Advances in Knowledge-Based Systems, Data Science, and Cybersecurity; Review Received 19-06-2025; Accepted 13-08-2025; Published 20-08-2025

Self-Attesting Intelligence: A Framework for Inherently Verifiable AI Systems

Zareen Hossain hossainzareen05@gmail.com
Dept of Law, Brainware University, India

Ritam Rajak ritamrjk@gmail.com
Department of CSE – AIML, Moodlakatte Institute of Technology, India

Indra Vijay Singh indravijay@mitkundapura.com
Department of CSE – AIML, Moodlakatte Institute of Technology, India

Tansuhree Das tanushree03552@gmail.com
Dept of Law, Brainware University, India

Abhinandan Pal Abhinandan799382@gmail.com
Dept of Law, Brainware University, India

Corresponding Author: Ritam Rajak

Abstract

The growth in the existence of complex, opaque Artificial Intelligence (AI) systems in other high-stakes societal fields, including finance, healthcare and law, has introduced a serious accountability gap. Their opaque decision-making logic also changes the way causation and liability can be traced according to legal principles of due process and undermines trust when the harmful results caused by these so-called black-box models are produced. Being a valuable means of producing human-readable explanations, the paradigm of post-hoc explainable AI (XAI) can, nevertheless and at times, turn into a source of unstable, partly misleading, and generally inadequate rationalizations that cannot be employed to support the evidentiary rigorous expectations of legal and regulatory review. This paper introduces a paradigm shift in the sense that subjective, explanatory post-hoc explanation is replaced by objective, a priori verifiability. We present Self-Attesting Intelligence, a new architectural framework that aims at guaranteeing correctness of an AI systems activity relation to a prescribed set of formal rules. It includes three underlying components, namely, a Declarative Knowledge Limiter (DKL) designed to translate legal and ethical rules into machine-enforceable format, a Constrained Inference Engine (CIE) which is the engine that enforces the rules in real-time during the decision process of the model, and an Attestation and Proof Generation layer that utilizes cryptographic techniques, namely, Zero-Knowledge Proofs (ZKPs) to generate unforgeable Certificate of Compliance in each decision. This certificate mathematically shows that the system has been then run within its constraints that are mandated by law without disclosing any sensitive input information or proprietary model information. Directly integrating compliance into the system through its design, such a

Citation: Ritam Rajak et al. Self-Attesting Intelligence: A Framework for Inherently Verifiable AI Systems. Advances in Knowledge-Based Systems, Data Science, and Cybersecurity. Research. 2025;2(2):15.

Page No.-291-309

framework reverses the relationship between accountability and the opaque model and where legal and regulatory scrutiny should be directed, on the rules established by people, rather than the poorly understood mechanism itself. We discuss the radical implications of this technology to facilitate the establishment of automated auditing, the redefinition of legal responsibility, and the establishment of a standard of care of AI development. Finally, we address the primary challenges to implementation, including the computational overhead of cryptographic proof generation and the normative difficulty of translating ambiguous ethical guidelines into formal logic, outlining key areas for future research.

Keywords: Self-Attesting Intelligence architecture, Declarative Knowledge Limiter, Constrained Inference Engine, Knowledge-Based Systems, Algorithmic Accountability.

1. INTRODUCTION

1.1 The Proliferation of AI in High-Stakes Societal Domains

Artificial intelligence (AI) and machine learning (ML) technologies have become central elements of the contemporary infrastructure systems, and they make independent or semi-autonomous decisions in areas associated with important choices of society [1]. And this is qualitative transformation in the governance and way of functioning of the society, the applied critical thinking is increasingly left to complex computerized systems. Alcoholic trading machines perform most of the market functioning in the finance sector at a pace that is uncontrollable by humans and the ML models also determine the creditworthiness and the possibility of insurance cost to the people which directly affect the economic life of millions [2]. Another sector with profound integration is healthcare where diagnostic AI, the next generation of diagnostic algorithms running on artificial intelligence (AI), has proven itself comparable or better to expert humans in advanced fields such as radiology, pathology, or genomics, where a misdiagnosis can literally be a matter of life or death [3].

In addition to its commercial use, governments and courts use predictive algorithms to serve various purposes, e.g., allocation of resources in cities and calculation of risks when deciding on pretrial bail and sentencing [4, 5]. The use of this application spreads into the operational use of critical infrastructure, the development of autonomous transportation networks, and it even enters the sphere of national security with AI being used to analyze intelligence and detect threats [6]. In both of them, the size, pace, and independence of AI-based decision-making bring about unprecedented efficiencies and capabilities surpassing those of human institutions. They do, however, usher in a different breed of risk which the current oversight systems are not designed to take on.

1.2 The "Black Box" Dilemma as a Barrier to Accountability and Trust

The rising level of complexity of these AI systems and specifically those based on the deep learning architectures with millions or billions of parameters have seen the establishment of what is known as black box problem [7]. The complexity of how these models work is also high-dimensional to a degree that even the model developers cannot comprehend, not mentioning external stakeholders or the subjects of the decisions that are being made [8]. This secrecy is not a shallow matter, this poses

the core challenge to accountability both in the legal and the social sense. When an autonomous system harms somebody, it is extremely hard to figure out causal relationships or responsibility, and thus legal frameworks that rely on notions related to intent, knowledge, and foreseeability become significantly less effective [9].

To address this concern, a new science known as Explainable Artificial Intelligence (XAI) was developed in order to make decisions more understandable in algorithmic based systems [10]. Nevertheless, LIME or SHAP are existing XAI methods that could only allow after-the-fact defense or point out the input that was influential in a model. Such explanations do not often result in the level of evidence that can be tolerated in a court of law or regulation [11, 12]. It has been demonstrated in the research that such explanations are partial, unstable, deceitful, or susceptible to manipulation by adversaries [13]. They usually provide a descriptive report of correlation but they are unable to provide the causal and counterfactual analysis which is currency of legal and scientific analysis [14]. Therefore, these ex-post rationalizations are not conducive to the trust that is needed to responsibly implement AI in society in the most sensitive spheres.

