onto the design of cryptographic proofs. Ensuring that a proof aligns with a verifiable claim in sufficient granularity requires its decomposition into three dimensions: the object (what element/s and characteristic/s of the system are being verified), the audience (for whom the proof is intended, such as regulators, users, or auditors), and the timing (when the proof should be generated, whether ex-ante or ex-post). This tripartite dissection allows proofs to be more aligned with governance-relevant claims. Overall, this expanded view recognizes that different stakeholders demand different forms of assurance for different elements at different points in the AI lifecycle, and that proofs must be tailored to these contexts. The Table 2 below illustrates this framework, showing how cryptographic primitives including zero-knowledge proofs (ZKPs) can be used to provide more targeted, context-sensitive guarantees that complement traditional approaches to transparency.

To understand how this fits within the broader category of Proofs for Private vs. Public Governance consider the following. Here, the *object of the proof (what is being proved)* determines whether the claim concerns data, models, compute, etc. *Public Governance Proofs* at this level demonstrate compliance with binding rules, such as GDPR requirements for data minimization, and thereby reinforce institutional trust in systems that operate under enforceable standards. By contrast, when

the object of proof is addressed only through voluntary measures or internal company protocols (e.g., a self-declared fairness audit or proprietary data quality badge), the result is a *Private Governance Proof*. These proofs can be valuable in private governance frameworks, where firms and industry bodies develop their own mechanisms to signal responsibility, but their contribution to insitutional trust is limited to market and reputational trust, since they may lack legal enforceability.

Dimension	Examples	Possible Application of Cryptographic Primitives
What is being proved?	**Data**	E.g. Verifying datasets used to train a model; Verifying fair data use; Verifing that it does not contain personal information, etc. [12, 15, 25]
	Model.	E.g. Verifying claimed model capabilities and performance characteristics. Applying zero-knowledge proofs to AI systems enables privacy-preserving verification of claimed properties and confirms that inference uses the same model weights audited previously. [12, 13, 16]
	Compute	E.g. TEEs can generate a signature or hash tied to a computation's inputs and outputs, proving correct execution or ownership without revealing sensitive data. Verifiers can also use TEEs to privately test model weights or other confidential data. [12, 39]
To whom is it proved?	**Users**	E.g. Proofs that sensitive data was not used in automated decision, proof that model used in job selection is not discriminatory, etc.
	Auditors	E.g. Proofs that verify claims tied to committed datasets, models, or processes enables independent checks and oversight.
	Regulators	E.g. Proofs of training data provenance, model fairness, data privacy, security mandates, etc.
When is the proof produced?	**Ex-ante**	E.g. Proofs that certify properties of AI elements before they are deployed, preventing unverified systems from entering use.
	Ex-post	E.g. Proofs that reconstruct compliance, ensuring post accountability without exposing sensitive information.

Table 2: Framework for Verifiable Proofs with Cryptographic Primitives

7 From Traditional Assurances to Proofs: a Paradigm Shift.

Cryptographic proofs and in particular zero-knowledge proofs (ZKPs) introduce a new paradigm for assurances in complex systems such as AI. Traditionally, assurance has depended on third-party audits, certification bodies, and standards organizations to verify that processes and products comply with safety, fairness, or security requirements. This approach, while foundational for trust in different sectors, depends on the credibility, neutrality and expertise of the human intermediaries involved. It also often requires the disclosure of sensitive data or proprietary information, which creates barriers to transparency and trust and may expose actors to legal risks, as illustrated by the recent Deloitte court case concerning private data exposure [40].

In general, applying cryptographic primitives to the verification of corporate AI-related claims marks a shift from trust in auditors and firms' own assurances to trust in cryptographic proofs. In theoretical scenarios, assurance no longer relies on third-party access to sensitive data, but instead on mathematically verifiable guarantees that can be independently validated by actors (assuming sufficient technical expertise). In more realistic setting however, intermediary bodies are required to translate these technical guarantees into intelligible explanations for relevant stakeholders. As such, these proofs do not eliminate the need for public and private governance structures, but they provide those structures with new tools that can enhance accountability, transparency, and trust in AI systems and their applications.

