What Is "Verified AI"?

The History Behind an Expanding Term and What It Means Today

A Historical Genealogy and Contemporary Taxonomy of AI Verification

Basil C. Puglisi, MPA
A Human-AI Collaboration
basilpuglisi.com

Working Paper · September 2026
DOI: https://doi.org/10.5281/zenodo.22924739

Abstract

"Verified AI" sounds like a settled category in 2026, but it is not. This paper traces the term in three parts, in the order its history ran. Part I follows verification in artificial intelligence from the verification and validation of expert and knowledge-based systems, the subject of a meeting series that began in 1988, through the 2016 formal-methods formulation of Verified Artificial Intelligence and the assurance frameworks that followed it.

Part II shows how that concept shaped the HAIA Ecosystem, from checking AI output against a second platform in 2023 to structured reasoning records, checkpoint governance, hardware-rooted evidence, and governed session records. It then sets out how this work defines Verified AI: AI work whose outputs and decisions leave a record that can be audited and reconstructed, with its evidence rooted in what a human can review and accept, be it in hardware or code.

Part III examines the 2026 trend, in which verification language spread across standards, audits, agent identity, and provenance, and it closes with Proof-of-Control, an agent-verification standard released for public comment in September 2026. The standard arrives late to the Verified AI conversation, and by its own terms it stops short of the proof stage, because it places the quality of human oversight outside its evidence. Machines can produce and check evidence, but proof of control, as this paper defines it, requires a named human who verifies that evidence and answers for the decision that follows.

Keywords: Verified AI, verification, validation, TEVV, AI assurance, formal methods, Responsible AI, AI Governance, Checkpoint-Based Governance, proof of control, hardware attestation, provenance, agent identity

Suggested citation: Puglisi, B. C. (2026). *What is "Verified AI"? The history behind an expanding term and what it means today* [Working paper]. Zenodo. https://doi.org/10.5281/zenodo.22924739

1. Introduction

Artificial intelligence has gathered a growing vocabulary of trust. Systems are called safe, responsible, trustworthy, aligned, assured, validated, certified, compliant, and increasingly verified. Each word carries an intuitive promise, and verification may carry the strongest one, because something that has been verified sounds as though its uncertainty has been settled. In artificial intelligence, that assumption is dangerous.

No single, universally accepted status called "Verified AI" exists today. The phrase has a specific history in computer science, yet verification language now spreads across technical, institutional, and commercial settings that neither verify the same thing nor rely on the same evidence. A model may be formally verified against a mathematical property, an organization may test a system against governance principles, and a third party may confirm a developer's stated performance figures. A cryptographic protocol may authenticate which agent sent a request, and a provenance system may record where an AI-generated image came from. Each process can legitimately claim verification, but none supports every conclusion the others imply.

The distinction matters more each year, because artificial intelligence is moving from isolated models toward systems that retrieve information, call software tools, generate media, interact with other agents, and take actions in digital environments. Verification is no longer only a question of whether an algorithm satisfies a mathematical property; it may also concern identity, authorization, provenance, process, performance, governance, or evidence that an action took place. The central question is shifting from "Is this AI verified?" to "What, exactly, has been verified, and who checked the proof?"

The term is trending in 2026, and the attention it now receives makes its history easy to lose. This paper restores that history in the order it happened. Part I traces Verified AI before the work examined here: the verification and validation of knowledge-based systems, the formal-methods tradition, the 2016 formulation that gave the term its research meaning, and the assurance frameworks that widened it. Part II shows where that concept appears in the HAIA Ecosystem, how it shaped the frameworks built there, and how this work now defines Verified AI.

Part III turns to the 2026 trend, in which verification language spread across standards, audits, agent identity, and provenance, and contrasts that trend with Proof-of-Control, a standard for verifying AI agents released for public comment in September 2026. The argument running through all three parts is simple but consequential: verification is evidence supporting a bounded claim, not a universal property called trustworthiness, and proof is only proof when someone checks it.

Part I: Verified AI Before This Work

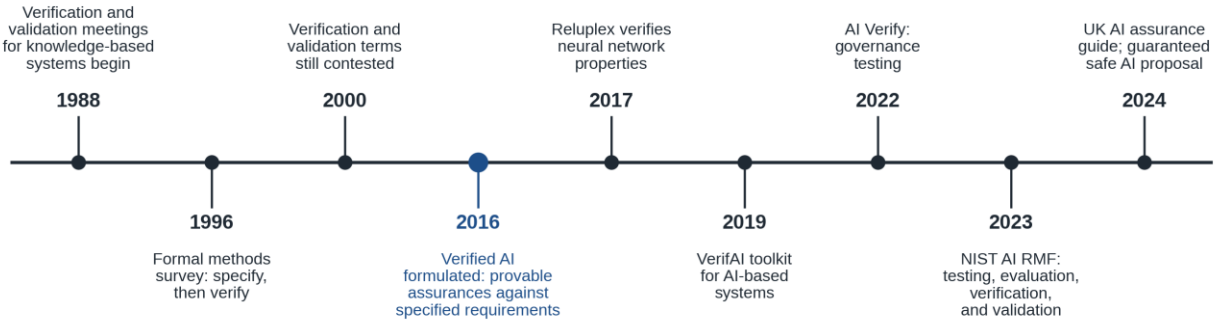

Figure 1. Verification in artificial intelligence before this work, 1988 to 2024. Compiled from the works cited in Part I.

2. Verification Before "Verified AI"

The history of AI verification begins before the modern vocabulary of machine learning safety. Expert systems and knowledge-based systems produced an early version of a problem that remains familiar. If a computer system is meant to reproduce expert reasoning, how does a developer establish that its rules are consistent, that its knowledge represents the domain accurately, and that its conclusions suit real-world use? Those questions produced a distinct literature on verification and validation.