1.3 The Verifiability Standard: A Framework for Inherent System Accountability

These inherent flaws of post-hoc explainability make it necessary to switch to a new paradigm, in which the systems themselves become accountable. In this paper, we discuss that to integrate AI safely and fairly in the social fabric, the benchmark should be objective verifiability of the explanation, this must be mathematical. It suggests a new type of systems, which are called Self-Attesting Intelligence, and which have an architecture designed in such a way as to be unable to generate an output which breaches a previously specified base of formal rules. When the term self-attesting is used, the connotation is that through every decision it makes, the system itself will generate its cryptographic attestation of compliance, a certificate that cannot be forged, asserting that it has run according to the governance by its mandatory rules.

Its driving idea is that it is possible to build artificially intelligent systems that are provably accountable by incorporating constraints directly into the reasoning process, as opposed to expressing a given choice, it is possible to construct a system that is provably accountable. Not a way of opening a black box to see what goes on inside it, this approach is a means of production of a system whose architecture is evidence of its compliance on its own. The framework has the potential to escape the present stalemate between innovation and responsibility and potentially a source of technical foundations underpinning the development of AI systems that it is possible to trust in a genuinely social way.

2. THE LEGAL AND REGULATORY IMPASSE OF OPAQUE AI SYSTEMS

2.1 The Challenge to Foundational Legal Principles

The application of opaque AI systems poses a direct challenge to the principles that the legal system of the Western world is based on, the right to due process and the possibility to determine a clear

causality chain to assign liability [15]. Procedurally, due process entitles one to be given fair warning and a chance to be heard prior to the state taking away his/her life, liberty or property [16]. The external setting through which a consequential decision is arrived at by the black box algorithm, destroys the very framework on which a meaningful hearing is held. In the absence of information concerning the reasoning process used in arriving at a decision, a person can not be able to question its legitimacy. No matter how incorrect it might be, the black box nature of the model makes rational objection impossible, thus making the right to be heard essentially nonexistent and resulting in what some scholars call an algorithmic due process violation [17, 18].

More so, tort and product liability law are based on the possibility to expose the causation and define accountability [9]. When an autonomous car is in an accident or when AI-driven financial instrument leads to loss in the market, it becomes an exercise in vagueness to determine where to lay a claim. This may be due to responsibility on the part of the developer, the data provider, the user or the system itself. The inability to interpret the choice of the AI poses a near-impossible burden of evidence in the hands of a claimant and it makes adjudging liability among the various stakeholders in the develop-deploy continuum extremely hard [19]. This liability law gap in liability cancels out the fundamental deterrent and reparative features of liability law.

2.2 The Inadequacy of Post-Hoc Explanations as Legal Evidence

The discipline of Explainable AI (XAI) pursues the goal of closing this accountability gap, but its current approaches are not typically up to the challenges of legal admissibility [12]. When subjected to the sanity test in a legal setting, post-hoc rationalizations which aim at justifying a decision once it has been made, bear some serious faults. To begin with, they do not necessarily tell the actual logic behind the model they explain; it may be a convenient but ultimately incorrect simplification of the model process or process explanation [14]. Courts need believable and factual accounts of instances, and an ex post approximation, which might not be an accurate description of how the decision process took place is not admissible [20].

Second, XAI-generated explanations have been proven to be in a tenuous state, that is, small, trivial changes in an input can result in a significantly different explanation of the same output, which makes them less reliable [13]. Third, such practices are subject to adversarial attacks, i.e., a system might be deliberately constructed to give benign-looking interpretations of malignant or partisan choices [21]. Such vulnerabilities complicate XAI methods in use now as a potential type of legal evidence, which should be solid and tamper-proof. Although they can reveal suggestive information, they are not the kind of information that can be adjudicated as their conclusion and verification is imperative.

2.3 Gaps in Current and Proposed Regulatory Frameworks

To address these issues, governors and regulatory organizations in many countries have started developing systems that regulate the field of AI usage. The most promising developments are new legislative frameworks, the recently adopted AI Act, applicable across the European Union, which introduces a risk-based approach and requires high-risk systems to be transparent to those who are affected by them [22]. Nevertheless, even these innovative rules leave the stalemate unsolved in full.

In a wide range of regulations, the availability of the information that is understood as sufficiently transparent or actually meaningful concerning the logic of an automated decision is vital [23].

Such a well-meaning demand tends to fall back on the same approaches to XAI whose legal and technical insufficiency has already been described. The scope of the current regulations fails to outline a technical description of what would describe a sufficiently meaningful explanation thus making an interlude between the regulatory intent and the technical reality. The frameworks are framed with high consideration on pre-market risk assessment and post-market surveillance with less acumen provided to the most important issue, the means of certifying overall internal compliance of a live-operating system. Regulatory oversight has not become a form of ongoing system rather than post-hoc, since there is still no technical process by which one can objectively and continuously determine whether a system is within its established legal and ethical domains.

This contradiction between the precise demands of law adjudication and the inherent inadequacies of the present explanatory practices is elaborated on in TABLE 1. The table is a direct comparison of the main requirements of legal evidence, including reliability, stability, and inability to manipulate by an adversary, to the established failures of post-hoc XAI explanations. The table summarizes the shortcomings of the common approaches to XAI in offering adequate assurances on implementing each standard to show the technical and jurisprudence schism that provides the rationale behind the paradigm shift.

Table 1: Comparison of Legal Evidentiary Standards and Characteristics of Post-Hoc XAI Explanations

Evidentiary Standard	Legal Requirement	Characteristic of XAI Explanations
Reliability/ Fidelity	Must be a true and accurate account of the facts.	Fidelity to the model's logic is not guaranteed; can be an inaccurate approximation.
Stability	Evidence should be consistent and dependable.	Explanations can be unstable; minor input changes can drastically alter the rationale.
Adversarial Robustness	Must be resistant to tampering or manipulation.	Susceptible to adversarial attacks; can be manipulated to hide bias or malice.
Verifiability	The basis of the evidence must be verifiable.	Inherently a post-hoc rationalization; not a verifiable proof of the decision process.

3. VERIFIABILITY AS A NEW PARADIGM FOR AI GOVERNANCE

3.1 Shifting from Subjective Explanation to Objective Proof

The weaknesses of traditional methods of explanation necessitate a paradigm change to how AI is governed. The argument to be presented in this paper is that the paradigm should shift towards objective verifiability as opposed to subjective explainability. Where explainability tries to present a causal sequence understandable to a human in order to justify a previously made decision, verifiability tries to generate mathematical proof that the decision-making process itself has met a series

of formally stated constraints [24]. This is a decisive difference explainability is an a posteriori interpretation whereas verifiability is an a priori guarantee of the present operation of a system.