8 Conclusion

This paper explores the use of cryptographic primitives, with particular attention to zero knowledge proofs, as a means of enabling independent and tamper proof verification of corporate AI claims without requiring disclosure of proprietary information. We contend that when supported by cryptographic proofs, AI related claims can establish a robust foundation for enhancing transparency and fostering trust within private AI governance. Although such mechanisms cannot replace formal regulatory oversight, they can significantly reinforce the legitimacy of self regulation by embedding verifiable

evidence into corporate practices. In doing so, cryptographic proofs reorient private governance away from symbolic or performative commitments and toward practices that are demonstrable and verifiable. More broadly, these developments represent a significant step toward governance regimes that combine trustworthiness with the capacity to reconcile proprietary protections and societal expectations of accountability and transparency.

References

[1] Anthropic. Claude's constitution. https://www.anthropic.com/news/claudes-constitution, 2023.

[2] European Parliament and Council. Regulation (eu) 2024/1689 of the european parliament and of the council on artificial intelligence (artificial intelligence act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689, 2024. Accessed: 2025-09-26.

[3] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2254–2272. ACM, June 2024. doi: 10.1145/3630106.3659037. URL http://dx.doi.org/10.1145/3630106.3659037.

[4] Benjamin S. Bucknall and Robert F. Trager. Structured access for third-party research on frontier ai models: Investigating researchers' model access requirements. Technical report, Oxford Martin School, Centre for the Governance of AI, October 2023. Available as PDF: https://cdn.governance.ai/Structured_Access_for_Third-Party_Research.pdf.

[5] Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Sandy Pentland, Arvind Narayanan, Percy Liang, and Peter Henderson. A safe harbor for ai evaluation and red teaming, 2024.

[6] Alejandro Tlaie and Jimmy Farrell. Securing external deeper-than-black-box gpai evaluations. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2503.07496.

[7] Asun Lera St.Clair. Trust in AI: The role of hybrid and private governance. In *Proceedings of the 57th ESReDA Seminar, October 23-24, 2019*, Valencia, Spain, 2019. ESReDA / Technical University of Valencia.

[8] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal III Daumé, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

[9] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3287560.3287596. URL https://dl.acm.org/doi/10.1145/3287560.3287596.

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2021. URL https://arxiv.org/abs/1803.09010.

[11] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte

Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward trustworthy ai development: Mechanisms for supporting verifiable claims, 2020.

[12] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, David Bau, Paul Bricman, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. Open problems in technical ai governance, 2025.

[13] Chhavi Yadav, Amrita Roy Chowdhury, Dan Boneh, and Kamalika Chaudhuri. Fairproof: Confidential and certifiable fairness for neural networks. In *2024 IEEE European Symposium on Security and Privacy (Euro S&P)*. IEEE, 2024. doi: 10.1109/EuroSP57164.2024.00030.

[14] Zhizhi Peng, Taotao Wang, Chonghe Zhao, Guofu Liao, Zibin Lin, Yifeng Liu, Bin Cao, Long Shi, Qing Yang, and Shengli Zhang. A survey of zero-knowledge proof based verifiable machine learning. *IEEE Access*, 12:104389–104408, 2025. doi: 10.1109/ACCESS.2024.3402193.

[15] Dami Choi, Yonadav Shavit, and David Duvenaud. Tools for verifying neural models' training data, 2023.

[16] Suppakit Waiwitlikhit, Ion Stoica, Yi Sun, Tatsunori Hashimoto, and Daniel Kang. Trustless audits without revealing data or models, April 2024. URL `http://arxiv.org/abs/2404.04500`.

[17] Simona Ramos, Vanesa Daza, and Alexandra Theben. Verifiable cryptographic proofs for ai governance: Bridging principles, regulation, and evidence. 2025.

[18] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. 2019.

[19] Joachim Baumann, Anikó Hannák, and Christoph Heitz. Enforcing group fairness in algorithmic decision making: Utility maximization under sufficiency. In *2022 ACM Conference on Fairness Accountability and Transparency*, FAccT '22, page 2315–2326. ACM, June 2022. doi: 10.1145/3531146.3534645. URL `http://dx.doi.org/10.1145/3531146.3534645`.

[20] Lawrence Lessig. *Code: Version 2.0*. Basic Books, New York, 2006.

[21] Mariarosaria Taddeo. Defining trust and e-trust: from old theories to new problems. *International journal of technology and human interaction (IJTHI)*, 5(2):23–35, 2009.

[22] Rachel Botsman. *Who can you trust?: How technology brought us together and why it might drive us apart*. Hachette UK, 2017.

[23] Simona Ramos and Fabio Pianese. Regulating transparency: A critique of the "free and open source ai" definition in the eu ai act, 2025.