North American meetings on the verification and validation of knowledge-based systems began in 1988. A 1998 report on two 1997 events described the planned AAAI-98 workshop as marking the first 10 years of those meetings (Antoniou et al., 1998). That 1998 workshop, cochaired by Daniel O'Leary and Alun Preece, collected papers on the verification and validation of knowledge-based systems as the field matured (O'Leary & Preece, 1998). Five years earlier, the 1993 AAAI workshop on the same subject already included work on the verification and validation of multiple-agent systems (Association for the Advancement of Artificial Intelligence, 1993).

The continuity is striking. Long before modern autonomous agents existed, researchers were working out how to verify systems built from distributed, interacting intelligent components, and the question of how separate reasoning parts could be checked as a whole was already on the table.

The terminology was never fully settled either. Mosqueira-Rey and Moret-Bonillo (2000) observed that most tools built to support verification and validation favored verification over validation and confined their analysis to a system's internal structures. Gonzalez and Barr (2000) went further, reporting that competing definitions of the two terms had multiplied and that their meaning, their differences, and their implementation remained deeply confused.

This history matters to the current debate. The ambiguity around "verified AI" is not simply the product of recent marketing or loose language. Verification and validation have always raised hard boundary questions about what is being measured, what counts as adequate evidence, and whether internal correctness establishes fitness for use. A system can satisfy its internal specification and still be unsuitable for its setting, and a system can appear to work well in observed conditions without

anyone showing that it satisfies every relevant property. That distinction grew more important as artificial intelligence moved from explicit rule systems toward systems that learn their behavior from data.

3. Formal Methods and the Meaning of Proof

A parallel history runs through formal methods, the mathematically based languages, techniques, and tools used to specify and verify software and hardware systems. In a 1996 survey, Clarke, Wing, and colleagues described formal methods as a way to specify systems rigorously and to determine whether implementations meet those specifications. They cautioned that applying formal methods does not by itself guarantee correctness (Clarke et al., 1996). Formal reasoning can expose inconsistency, ambiguity, and incompleteness, yet its conclusions still depend on the specifications, models, assumptions, and process behind them.

That caution remains central to AI verification. Formal verification does not answer an open question such as whether an AI system is trustworthy; it answers a bounded one, namely whether property P can be shown to hold for system X given assumptions A and specification S. The structure is powerful precisely because it is constrained.

Traditional software verification benefits from systems whose behavior is explicitly programmed. Machine learning complicates the task, because learned components take their behavior from training processes and data rather than from hand-written instructions alone, and their behavior can depend on high-dimensional environments that resist complete modeling. The meeting of these two traditions, formal proof on one side and learning systems on the other, set the stage for the modern research program called Verified AI.

4. 2016: The Modern Formulation of Verified Artificial Intelligence

The year 2016 marked a turning point for AI safety and verification. On June 21, 2016, Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané posted *Concrete Problems in AI Safety* (Amodei et al., 2016). The paper recast AI safety around practical failure modes in machine learning, among them unintended side effects, reward hacking, scalable supervision, safe exploration, and distributional shift. A week later, on June 28, Carnegie Mellon University and the White House Office of Science and Technology Policy held a public workshop on safety and control for artificial intelligence (Carnegie Mellon University, 2016). Its question was how increasingly capable systems could be engineered to operate in a safe, controlled manner.

Against that background, Sanjit Seshia, Dorsa Sadigh, and Shankar Sastry posted *Towards Verified Artificial Intelligence* on June 27, 2016, and restated the argument in *Communications of the ACM* in 2022 (Seshia et al., 2016, 2022). Their definition leads any serious account of the term. Verified AI, in their formulation, is the goal of designing AI systems that have "strong, ideally provable, assurances of correctness with respect to mathematically specified requirements" (Seshia et al., 2022, p. 46).

Under that definition, Verified AI is not a badge applied to a generally trustworthy system; it is a research objective grounded in formal methods. The authors framed the work around five challenges

(Seshia et al., 2022). They are modeling the environment in which an AI system operates, writing formal specifications of intended behavior, building useful models of learning systems, developing computational engines that scale, and designing systems that are correct by construction. Those challenges show why AI verification differs from conventional software verification. The AI system is only part of what must be understood, since the environment, the specification, the learned component, the feasibility of proof, and the development process all shape the result.

The most defensible historical conclusion is not that verification in AI began in 2016. Instead, 2016 marks a clear formulation of Verified AI as a modern research program that joined formal methods to artificial intelligence and machine learning.

5. From Concept to Technical Program

The years that followed produced concrete methods for verifying machine learning systems. Katz and colleagues introduced Reluplex, a satisfiability modulo theories solver for verifying properties of deep neural networks built on rectified linear units, and evaluated it on a prototype of an airborne collision avoidance system for unmanned aircraft (Katz et al., 2017). Reluplex did not make neural networks "verified" in any general sense; it showed that particular properties of particular networks could be proven, or counterexamples found, under defined conditions.

Dreossi and colleagues extended the work through VerifAI, a toolkit for the formal design and analysis of systems with AI and machine learning components (Dreossi et al., 2019). It pairs formal specifications with simulation, falsification, fuzz testing, parameter synthesis, counterexample analysis, and data augmentation. The shift matters conceptually, because verification increasingly became one part of a spectrum of evidence-generating methods rather than a binary choice between mathematical proof and no assurance at all.

The line continues in proposals for guaranteed safe AI. Dalrymple and colleagues described an approach built from three components: a world model, a safety specification describing acceptable effects, and a verifier that produces an auditable proof certificate showing the specification holds relative to the model (Dalrymple et al., 2024). The approach descends directly from the formal Verified AI tradition, and it also exposes that tradition's limit. A proof establishes what follows from the specified world model and safety specification, but it cannot guarantee that the model captures reality or that the specification captures every concern that ought to matter. Verification can be extremely strong without being universal.