The point is not anymore to know how the reasoning of an unconstrained model works but rather that the operations of that model can be guaranteed not to go outside a given space, determined to be acceptable. With the use of formal methods, one can get a definite either-or answer to the question of correctness: either the system worked according to its rules, or it did not [25]. It is a big step on the foundation of accountability that claims such high-stakes uses are moved to quantitative replicating statements.

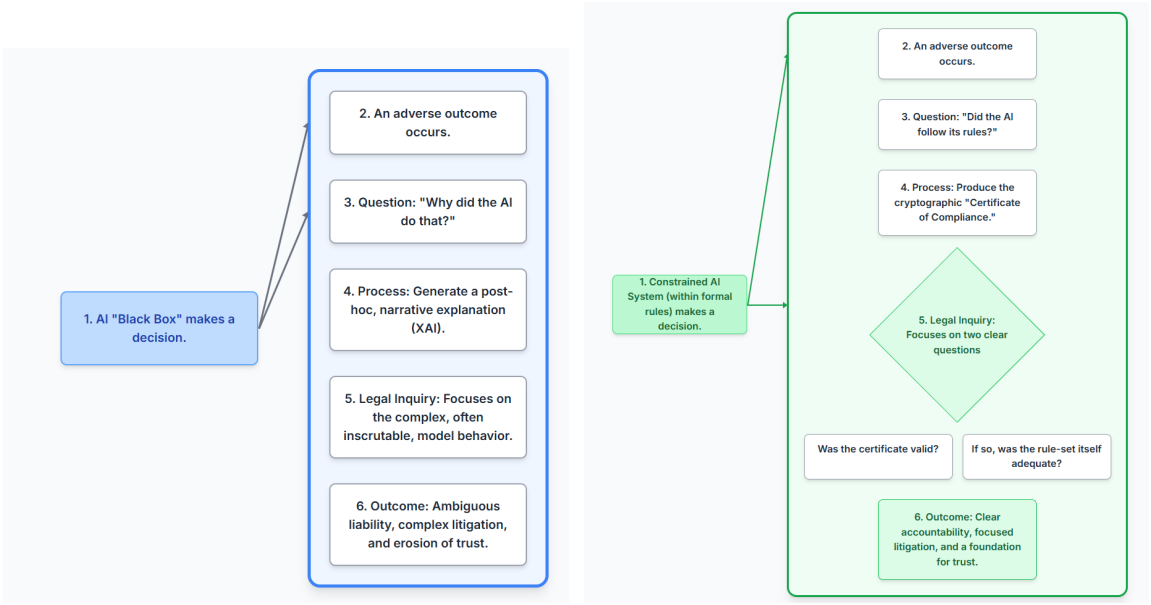

Figure 1: (a) Comparison of AI Accountability Paradigms. (b) Proposed Paradigm: Inherent Verifiability

The difference between the two paradigms of law and its implications on the process of legal inquiry are depicted in Figure 1. The paradigm described by the current reactive approach to post-hoc explainability (Part A) is compared with the postulated proactive approach to inherent verifiability (Part B). An erroneous output of the black box AI in the existing model triggers the investigation that is complex and inconclusive as the AI model is incomprehensible in its actions. In the model proposed, the questioning is diverted. Whether or not the AI has acted in accordance with its rules becomes the key question that can be answered by observing a cryptographic Certificate of Compliance. This refocuses the legal inquiry towards the following two more definite questions: was the certificate valid, and the rule-set itself satisfactory? This is now a systematic procedure that turns an obscure research process into a restricted, evident research, which provides a basis to certain accountability and trust.

3.2 Establishing Mathematical Verifiability as a Higher Standard of Care

Such existence of systems endorsable by architecture has direct consequences to the law of standard of care, a legal threshold, which stipulates the degree and diligence that an entity ought to take up to ensure that he or she does not cause harm to other individuals [26]. In the case of high-risk AI systems, one may argue that the standard of care will have to change so it includes considering applications of designing provably safe, and compliant architecture to be used. In this view, choosing to adopt a less transparent, unaccountable "black box" type of model in an essential application may be said to be a violation of duty in case a reduction in adopting a verifiable alternative is technically a possibility [27].

This can be compared to the principles that have been proved in the traditional engineering fields. A civil engineer is required to utilize materials and designs that are certified to resist particular, calculated stresses. On the same note, a developer of a critical AI system should be assumed to apply architectures where the operations are known to be confined to the formally specified structure of logical and ethical limits. The standard of care so gets uplifted beyond just scrutinising the results of a model but it is also about ensuring that the integrity of the internal processes are verifiable.

3.3 Implications for Redefining Legal Liability and Corporate Responsibility

The paradigm of verifiability elucidates the terrain of legal liability in essence. If verifiable AI systems would be to produce an unfavorable result, the legal discovery would change from trying to decipher the incomprehensible inner workings of the model and instead focus on the human-authored rules that define the operation. This type of architecture puts the blame on the more discrete human agents rather than the indecipherable algorithm [19].

Should the system have generated a valid certificate of compliance demonstrating that it acted in accordance with its rules, but nonetheless harm none the less has been caused then liability would most likely find its home in the insufficiency of the rules. The burden of such responsibility would then pass onto the office that designed that audit-rule-set, or implemented that rule-set, or audited them, be it an ethics-committee of the company, a group of legal engineers, or a regulatory agency. This provides a direct line of completion of accountability that is open to scrutiny and not veiled in technical darkness. Corporate governance of AI will thus have to be made over to include formidable operations on designing, validation, and continuous updating of formal constraints guidance that their intelligent designs operate within [28].

3.4 The Potential for a "Safe Harbor" Status for Verifiable Systems

A legal safe harbor [29] would be a strong regulatory incentive to increase the speed at which these more secure systems are adopted. Regulators can fix in place the organizations that are using AI systems that are certifiable via a verified regimen to some measure of shelter against any form of liability, including punitive damages in case of the occurrence of an unanticipated bad event. The basis in which an organization could have such a status would be a demonstration by such organization to have shown that the system was in use with a rule-set that was reasonable and safely designed to be safe and in compliance at the time it was formulated.

This safe harbor would exempt complete exemption since there would be the liability of negligence in the design of the rules. In any case it would generate a quite potent economic incentive of investing in more rigorous, verifiable architectures. It would provide an economic incentive to proactive risk management and good design in place of the existing practice of rolling out systems and trying to justify their actions ex post facto, introducing a market pressure where commercial and societal advantage become one as well as an environment of trustworthy AI [30].