[24] D. Bitan, R. Canetti, S. Goldwasser, and R. Wexler. Using zero-knowledge to reconcile law enforcement secrecy and fair trial rights in criminal cases. In *Proceedings of the 2022 Symposium on Computer Science and Law (CSLAW '22)*, pages 9–22. Association for Computing Machinery, 2022. doi: 10.1145/3511265.3550452.

[25] Yonadav Shavit. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring. *arXiv preprint*, 2023. URL `https://arxiv.org/abs/2303.11341`.

[26] Tobin South, Alexander Camuto, Shrey Jain, Shayla Nguyen, Robert Mahari, Christian Paquin, Jason Morton, and Alex 'Sandy' Pentland. Verifiable evaluations of machine learning models using zksnarks, February 2024. URL `http://arxiv.org/abs/2402.02675`.

[27] Haochen Sun, Jason Li, and Hongyang Zhang. zkllm: Zero knowledge proofs for large language models, April 2024. URL http://arxiv.org/abs/2404.16109.

[28] Oded Goldreich. Secure multi-party computation (final (incomplete) draft, version 1.4). Technical report / working draft, Weizmann Institute of Science, October 2002. URL https://www.wisdom.weizmann.ac.il/~oded/PSX/prot.pdf. Originally drafted June 1998, revised October 27, 2002.

[29] U. Feige, A. Fiat, and A. Shamir. Zero-knowledge proofs of identity. *Journal of Cryptology*, 1(2): 77–94, 1988. doi: 10.1007/BF02351717. URL https://doi.org/10.1007/BF02351717.

[30] Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. Trusted execution environment: What it is, and what it is not. In *2015 IEEE 14th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 57–64, Helsinki, Finland, August 2015. doi: 10.1109/Trustcom.2015.357. URL https://doi.org/10.1109/Trustcom.2015.357.

[31] Rob Viglione. Standards for zero-knowledge proofs will matter in 2025. https://cryptoslate.com/standards-for-zero-knowledge-proofs-will-matter-in-2025/, 2025. Publicado 15 de febrero de 2025; opinión / artículo en línea.

[32] European Parliament and Council. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679, 2016. OJ L 119, 4.5.2016, p. 1–88. Accessed: 2025-09-26.

[33] A. D. Selbst and J. Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 2017. doi: 10.1093/idpl/ipx022.

[34] Edoardo Celeste. *Digital Constitutionalism: The Role of Internet Bills of Rights*. Routledge, 1 edition, 2022. doi: 10.4324/9781003256908. URL https://doi.org/10.4324/9781003256908.

[35] Simona Ramos and Morshed Mannan. Watch the gap: Making code more intelligible to users without sacrificing decentralization? In *2022 IEEE 24th Conference on Business Informatics (CBI)*, pages 133–139. Institute of Electrical and Electronics Engineers, 2022. doi: 10.1109/CBI54897.2022.10059. URL https://doi.org/10.1109/CBI54897.2022.10059.

[36] Stibbe Law Firm. Cjeu confirms "right of explanation" in battle between trade secrets and algorithmic transparency. Stibbe Insights, March 2025. URL https://www.stibbe.com/publications-and-insights/cjeu-confirms-right-of-explanation-in-battle-between-trade-secrets-and. Accessed 25 September 2025.

[37] Andreas Liesenfeld and Mark Dingemanse. Rethinking open source generative ai: Openwashing and the eu ai act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pages 1774–1784. ACM, 2024. doi: 10.1145/3630106.3659005. URL https://dl.acm.org/doi/10.1145/3630106.3659005.

[38] High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy ai. Technical report, European Commission, 2018. URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. High-Level Expert Group on Artificial Intelligence, set up by the European Commission.

[39] Huili Chen, Cheng Fu, Bita Darvish Rouhani, Jishen Zhao, and Farinaz Koushanfar. Deepattest: An end-to-end attestation framework for deep neural networks. In *Proceedings of the 46th International Symposium on Computer Architecture (ISCA '19)*, pages 487–498, New York, NY, USA, June 2019. Association for Computing Machinery. ISBN 9781450366694. doi: 10.1145/3307650.3322251. URL https://doi.org/10.1145/3307650.3322251.

[40] Alexander Schmalenberger. Analysis of the cjeu judgment in case c-413/23 p (edps v srb). —, 2025. Case commentary.