6. AI Verify: Governance Testing as Verification

In 2022, Singapore's AI Verify added a governance branch to the story. The Infocomm Media Development Authority and the Personal Data Protection Commission launched it on May 25, 2022 as an AI governance testing framework and toolkit (Personal Data Protection Commission Singapore, 2022). It was built for organizations wishing to show responsible AI through a combination of technical tests and process checks. The framework assesses systems against 11 governance principles: transparency, explainability, repeatability and reproducibility, safety, security, robustness,

fairness, data governance, accountability, human agency and oversight, and inclusive growth with societal and environmental well-being (AI Verify Foundation, n.d.).

AI Verify was designed around self-assessment. At launch, developers and owners tested their own systems against the principles and produced reports for their stakeholders, and the toolkit now supports both self-assessment and independent third-party testing (Infocomm Media Development Authority, 2022; AI Verify Foundation, n.d.). That design choice bears directly on the question this paper keeps returning to: who checks the proof?

AI Verify and Verified AI are related in language but not in epistemic meaning. The formal-methods formulation asks whether specified properties can be rigorously established, while governance testing asks whether structured technical and process evidence shows how a system performs against governance criteria. Both involve verification, and they do not support interchangeable conclusions. The distinction grows more important as "verified" moves into standards, procurement, certification, and public communication, where a reader may assume a stronger claim than the underlying assessment supports.

7. Verification Expands Into AI Assurance

During the 2020s, verification became embedded in the broader vocabulary of AI assurance. The NIST AI Risk Management Framework, released on January 26, 2023, does not treat verification as a final status (Tabassi, 2023). It offers organizations a resource for managing the risks of designing, developing, deploying, and using AI systems, with testing, evaluation, verification, and validation running across that lifecycle. The question shifts from whether an algorithm satisfies one formal property to whether enough evidence exists to support decisions about a system operating in context. NIST's shorthand for that work is TEVV: testing, evaluation, verification, and validation.

The United Kingdom has built its own assurance vocabulary beyond formal verification. The Department for Science, Innovation and Technology published *Introduction to AI Assurance* on February 12, 2024, placing assurance inside a wider governance ecosystem. Its mechanisms range from risk and impact assessments to bias and compliance audits, conformity assessment, performance testing, and formal verification (Department for Science, Innovation and Technology, 2024). The difference between these vocabularies is not merely semantic. Formal verification seeks proof of specified properties, while AI assurance seeks enough evidence to justify confidence or a decision in a particular context, so verification can be one part of assurance without being the same thing.

8. Verification and Validation Are Not the Same

A common engineering shorthand separates two questions. Verification asks whether the system satisfied its specified requirements, and validation asks whether those requirements, and the system built to them, suit the intended use. The distinction is useful even though exact definitions differ across standards and fields, and it matters more for artificial intelligence because the specification itself may be incomplete.

Consider a hiring model whose formal requirement states that candidates scoring above a threshold advance to review. A verification procedure might establish conclusively that the software implements that rule. It would not establish whether the score measures job suitability, whether the features behind it create discriminatory effects, whether the training data represents the applicant population, or whether the process meets legal obligations. The system can be verified against its specification while the specification remains inadequate, which is the specification problem in practical form. Formal correctness therefore cannot stand in for validation, governance, legal review, or human judgment.

9. What a Verification Claim Must Name

The history above calls for a definition of the activity itself:

> **AI verification is the process of establishing evidence that a specified claim about an AI system, component, actor, output, process, or associated artifact satisfies explicitly stated criteria, using a defined verification method, within a defined scope.**

The definition does not require every method to be a formal proof. It requires that the evidence state what it supports.

A narrower reading of the label follows from it. A claim that an AI system or artifact is "verified" means something only to the extent that its parts can be identified. Those parts are its object, the claim evaluated, the criteria applied, the evidence considered, the method used, the verifier, the scope of the conclusion, and the period of validity. The sequence forms a verification record: Object, Claim, Criteria, Evidence, Method, Verifier, Scope, and Validity. Object identifies what was verified, and Claim states what is asserted about it. Criteria name the requirements or thresholds applied, Evidence names the information behind the determination, and Method names how that evidence was assessed. Verifier names who or what performed the assessment, Scope marks which conclusions are and are not supported, and Validity records when the evidence was generated and what changes would require a fresh assessment.

The record does not resolve every dispute in AI assurance. It does something more basic, which is to stop the word "verified" from carrying more meaning than its evidence supports.