4. THE ARCHITECTURE OF A SELF-ATTESTING INTELLIGENCE

4.1 High-Level System Overview and Workflow

Self-Attesting Intelligence is a framework that aims at the operations of an AI that can be proven to be compliant with a formalized set of rules. This has been done by adopting a design philosophy called trust by design, where checks aimed at ensuring the compliance are not added as an afterthought, but are directly integrated into the main computational algorithm. The three components of the architecture can collapse to three distinct, synergistic parts: the Declarative Knowledge Limiter (DKL) that is the normative authority, the Constrained Inference Engine (CIE) whose role is the operational enforcer and the Attestation and Proof Generation Layer that is the cryptographic guarantor.

The workflow of the system, as Figure 2 shows, is a well-organized and step-by-step process. This end-to-end flow is depicted in the diagram, and it starts with two external things being ingested: general legal and ethical rules, and the concrete data about an intended decision. Declarative Knowledge Limiter (Component I) is used to first convert abstract rules in a machine enforcible "Formal Rule-Set." This set of rules, and the input data, is then input into the Constrained Inference Engine (Component II). The diagram indicates that the CIE has two outputs to its two inputs namely final, compliant decision and the internal trace of less significant Computation Trace. This trace which constitutes a full record of the decision making process is safely handled by the Attestation & Proof Generation Layer (Component III). It is the last component to produce a so-called cryptographic Certificate of Compliance and ends up being the final check on the workflow, as it creates a tamper-proof, verifiable record of the integrity of the decision.

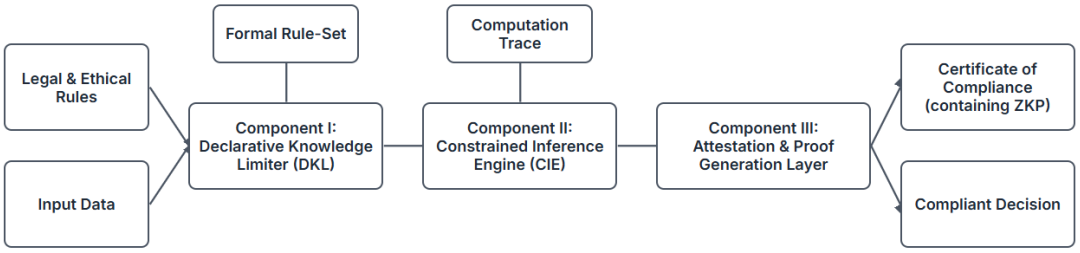

Figure 2: The Architectural Framework of a Self-Attesting Intelligence

4.2 Component I: The Declarative Knowledge Limiter (DKL)

The Declarative Knowledge Limiter (DKL) is the core element and provides the conversion of the high level, and generally rather ambiguous, policies in the legal, ethical and operational domain into a precise logically consistent representation that can be executed by machines. Its main purpose is to allow automating the knowledge acquisition bottleneck usually referred to as understanding how to capture complex human rules in a formal system [31]. It consists of two processes which are: knowledge engineering and formal representation.

During the knowledge engineering phase, the domain experts (lawyers, ethicists or compliance officers) give structured definitions of the constraints through a formal interface. Such rules may be straightforward, clear (in the form of a legal requirement, like, e.g. "A loan application decision cannot rely on a protected demographic attribute,") or they may be complex operations (such as the speed of a robotic arm cannot be faster than X when sensor Y is on). Mitigation of ambiguity To confront this level, this stage can use controlled natural language or domain specific language which directly correlates with logical primitives. This pre-planned input is essential in solving the widely reported problem of conversion of discretionary human norms into precise logic. This translation, of normative to the formal, is done using methodologies developed by the DKL to support versioning of evolving standards, working groups of experts (using collaborative authoring tooling).

These inputs are implemented during the formal representation stage as dynamic knowledge base using established formalisms [32]. An example would be an ontology, a formal, explicit description of the meaning of concepts and their properties; it would specify what we mean by the term "protected demographic attribute", and how we relate this term with another term, so-called "loan application." The rules are in turn coded in expressive logical languages. Descriptions and relationships of classes can be defined in Logics as the Attributive Language with Complements (ALC) which can be used, and rules be dependent to events or changing over time, to a linear Temporal Logic (LTL) [33]. The output is not a stand-off document but a queryable knowledge graph which is the source of truth to the Constrained Inference Engine.

4.3 Component II: The Constrained Inference Engine (CIE)

The Constrained Inference Engine (CIE) is the working-backbone of the architecture. It performs the inference process of the AI model within the framework of the rules that are very strictly obeyed according to the DKL. The designed component is a hybrid system tightly coupled, thus alleviating the brittleness of a conventional expert system and the obscurity of ML models with zero constraints [34]. The essence of CIE is that it embraces active imposition of constraints as a component of the inference process, and not as an ex post facto sieve.

Rather than construct a solution with the model, then checking that such a constructed solution falls within compliance, the CIE finds the combination of the rule-set within the DKL to be computationally sensitive. How that constraint will actually be enforced is a matter of the architecture of the ML model. In the case of deep neural networks, the rule-set can be combined in the form of policies to prune the computational graph. As an example, in some cases when a rule prohibits some specific combination of features impinging on an output, it is possible to dynamically determine the respective weights in the network to zero during the forward pass, making such unwanted inference

path computationally infeasible [35]. In the case of reinforcement learning agents, the rule-set may specify a special notion of an "action mask" such that at any particular step any action that would break a safety/legal constraint is considered to be unavailable to the agent at that step [36]. Such deep integration will make the resulting output not compliant under any circumstances but the system can only produce the resulting output under its specified constraints.

4.4 Component III: The Attestation and Proof Generation Layer

The CIE will ensure compliance within its own circles but making such compliance externally verifiable is the Attestation Layer, in a way that does not affect security and privacy. This is done by employing sophisticated cryptographic schemes to do so, the most prominent being Zero-Knowledge Proofs (ZKPs) [37]. A ZKP is a quite strong cryptographic method, which enables one party (primarily called the prover) to provide convincing to the other party (primarily called the verifier) that a sentence is valid, without revealing (recommending any) information more than the truth of the very utterance.

Such an architecture is structured in that all of the running of a CIE on a particular decision, the data fed to it, the calculations of the model, and the enforcement of the constraints in the DKL, are captured in one big, structured mathematical formalism, e.g. an arithmetic circuit [38]. The prover, which is the Attestation Layer, takes this circuit and forms a compact non-interactive proof, e.g., a zk-SNARK [39, 40]. A small amount of data demonstrating in a mathematical way that the whole calculation has been carried out correctly and in a manner that holds within all the limitations.

Such a proof is in turn packaged into a signed document known as the "Certificate of Compliance," which contains a cryptographic hash of the rule-set version that was used, a timestamp, and a globally unique transaction ID. The verifier may then be an external auditor or regulator or court. They can be cryptographically sure of the fact that the AI system ran within its own rules on each particular transaction simply by running a fast verification algorithm on the certificate. Importantly, if the verifier wants to verify this, he never has to know the proprietary model of the organization, or ever view the sensitive data being inputted, or even he does not have to re-compute the expensive computation. This attains a different level of cryptographic guarantee.

4.5 Hypothetical Case Study: AI-Driven Loan Approval

To demonstrate how the framework can be operationalized, take the case of an AI system utilized by a bank to make decisions on loan approval where one of the important legal considerations is that of non-discriminatory lending. The first step is the Rule Formalization process; in this step, a legal engineer collaborates with compliance officers to formalize the following wordings of the regulation into the DKL, e.g., "In any loan application L, the decision of the output D shall be computationally independent of the applicant protection attributes A (e.g., race, gender)." Next, in the process of Constrained Inference, the CIE operates this rule before the information is presented to the ML model when a new application is being processed. It isolates the attributes which are required to be secured and then applies a security process such as dynamic weight masking to make sure that these particular inputs do not affect the later levels of neural network. Although they could still use correlated information (such as zip code), the main attribute that is protected is verified to

have no part in the core decision pathway. Lastly, all components of the process are put together in an Attestation and Verification process that forms a ZKP. After such examination, the system states its decision ("Approved" or "Denied") with a further "Certificate of Compliance." In case an auditor raises queries to the decision, the bank presents the certificate. This proof can be verified by the auditor, and can mathematically demonstrate that the data about the applicants was not used in the decision logic (without the auditor seeing any of the personal declaration of the applicants or without the model used at the bank being disclosed). In case the test is sound and there is suspicion of harm after all, the magnitude of the investigation would then be directed towards whether the rule was satisfactory.

5. ANALYSIS OF SYSTEM PROPERTIES AND PERFORMANCE

5.1 Formal Guarantees: Verifiability, Soundness, and Non-Repudiation

The Self-Attesting Intelligence architecture is developed to offer a group of strong stipulations which are mathematically supported that becomes the origin of its credibility. The first ensures that the guarantee is verifiability This is the fact that enables an independent third party to confirm that a particular decision followed the active rule-set by checking the "Certificate of Compliance" using a rapid mathematical process. No access to the proprietary model, the sensitive input data and a re-running of the computationally expensive inference procedure is required to do this process. It implements a post-factum-auditable auditing service which is part and parcel of the Zero-Knowledge Proof (ZKP) embedded in the certificate.

The second assurance is healthiness Soundness Computer A sound cryptographic system is a system whose verification carries with it the property that it is infeasible, given a false statement, to generate a valid verification of that false statement by a malicious or malfunctioning system [37]. In this context, it implies that the Constrained Inference Engine (CIE) cannot issue a valid certificate over a decision that, in fact, went outside the rules that the CIE is informed about. Having the ZKP scheme with mathematical integrity is a solid protection against deviations; any manipulations will be reflected in the lack of compliance with the computational path by the set of formal rules, which will be rejected immediately by the verifier. This is to ensure that the system cannot lie that it does not comply.

Non-repudiation, the third guarantee, is implemented by usual public-key cryptography. The organization putting the AI system into operation signs each of the so-called notification of compliance documents digitally. The tagging of such a rule-set with this signature, coupled with an un-forgeable timestamp and an un-forgeable cryptographic hash of the version number of the rule-set that was applied, forms a tamper-evident and non-deniable record. An operating entity can not later reject that a particular decision was taken in a particular set of rules at a particular time and a third party cannot go and forge a certificate and say it was done by the system. This forms a verifiable audit log, which is critical to legal and regulatory processes.

5.2 Privacy and Security Analysis of the Attestation Process

One of the fundamental design principles of the framework is decoupling accountability and data exposure. This is done by using the zero-knowledge property of the attestation procedure that will offer strong privacy safeguards to people and also protect intellectual property of the organization [39]. An audit of a decision such as whether or not prohibited discriminatory factors have been used in a loan application can be effectively carried off without the audit, regulatory, or indeed any other party ever glimpsing the sensitive personal data of the applicant. The evidence shows that the rule was adhered to but nothing of the data on which the rule was used is given.

At the same time, the ML model applied by an organization is commonly a major trade secret and competitive advantage. The ZKP affirms the behavior of the model on a certain computation without exposing architecture, parameters, as well as the weights. Such dual security becomes a vital requirement of integrating the use of accountable AI in highly regulated sectors like healthcare (affecting HIPAA) and financial sector (affecting GDPR and other data protection policies) [41]. Also, this provides security: this also makes the system resistant to model extraction and inversion attacks, where an adversary tries to steal the model or reverse-engineer training data by asking it lots of questions.

5.3 Performance and Scalability: Analysis of Computational Overhead

The feasibility of such a framework implementation resides in a way of overcoming computational load which is being added as a result of the cryptographic proof generation. The longest performance bottle neck is Attestation Layer. Encoding the inference path of a contemporary, large-scale ML model into a related arithmetic circuit is not a trivial effort, and the amount of time it takes to form a ZKP is proportional to the amount of that circuit [38, 42]. In the case of deep neural networks having millions or even billions of parameters, the extra latency incurred by producing a proof may be untenable in real-time applications like autonomous vehicle driving, high-frequency trading, etc.

Nevertheless, this overhead can be alleviated, due to the existence of a number of design techniques and technology. First, in non-latency applications, decisions may be processed in batches and the cost of one larger proof may be spread out into thousands of individual transactions. In the second place, ZKP schemes are still under development, and there is continued research in speeding up the process of generating proofs (using, e.g., recursion, as in STARKs) and designing specialized hardware accelerators (FPGAs and ASICs) to carry out the underlying computations with unprecedented efficiency [43, 44]. Third, optimization methods of AI models like network pruning and quantization can also be applied to make the model size smaller, which also simplifies the corresponding arithmetic circuit and helps to build photonic-SI Proof faster.