Part II: Verified AI in This Work

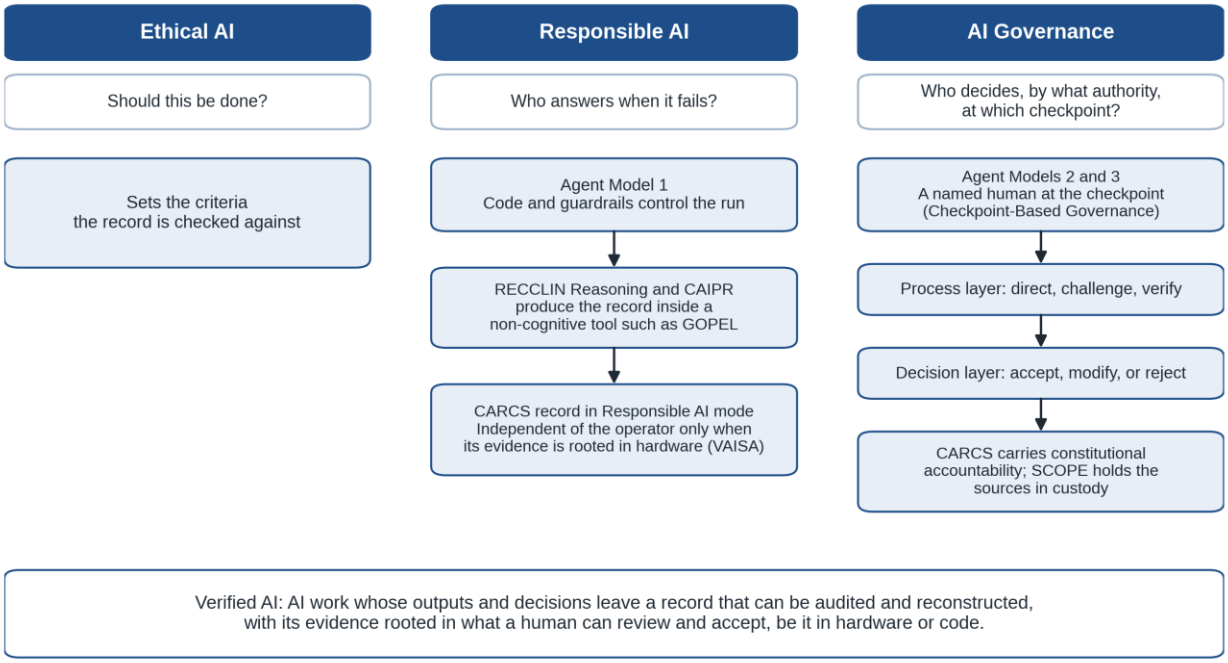

Figure 2. Verified AI in the HAIA Ecosystem across Ethical AI, Responsible AI, and AI Governance. Compiled from the works cited in Part II.

10. Checking the Work: Factics and RECCLIN Reasoning

The verification record that closes Part I already runs in practice in this work as structured output. Factics, the method at the base of the HAIA Ecosystem, pairs every fact with a tactic and a measurable outcome. In 2023, its source discipline moved AI work onto a second platform, because a single platform returned answers without reliable sources (Puglisi, 2026h). A public article on February 1, 2024 described that practice, naming fabrication as the problem, using a second platform to validate sources, and holding that neither platform is an authority (Puglisi, 2024). Applied to every AI output, the same discipline became RECCLIN Reasoning, which returns each response in 10 fields: Role, Task, Output, Sources, Conflicts, Confidence, Expiry, Fact to Tactic to KPI, Recommendation, and Decision (Puglisi, 2026m).

The eight elements of a verification claim map onto that structure, and onto the source record kept by SCOPE, with little strain.

Verification record	RECCLIN Reasoning field	SCOPE full source custody field
Object	Task	Source title, author or organization, identifier
Claim	Output	Exact claim used
Criteria	Fact to Tactic to KPI	Claim support classification
Evidence	Sources	Page or section; archive link
Method	Role	Status check method
Verifier	Decision	Author diligence note
Scope	Conflicts and Confidence	Source type; claim support classification
Validity	Expiry	Date accessed; later review date; later status change

The mapping is structural rather than an equivalence: each field carries the job of the corresponding element in its own setting. RECCLIN Reasoning adds two fields the eight-element record lacks. Conflicts preserves dissent rather than averaging it away, and Decision leaves the final choice to a person rather than to the system that produced the output. When the work runs across several platforms at once under HAIA-CAIPR, each return keeps that structure, and the dispatch record declares how independent the returns really are. Two returns from different models on the same platform count as two returns that share an access route, which the independence profile records (Puglisi, 2026h).

HAIA-CARCS turns the session behind that output into a governed record. It is a human-triggered protocol that produces a 10-section document capturing what happened, who decided, what the evidence was, and what comes next. Each human decision is classified as a Corrective Override, a Creative Supersession, a Checkpoint Confirmation, or a Deferred Decision (Puglisi, 2026a). The published protocol states the reason for that typing directly: "Governance that cannot tell the difference between a human who caught an error and a human who approved without reading is not governance" (Puglisi, 2026a).

The same record serves both tiers of practice. The May 2026 revision of the protocol states: "When CBG is active, the CARCS record carries constitutional accountability. When it is not, the record documents Responsible AI mode and declares the absence of named human authority" (Puglisi, 2026c). CARCS also grades its own integrity. An Attestation Grade record carries the practitioner's signed attestation, while a Hash Verified record carries cryptographic chain integrity. The protocol is candid about the gap between them: "A locally computed hash stored in operator-controlled storage is self-attested integrity, not externally verified chain of custody" (Puglisi, 2026a).

11. Ethical AI, Responsible AI, and AI Governance

The three tiers of the HAIA Ecosystem sort the verification problem by who owns each part of it. Ethical AI asks whether something should be done, and it supplies the criteria any record is checked against. Responsible AI asks who answers when a system fails, and it works through internal controls, traceability, and machines checking machines, yet it cannot name a person who answers for an individual output. AI Governance asks who decides, by what authority, and at which checkpoint,

and it places a named human with binding authority over the output. Verification of the machine sits in the second tier, while judgment about whether the criteria were right, and the duty to answer for the result, sit with people in the first and third.

12. Responsible AI: Agent Model 1

The HAIA-RECCLIN agent architecture, published in February 2026, defines three operating models that differ by checkpoint density and automation level (Puglisi, 2026i). They are Agent Model 1, Agent Responsible AI; Agent Model 2, Agent AI Governance; and Agent Model 3, Manual Human AI Governance. Agent Model 1 is automation. The agent runs the full pipeline without stopping, sends each functional role to several platforms, and delivers one package at the end, so the machine checks the machine and convergence across platforms constrains any single platform's instability. The specification treats the human review at the endpoint as "an informal safety valve, not a formal governance control" (Puglisi, 2026i).

Running consequential work on automation is a choice that belongs to the organization. Responsible AI has a legitimate place where stakes allow process controls to suffice and outputs remain reversible, and organizations will choose it for financial reasons wherever the risks do not outweigh the rewards. The problem arises only when that factory-quality process is presented as governance.