5.4 The Trade-Off between Rule Complexity and System Efficiency

There is a trade-off underlying the framework between the logical expressiveness of rules in the DKL and the performance of the entire system. The more subtle and complicated the rules, particularly those involving intricate temporal conditions, nested conditional expression, or intricate cross-feature relationships, the bigger and more convoluted the constraints the CIE will be required

to ensure, and the Attestation Layer will be necessitated to demonstrate [33]. There is a direct computational cost payable to this increase in rule complexity, which affects both the inference latency within the CIE and, more so, the proof-generation latency at the Attestation Layer.

This fact brings out the need to have a tractable engineering way of working with rule-set design, balancing objectives of compliance with a budget on performance. This needs intensive, round trip co-designing between engineers and subject-matter specialists (lawyers, ethicists). The normative goals are defined by the domain experts and the costs of carrying out that formalisation are modelled by engineers. This can be invoked when it is necessary to estimate a highly complex rule with a fixed set of easier to compute constraints that accomplish the same practically feasible result. Figure 3 provides such a trade-off, conceptualized by means of the relationship between the rule complexity and computational cost. It can be seen in the diagram that, although the CIE seems to be apt at increasing inference latency only modestly compared to the rule complexity, the time spent in generating the proof at the Attestation Layer is exponentially higher. The idea with the practical application is that in this curve, there is the most "Optimal Operating Zone" that can give the best performance as close to the original human intent as possible, as well as achieving the performance and scalability demands of a system.

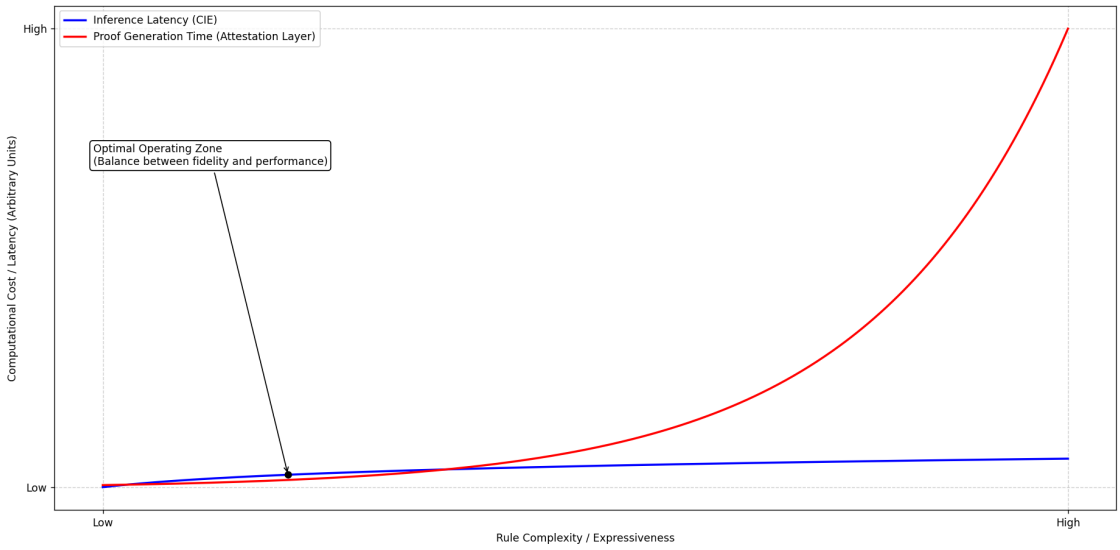

Figure 3: The Performance Trade-Off Between Rule Complexity and System Efficiency.

6. THE TRANSFORMATIVE IMPACT ON REGULATION AND DEVELOPMENT

6.1 For Industry: Integrating Compliance into the AI Development Lifecycle

The Self-Attesting Intelligence framework is a paradigm shift in the manner in which organizations plan to develop and put in place AI systems. Compliance is no longer a siloed and reactive activity

that the legal and auditing programs undertake once a product is out there. Rather it is an organised, proactive, ongoing component of development lifecycle itself. It can be likened to cultural and technical revolution in the current world of software engineering as DevOps and DevSecOps bubble up, wherein operations and security are integrated into every step of the process [45]. In this new schema it may be called DevComplianceOps and the formal rule-set produced by the Declarative Knowledge Limiter (DKL) is one of the first-class development artifacts, equal in importance to the model architecture or the training data.

Engineers and data scientists would create and predict models that could be tested not solely on the standard performance evaluation metrics, such as accuracy, but against the formal constraints of the rule-set as well. The Certificate of Compliance that the Attestation Layer generates is a quality gate that can no longer be negotiated on in the continuous integration and continuous deployment (CI/CD) pipeline. Any model build that does not create a valid proof of a set of predetermined test cases such as known edge cases, high-risk scenarios, would automatically fail. This ensures that a non-compliant system never sees a staging or production system. In this shift-left view of regulatory risk, in which possible problems are detected and resolved at their earliest stages of development, the possibility of an expensive regulatory fine is dramatically reduced, external regulatory audits become far easier, and makes the eventual passing of a system in regulatory compliance not only easier, but also makes a veritable chain of immutable and timeless evidence, as to how a system has operated in compliance since its very first startup.

6.2 For Regulators: Enabling Automated Auditing and Continuous Monitoring

In the case of regulatory bodies, this will serve as an additional tool of oversight not reliant on a particular geographic location with the capability to scale across locations in far more efficient and practically sound forms than the current oversight methods allow. Traditional auditing of AI systems can be a mixture of corporate documentation review, subjective impact analyses, and statistical analysis of historical performance all of which are time-consuming, one time efforts that cannot give a full perspective of internal logic of a system [46]. The formula of a quantitative verification is made up of the “Certificate of Compliance”.

This makes possible some sort of continuous and automatic monitoring of compliance. A regulator might be provided with programmatic and secure access to query one or more deployed AI ecosystems, and receive its cryptographic certificates regarding a given decision or over a given period of time. One such example is an agency creating a compliance dashboard that continuously ascertains the certificates of adherence to an anti-discriminatory lending regulation despite its machine-level of advancement by its lending AI used by a financial institution to evaluate trend-making lending decisions at the transaction level. In case the system fails to ever generate a valid certificate, or even if it generates a certificate against the obsolete or not-approved rule-set, an automatic alert that calls attention to regulatory review will be triggered. This transforms the regulatory paradigm to be of persistence and cryptography instead of haphazardly occurring and manual supervision, providing the capability to supervise much more comprehensively with fewer overseeing resources and offers the potential to identify and respond to non-compliance at it occurs and not months or years afterward.

6.3 For the Legal Profession: A New Form of Digital Evidence

The emergence of a cryptographically secure, mathematically verifiable Certificate of Compliance will make a tremendous difference in disputes in which an AI system is involved. The certificate is a novel type of digital evidence that has integrity, verifiability, and non-repudiation qualities, which occur to be much stronger than those of traditional system logs or the testimony, which is arrived at after the fact by experts [20]. Litigation involving AI-related damages would most likely make the certificate into a critical, dispositive piece of evidence. It was cryptographically sound, and its soundness could be verified standardly in mathematics, which was a radical change in legal proceedings.

Rather than involving an exercise of trying to understand the inner logic of a black box, litigation would be focused on two more pragmatic and people-centric questions; first, was the rule-set that informed the operation of the AI acceptable and legally sufficient? And second, did the party who came up with and authorised those rules themselves exercise reasonable care? This gives a clear path of enquiry and it makes the process of conducting law more effective and evident. It would also create a new type of technical competence in the legal profession. As opposed to speculative models of model behavior, security of cryptographic schemes or of logical consistency of the formal rule-sets would be testified by experts and the lawyers, judges, and experts would need to adjust to these new forms of evidence to be able to litigate and adjudicate cases [47].

6.4 The Emergence of New Professional Roles

Introducing this framework would require the design of new highly interdisciplinary proficient roles that would connect technology and ethics with the sphere of law. One of them would be the Legal Engineer. The role of this person or group would be the risky one of interpreting high level legal codes, corporate regulations and ethical principles to create the deductive reasoning demanded by the Declarative Knowledge Limiter, in the formally precise language demanded. This role requires an interesting mixture between the thorough mastery of a legal or ethical field by one side, and mastery of formal techniques, formal logic and knowledge representation languages by the other [48]. They would be the engineers of the formal rule-sets that would make AI behave.

In the same spirit, the job of the Algorithmic Auditor would change substantially. Rather than just focusing on statistical results as a key indicator of bias, this new auditor would be a technically competent professional who would be able to mathematically prove cryptographic compliance certificates. The toolkit of theirs would grow beyond the statistical packages into that of the cryptographic verifiers and formal methods softwares. Their role would be to endorse malice-free nature of the attestation machinery itself, ensure logical soundness of the rule-sets prior to their use, and to audit the certificates streams of live systems in real time. The introduction of these roles is symbolic of further professionalization of AI governance, where it is deemed as high-level policy work to a rigorous, technical and auditable engineering activity.

7. CONCLUSION

This paper has challenged the battle between the dissemination of powerful and without any clarity machine intelligence systems and the underlying needs of legal and societal responsibilities. It has presented the case that the current paradigm of post-hoc explainability though useful in interpretation, is inadequate to offer a stable basis through which governance should be carried out in high-stakes areas. This paper has replaced it with a system design called Self-Attesting Intelligence based on the responsibility of systems to operate in an architectural design. This combination of a formal rule-set (DKL), a restricted process of inference (CIE), and cryptographic mechanism of attestation (ZKP) has the effect of moving the margin of interpretation subjectively, to that mathematics that is objective.

The approach has three transformative advantages. It, first, puts in place a readable and audit-able chain of responsibility, pushing liability to the human-made rules as opposed to the opaque model. Second, it would facilitate a novel type of efficient, automated, and constant regulatory monitoring which would be based on cryptographic certitude. Third, it gives the legal system access to new form of powerful digital evidence that will be able to simplify litigation and direct legal investigation.

Although there is clear promise in the presentation of the Self-Attesting Intelligence framework, whether or not it becomes a reality is contingent upon what major gaps in its application need to be bridged, which also serve as the jumping-off point to subsequent studies. Technically, the biggest concern that still exists is the computational overhead and scalability of generating Zero-Knowledge Proof applied to the massive-scale models, which is used in real-time applications. Additional improvements of cryptographic algorithms and full-purpose hardware will be paramount in breaking this bottleneck.

On the normative side, the integrity of the framework exists entirely on the strength and faithfulness of the rules contained in the Declarative Knowledge Limiter. The task of articulating vague, conflicted legal and ethical ideals in black and white, in machine terms, as it were, is significant in its own turn. This translation process has to be developed with the use of robust methodologies and collaborative tools to make sure that the formalized rules will not reek of the novel biasness and will best show the original intent. A rule-set is no better than the principles that it represents; the garbage-in, garbage-out principle is relevant here as well.

Lastly, the implementation of such a paradigm is a social-technological problem. A purely technical fix, even assuming a forward-looking elegance, will never work without a concomitant evolution of the legal and regulatory frameworks, along with professional norms. The framework suggested in this paper is by no means ignobly touted as a panacea but as a sound technical basis on which more secure and righteous structures of governance can be established. Its main selling point is the fact that it shows that it is possible, to design AI systems, such that accountability is not a second-order property texture but rather amenable of first-order verification. Such a symbiotic co-design where the values desired in law and the places where human values are embedded in legal rules influences what is possible in technology design and vice versa, is what must happen to build a trustworthy AI ecosystem.

References

[1] Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 4th ed. Pearson. US. 2021.

[2] Jordan MI, Mitchell TM. Machine Learning: Trends Perspectives and Prospects. Science. NIH. 2015;349:255-260.

[3] Topol EJ. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. Nat Med. 2019;25:44-56.

[4] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?utm-source=chatgpt.com.

[5] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science. NIH. 2019;366:447-453.

[6] Jobin A, Ienca M, Vayena E. The Global Landscape of AI Ethics Guidelines. Nat Mach Intell. 2019;1:389-399.

[7] Burrell J. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. Big Data Soc. 2016;3.

[8] Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell. 2019;1:206-215.

[9] Wachter S, Mittelstadt B, Floridi L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. Int Data Privacy Law. 2017;7:76-99.

[10] Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. 2018;6:52138-52160.

[11] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst. 2017;30:4765-4774.

[12] Ribeiro MT, Singh S, Guestrin C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York USA: ACM. 2016:1135-1144.