The tool that would carry an Agent Model 1 run is a non-cognitive enforcement layer such as GOPEL, which performs seven deterministic operations (dispatch, collect, route, log, pause, hash, and report) and evaluates no content. GOPEL was published as reference code on February 23, 2026, and its repository describes it as a working concept that is not yet proven or validated; it has not been deployed (Puglisi, 2026f, 2026g). In an Agent Model 1 run, RECCLIN Reasoning structures each output, CAIPR dispatches across models inside that layer, and the CARCS record that results is the proof of control, declared as Responsible AI.

Agent Model 1 still carries a formal control, and it points back to a person. Checkpoint-Based Governance named automation bias drift, in which reviewers progressively defer to AI recommendations without critical evaluation, as a critical governance failure in September 2025 (Puglisi, 2025c). Checkpoint-Based Governance requires passive acceptance to be detectable at every checkpoint rather than merely discouraged (Puglisi, 2026b). For Agent Model 1, the agent architecture tracks approval and reversal rates across review cycles, leaving the specific thresholds to implementation, and when those signals trip, the work escalates from Agent Model 1 to Agent Model 2 (Puglisi, 2026i).

Code also carries a limit that no amount of engineering removes. GOPEL's own test suite shows that "a well-crafted lie, one that contains no injection patterns, no Unicode anomalies, no delimiter tricks, and passes all structural checks, moves through the pipeline undetected," and it names the defense as the human checkpoint (Puglisi, 2026f). The operator also writes the code, runs it, and signs its logs, so an Agent Model 1 record built on code alone carries the operator's word. That limit leads directly to hardware.

13. Hardware and Operator Independence

The Verified AI Inference Standards Act, a model legislative proposal published on March 6, 2026 as the fifth document of the AI Provider Plurality Congressional Package, names the gap that code cannot close (Puglisi, 2026l). Data sent to an external AI platform is encrypted in transit and then decrypted inside a processing environment the sender cannot inspect. VAISA calls that interval the Invisible Moment: "This window, between when data leave a trusted boundary and when a response returns, is the Invisible Moment." Its diagnosis is one sentence long: "Contractual promises govern what should happen. Nothing verifiable proves what did."

VAISA's answer is hardware, built on the remote attestation architecture the IETF set out in RFC 9334. In that architecture, an attester produces evidence, a verifier appraises it against policy, and a relying party decides what the result permits (Birkholz et al., 2023). Under VAISA's hardware profile, "The attester cannot fabricate a valid attestation quote without access to the hardware private key held inside the TEE." The act continues: "The evidence is cryptographically bound to the physical hardware state at the moment of attestation" (Puglisi, 2026l). The proposal would require a compliant attestation quote to carry five elements. They are a hardware vendor signature, an enclave measurement of the approved code image, a freshness nonce, a trusted computing base version, and a signed processing receipt that binds the specific transaction to the attested environment. VAISA draws the contrast with contracts in terms that carry straight across to operator-written code: "The BAA governs the aftermath. Attestation governs the conditions."

Hardware-rooted evidence serves both tiers. For Agent Model 1, it makes the proof of control independent of the operator's word, because a valid quote requires a hardware key the operator does not hold. For AI Governance, it gives the checkpoint evidence rather than a provider's word; as VAISA puts it, without attestation "every checkpoint rests on the AI provider's self-report. That is not governance. It is trust with a governance label" (Puglisi, 2026l). The act frames the whole shift in its Appendix B: "That shift from assurance to evidence is the difference between governance that is designed correctly and governance that functions correctly."

VAISA's own profile system shows where hardware ends and the human begins. Profile 3, Human-Gated Emergency Processing, applies when neither attestation nor sufficient data minimization is available: "A mandatory human pause gate activates. A named human arbiter must review and explicitly approve the inference dispatch" (Puglisi, 2026l). The approval is logged with the arbiter's identity, the timestamp, the data class, the volume of records, and the justification. Each authorization is limited to 72 hours and reviewed within 14 days. Hardware verifies what it can, and where hardware cannot prove the conditions, a named human decides. That is the bridge between Verified AI under Responsible AI and Verified AI under AI Governance.

Hardware has its own boundary. The GOPEL Confidential Processing Extension states it plainly: "Attestation proves the environment, not the runtime behavior" (Puglisi, 2026e). Hardware attestation also moves trust to the silicon vendor and its provisioning chain, which carry their own residual risks. Hardware can make the machine's record tamper-evident under stated trust assumptions, but it cannot create the governor.

14. AI Governance: Agent Models 2 and 3

Agent Model 2 is the hybrid. The agent handles dispatch, collection, and routing, pauses after each functional role, and waits for a named human to approve before the next role begins, with a non-cognitive enforcement layer such as GOPEL holding the pauses in place (Puglisi, 2026i). Like Agent Model 1, it is specified and built as reference code but not deployed. Agent Model 3 runs with no agent at all: the human dispatches, collects, and routes by hand, and it is the model in use today. *Governing AI: When Capability Exceeds Control*, published in November 2025, was produced this way. Its closing note on production records that roles were assigned, checkpoints were logged, dissent was preserved, and human arbitration made the final decisions. It adds that readers "can audit the process and reuse the same structure in their own programs" (Puglisi, 2025d).

Checkpoint-Based Governance supplies the authority in both models, and it works in two layers. In the process layer, the named human directs, challenges, and verifies the work and approves it to continue from one step to the next, an approval that moves the work forward and closes nothing. In the decision layer, the checkpoint closes on accept, modify, or reject, and reject ends the process; the decision produces a record of who decided, on what evidence, and what dissent ran against it (Puglisi, 2026b). The named human stands accountable through four channels no machine component carries: moral, professional, civil, where negligent judgment may result in lawsuit and personal liability, and criminal, where gross negligence may result in prosecution. Human in the loop does not meet this standard, because presence and participation do not require authority or accountability; as a June 2026 analysis put it, "presence is not authority, and participation is not accountability" (Puglisi, 2026n). The same analysis explains why no training regime or prompt closes the gap: "Programming improves the disposition. It does not create the governor."

A named human can also approve without reading, and the framework does not pretend otherwise. Checkpoint-Based Governance treats passive acceptance at the checkpoint as a governance failure and requires it to be detectable rather than merely discouraged (Puglisi, 2026b). That requirement answers the strongest scholarly critique of human oversight. Green (2022) examined 41 government policies requiring human oversight of algorithms and found that people often fail at the oversight those policies assign, which lets the policies legitimize flawed systems. Checkpoint-Based Governance accepts the diagnosis and answers it with authority plus a record: the human holds binding authority, every decision is typed and logged, and the pattern of decisions is itself monitored. The framework is in production across three books, the HAIA frameworks, and published content, and it is not yet independently validated, peer-reviewed, or deployed at enterprise scale.

Under AI Governance, the CARCS record carries constitutional accountability, and its decision types separate a governor who caught an error from one who approved without reading. That is what makes a governed decision attributable rather than merely auditable. The governor the definition calls for is qualified by capability rather than credentials alone, holding the competence to direct, challenge, verify, and own work done with AI without needing to out-know the specialist machine.

HAIA-SCOPE holds the sources in custody. Its premise is direct: "Your citations are only as strong as the record behind them" (Puglisi, 2026k). SCOPE records what an author verified about each cited source at the time of use, and it works in three tiers. A Basic Source Check confirms that the source

exists and matches its citation. A Claim Support Record classifies whether the source supports the claim directly, indirectly, or only as background, and a Full Source Custody Record adds an archive link, a retraction status check, and a diligence note. Its evidentiary hierarchy ranks preservation methods by how independent their timestamps are, from an Internet Archive snapshot the author cannot alter down to a copy-paste note. SCOPE answers the problem of decaying evidence, since a source can fail after good-faith use through retraction, link rot, content drift, or exposed fabrication (Puglisi, 2026d), and a preserved record keeps the original check open to audit. When a cited page later disappears, SCOPE's verdict is exact: "The medium failed. The author did not." It also states its own limit: "The SCOPE record does not immunize an author. It creates reviewable evidence of what the author checked, when the author checked it, and whether the author's reliance was reasonable under the conditions available at the time" (Puglisi, 2026k). SCOPE is human-triggered, and the author decides which tier each citation receives, so it is the one record in the stack that always carries a human verifier.

15. How This Work Defines Verified AI

Section 4 set out the formal meaning: the research goal Seshia, Sadigh, and Sastry named in 2016, designing AI systems with strong, ideally provable assurances of correctness against mathematically specified requirements. The work traced in this part grew from that goal into practice, and for the practitioner, Verified AI now means:

> **Verified AI is AI work whose outputs and decisions leave a record that can be audited and reconstructed, with its evidence rooted in what a human can review and accept, be it in hardware or code.**

Within the HAIA Ecosystem, Verified AI is defined as follows:

> Under Responsible AI, Verified AI is an Agent Model 1 run in which code and guardrails control the system, RECCLIN Reasoning and CAIPR produce the record inside a non-cognitive tool such as GOPEL, and the CARCS record is the proof of control, independent of the operator only when its evidence is rooted in hardware, as VAISA specifies. Under AI Governance, Verified AI is Agent Model 2 or Agent Model 3 work in which a named human directs, challenges, verifies, and owns the output at a checkpoint under Checkpoint-Based Governance, the same record carries constitutional accountability, and SCOPE holds the sources in custody. Ethical AI sets the criteria the record is checked against. In every case, the proof is proof only when someone checks it, and that someone is human.

Three words in these definitions carry specific weight. A record is auditable when a third party can inspect it, and that holds in both tiers, because both leave a record. A decision is attributable when a named human stands behind it, and that belongs to AI Governance alone. A decision path is reconstructable when the author can explain the work, defend it, and rebuild the decision path and its evidentiary basis. That test does not extend to reproducing a model's internal reasoning, which is frequently impossible (Puglisi, 2026h). Throughout, machine verification and the named human's

verification stay distinct: machines produce and check evidence, and the human verifies that evidence and decides.

16. The Dated Record of This Work

The dates below come from public pages, a published book, and public repositories. They show the order of publication and nothing more.

Date	Public artifact	What the artifact proposed, described, or documented
February 1, 2024	"Factics Make Us More Intelligent," basilpuglisi.com	A second platform for source validation, with neither platform treated as an authority
September 23, 2025	Checkpoint-Based Governance implementation framework, basilpuglisi.com	Human arbitration at defined checkpoints; automation bias drift named as a governance failure
October 30, 2025	"The Case for AI Provider Plurality in Evidence-Based Research," basilpuglisi.com	The provider plurality argument on which GOPEL was built
November 2025	*Governing AI: When Capability Exceeds Control*	A book produced under manual human governance, closing with an account of its own production
December 1, 2025	Checkpoint-Based Governance constitution, GitHub	The constitutional text of the checkpoint architecture
December 21, 2025	Kimi case study, GitHub	A published governance case record
February 3 and 4, 2026	Agent architecture papers and specification, GitHub	Three operating models, with automated agent runs defined as Responsible AI and automation bias detection signals specified
February 17, 2026	Prior art provenance record and Integration Gap Evidence Record, GitHub	Dated custody of prior-art evidence
February 23, 2026	GOPEL reference code and "GOPEL: The Code Behind the Policy," GitHub and basilpuglisi.com	Non-cognitive enforcement built as software, with the human checkpoint named as the defense against deception
March 6, 2026	Verified AI Inference Standards Act, basilpuglisi.com	Per-transaction hardware attestation at the inference boundary, with a human gate where attestation is unavailable
April 23, 2026	HAIA-CARCS, basilpuglisi.com	The governed session record and its decision types
June 2, 2026	HAIA-SCOPE, basilpuglisi.com	Source custody for published work
June 12, 2026	"Why You Cannot Program or Prompt Governance Into AI," basilpuglisi.com	The case that governance lives outside the model

Sources for the table: Puglisi (2024, 2025a, 2025b, 2025c, 2025d, 2026a, 2026f, 2026g, 2026i, 2026j, 2026k, 2026l, 2026n).