[13] Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In: Proceedings of the 2020 AAAI/ACM Conference on AI Ethics and Society. New York USA. ACM. 2020:180-186.

[14] https://originalstatic.aminer.cn/misc/pdf/Molnar-interpretable-machine-learning_compressed.pdf.

[15] Kroll JA, Huey J, Barocas S, Felten EW, Reidenberg JR, et al. Accountable Algorithms. Univ Pa Law Rev. SSRN. 2017;165:633-705.

[16] https://www.scirp.org/reference/referencespapers?referenceid=2288614.

[17] Citron DK, Pasquale F. The Scored Society: Due Process for Automated Predictions. Wash Law Rev. 2014;89:1-33.

[18] Wachter S, Mittelstadt BD. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. Columbia Bus Law Rev. SSRN. 2019;2019:494-620.

[19] Das S, Kar SP, Sil S, Molla AR, Rajak R, Chaudhuri AK. A Multifaceted Approach to Understanding Mental Health Crises in the COVID-19 Era: Using AI Algorithms and Feature Selection Strategies. In: Khang A editor. AI-driven innovations in digital healthcare: emerging trends challenges and applications. IGI Global Scientific Publishing. ResearchGate. 2024:97-119.

[20] A. M. González de Miguel and A. Sarasa-Cabezuelo, ”A Global Approach to Artificial Intelligence,” in IEEE Access, 2025;13, :76946-76962.

[21] Ghorbani A, Abid A, Zou J. Interpretation of Neural Networks Is Fragile. In Proceedings of the AAAI conference on artificial intelligence . 2019;33:3681-3688.

[22] European Commission. . Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). COM/2021/206 final.2021.

[23] Kaminski, M. E. The Right to Explanation, Explained. Berkeley Technology Law Journal. 2019;34:189–231.

[24] Swaroop A., Abhishek S., Chandra G.,Prakash S.,Yadav S.K., et al., ”A Comprehensive Overview of Formal Methods and Deep Learning for Verification and Optimization”. International Conference on Decision Aid Sciences and Applications (DASA), Manama, Bahrain, 2024:1-6,

[25] Meng MH, Bai G, Teo SG, Hou Z, Xiao Y, et al,. Adversarial Robustness of Deep Neural Networks: A Survey From a Formal Verification Perspective. IEEE Transactions on Dependable and Secure Computing. 2022.

[26] Hacker, P., Krestel, R., Grundmann, S. et al. Explainable AI Under Contract and Tort Law: Legal Incentives and Technical Challenges. Artif Intell Law 2020;28:415–439 .

[27] Kar SP, Molla AR, Das S, Rajak R, Sil S, et al. Identification of Insecurity in COVID-19 Using Machine Learning Techniques. In: Khang A editor. Medical Robotics and Ai-Assisted Diagnostics for a High-Tech Healthcare Industry. IGI Global Scientific Publishing. 2024:239-256.

[28] Dignum V. Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer Nature. 2019.

[29] Parmadi, B. D. Cyber Safe Harbor 4.0: Advancing Ethics and Professionalism in Indo-nesia’s Digital Landscape. Eduvest - Journal of Universal Studies, 2024;4:11012–11033.

[30] Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center Research Publication. SSRN Journal. 2020.

[31] Hogan A, Blomqvist E, Cochez M, D’Amato C, Melo GD, et al,. Knowledge Graphs. ACM Computing Surveys (Csur). 2021;54:1-37.

[32] Hogan A, Blomqvist E, Cochez M, D'amato C, de Melo GD, et al. Knowledge Graphs. ACM Comput Surv. CSUR. 2021;54:1-37.

[33] Governatori G, Rotolo A, Sartor G, Gabbay D, Horty J, et al,. Logic and the law: philosophical foundations, deontics, and defeasible reasoning. Handbook of Deontic Logic and Normative Reasoning. 2021;2:655-760.

[34] Garcez A. D'a, Lamb LC. Neurosymbolic AI: The 3rd Wave. 2020. arXiv preprint :https://arxiv.org/pdf/2012.05876.

[35] Cingillioglu N, Russo A. Deeplogic: Towards end-to-end differentiable logical reasoning.2018. arXiv preprint arXiv:https://arxiv.org/pdf/1805.07433

[36] Pore A, Corsi D, Marchesini E, Dall'Alba D, Casals A, et al,. Safe reinforcement learning using formal verification for tissue retraction in autonomous robotic-assisted surgery. In2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) .IEEE.2021:4025-4031.

[37] https://eprint.iacr.org/2019/953

[38] Corsi D, Marchesini E, Farinelli A. Formal verification of neural networks for safety-critical tasks in deep reinforcement learning. InUncertainty in Artificial Intelligence.PMLR. 2021:333-343.

[39] Bowe S, Gabizon A, Miers I. Scalable Transparent and Post-Quantum Secure Computational Integrity. In: Proceedings of the ACM SIGSAC conference on computer and communications security. 2018;2018:46.

[40] file:///C:/Users/HP/Downloads/Explaining

[41] Price WN, Cohen IG. Privacy in the Age of Medical Big Data. Nat Med. 2019;25:37-43.

[42] Bünz B, Fisch B, Szepieniec A. Transparent SNARKS From DARK Compilers. In: Canteaut A, Ishai Y, editors. Advances in cryptology—EUROCRYPT. Springer. 2020:677-706.

[43] Stark W. The Sociology of Knowledge :Toward a Deeper Understanding of the History of Ideas.1st Ed.Routledge, New York.1991.

[44] Liang J, Hu D, Wu P, Yang Y, Shen Q, Wu Z. SoK: Understanding zk-SNARKs: The Gap Between Research and Practice. arXiv preprint arXiv:https://arxiv.org/pdf/2502.02387.

[45] Perwej Y, Abbas SQ, Dixit JP, Akhtar N, Jaiswal AK. A systematic literature review on the cyber security. International Journal of scientific research and management. 2021;9:669-710.

[46] Raji ID, Smart A, White RN, Mitchell M, Gebru T, et al,. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. InProceedings of the 2020 conference on fairness, accountability, and transparency 2020:33-44.

[47] SBanik S, Rajak R. Fake Review Detection: Taxonomies Benchmarks and Intent Modeling Frameworks. 2025;10.

[48] Soavi M, Zeni N, Mylopoulos J, Mich L. From legal contracts to formal specifications: A systematic literature review. SN Computer Science. 2022;3:345.