The CARCS and SCOPE entries mark when each practice received a published name, not when the practice began. The book's production account in November 2025, the case study and raw-data preservation records that followed, and the prior-art custody records of February 2026 show both practices running for months before either was formalized.

Part III: The 2026 Trend and Proof-of-Control

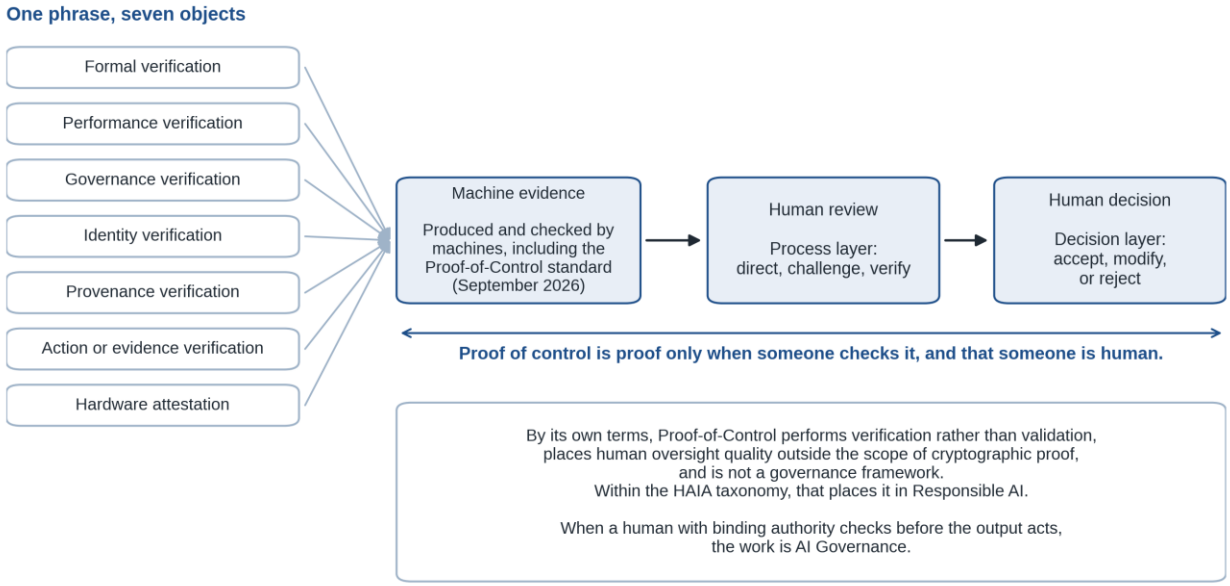

Figure 3. The 2026 verification trend and where proof of control is completed. Compiled from the works cited in Part III.

17. Verification Language in 2026

Verification language became especially visible during 2026, across standards bodies, auditors, and vendors. In January, the AI Verification and Evaluation Research Institute launched to press for independent audits of frontier AI developers (Kahn, 2026). On February 6, the ISO/IEC draft technical specification 42119-3, which addresses verification and validation analysis of AI systems, was registered for formal approval. It was still listed as a draft technical specification in September (International Organization for Standardization & International Electrotechnical Commission, 2026).

On August 7, 2026, NIST released the initial public draft of NIST AI 200-2, the TEVV-Athlon Framework for Evaluating AI Systems, with comments open through October 6, 2026 (Phillips et al., 2026). The first item in its request for input concerns the definitions and uses of the terms testing, evaluation, verification, and validation. Nearly four decades after the first verification and validation meetings for intelligent systems, a leading standards institution is still asking where the boundaries belong. The sections that follow show how far the word has traveled while that question stays open.

18. Independent Performance Verification

Commercial assurance has taken up the term as well, in the form of independent confirmation of performance claims. The British Standards Institution offers an AI performance assessment that independently verifies whether an AI system or component meets the performance, bias, and robustness claims its manufacturer makes, measured against metrics drawn from international standards (British Standards Institution, n.d.). The same service page carries the formal-methods definition of Verified Artificial Intelligence nearly word for word, a clear example of the term stretching from mathematical proof to commercial measurement.

This is not formal verification in the Seshia sense, yet it is recognizable verification with a different object. Instead of asking whether a system mathematically satisfies a formal property, the verifier asks whether the system achieves the accuracy its developer claims or keeps observed bias within a defined threshold.

Performance verification also exposes a feature that recurs throughout AI verification: its conclusions can expire. A model update, data shift, prompt change, retrieval change, new integration, or altered operating environment can make earlier evidence unrepresentative of current behavior, which is why performance verification increasingly meets continuous monitoring rather than one-time certification.

19. Agentic AI Creates a New Verification Problem

AI agents produce the sharpest change in how the word "verified" is used. An agent that interacts with websites, application programming interfaces, software tools, payment systems, or other agents raises a basic identity problem, because another system needs to know which agent is making a request.

HUMAN Security released its open-source HUMAN Verified AI Agent project on July 21, 2025 to show cryptographically authenticated communication between agents and digital services (Diamant & Elias, 2025). The project uses HTTP Message Signatures so that a gateway can confirm which agent sent a given request. The object verified here differs fundamentally from the object in formal Verified AI. The verification establishes that a communication was authenticated as coming from a particular agent. It does not establish that the agent is correct, that it is safe, that its reasoning is sound, that it holds authority for every decision it attempts, or that its objective is desirable. Identity verification matters, but it is not behavioral verification.

A similar model appears in decentralized agent infrastructure. Concordium launched its Agent Registry in late May 2026, giving agents an on-chain identity tied to a verified human or organizational owner and offering a "Verified by Concordium" badge (Concordium, 2026a). Concordium describes the limit of that badge plainly: it tells a counterparty that a verified person or business answers for the agent, not that the agent will behave (Concordium, 2026b). That candor makes the point this paper develops. In the 2016 formal-methods sense, verification asks whether a system can be shown to satisfy a requirement, while in the agent-identity sense it asks which agent this is and who stands behind it. Both questions matter, and answering one does not answer the other.

20. Provenance Creates Yet Another Verification Layer

Generative AI has also turned provenance into a verification problem. The Coalition for Content Provenance and Authenticity maintains the Content Credentials specification, at version 2.4 as of April 2026 (Coalition for Content Provenance and Authenticity, 2026). The specification defines a model for storing and accessing cryptographically verifiable information about where digital content came from and how it changed. The object verified is not the truth of the content; it is information associated with the content's history.

That boundary grows more important as generative media spreads. On May 19, 2026, OpenAI announced conformance with the C2PA standard, SynthID watermarking for images through a

partnership with Google, and a preview of a public verification tool. On July 31, 2026, it extended the watermarking to audio and opened API access for verification (OpenAI, 2026). OpenAI also cautions that metadata can be stripped, lost through uploads, or broken by transformations such as screenshots, so provenance signals cannot carry the full weight of trust on their own.

The distinction illustrates the larger argument of this paper. Verification may establish that a provenance record is cryptographically bound to an asset, but it does not establish that a claim inside the asset is true. Likewise, confirming that media came from a particular AI system does not establish that the media depicts an event accurately, that rights were properly held, that the surrounding context is accurate, or that the content was never used to deceive. Provenance verification is another bounded claim.

21. One Phrase, Multiple Objects

By 2026, verification language around AI can be organized into at least seven distinct categories.

Verification domain	Primary object	Typical question	Evidence or method
Formal verification	System or component property	Does property P hold under specification S?	Mathematical specification, theorem proving, model checking, formal analysis
Performance verification	Performance claim	Does the system perform as claimed?	Testing, ground truth, metrics, independent assessment
Governance verification	Process or control	Has the system or organization met defined governance criteria?	Documentation, technical tests, process evidence, audits
Identity verification	AI agent or operator	Is this agent who it claims to be, and who stands behind it?	Cryptographic signatures, registries, identity credentials
Provenance verification	Digital asset or output	What can be established about where this content came from and how it changed?	Cryptographic manifests, Content Credentials, watermarking
Action or evidence verification	Agent action or transaction	Can evidence establish that a claimed action occurred within defined conditions?	Logs, signed records, proofs, attestations, transaction evidence
Hardware attestation	Execution environment and evidence key	Did the run happen in an approved, unaltered environment, and could the operator have forged the record?	Hardware-signed attestation quote, enclave measurement, freshness nonce, TCB version, signed processing receipt

These categories overlap, but they should not be collapsed. A formally verified component may lack verified provenance, a cryptographically authenticated agent may run an unverified model, and a system may pass a governance assessment while offering no mathematical guarantee about its behavior. An independently verified performance figure may go stale after a model update, and AI-generated media may carry valid provenance data while making a false assertion. The last row asks a question the others do not: whether the evidence is bound to a measured execution environment and to key material the operator does not hold. Unless a claim names its object and scope, the phrase "verified AI" says very little.

22. The Major Conflicts Behind "Verified AI"

The contemporary debate is not about whether verification is useful. The substantive disagreements concern what evidence justifies the term, how far its conclusions reach, and what a verified status permits others to infer.

Proof Versus Evidence

The formal-methods tradition places mathematical proof near the center of verification, while contemporary assurance often relies on empirical testing, benchmarks, audits, red teaming, conformity assessment, documentation, simulation, or statistical analysis. These methods produce evidence of different strength. Calling all of them "verification" without qualification erases real differences between proof, measurement, observation, and procedural assurance.

Component Versus System

AI rarely operates alone. A deployed system may combine a foundation model, system prompts, retrieval infrastructure, databases, software tools, external interfaces, security policies, human operators, and organizational processes, and a verified component does not make a verified system. The ISO/IEC draft technical specification 42119-3 reflects this wider view. It addresses verification and validation analysis for AI systems that include both AI components and their interaction with non-AI components, using formal methods, simulation, and evaluation (International Organization for Standardization & International Electrotechnical Commission, 2026).

Correctness Versus Trustworthiness

Verification establishes evidence relative to criteria, and trustworthiness is broader. It depends on safety, security, privacy, fairness, reliability, accountability, transparency, legality, human control, institutional legitimacy, and fitness for a particular context, and no single verification result resolves all of those questions.

Self-Verification Versus Independent Verification

Who performs the verification is its own conflict. A developer may produce technically valid evidence about its own system, yet a system in which the developer defines the requirements, chooses the measures, runs the tests, interprets the evidence, and grants itself a "verified" label carries an obvious independence problem. The 2026 push for independent audits of frontier developers turns on this point (Kahn, 2026). Independence alone does not establish quality, however, since a weak criterion checked independently remains a weak criterion. The useful question asks who defined the claim, who selected the evidence, what method was used, what conflicts existed, and what conclusion the evidence actually justifies.

Static Versus Continuing Verification

Traditional certification suggests a status that holds for a period, and modern AI strains that assumption. Models change, providers adjust behavior, retrieval corpora shift, plugins and tools change, threats evolve, user populations change, data distributions drift, and agents gain new permissions, so verification is increasingly bound to a time and a configuration. A system verified as version 2.1 in environment E on a given date has not thereby been verified as version 2.4 running

with different tools. The evidence itself can decay as well. A source that supported a claim on the day it was checked may later be retracted, edited, or deleted, which leaves the original verification impossible to audit unless someone preserved what was checked.

23. Verification Is Not a Synonym for Truth

Perhaps the most important boundary runs between verification and truth, because verification always needs both an object and a proposition. One can verify identity without verifying behavior, provenance without verifying factual accuracy, and performance without verifying fairness. One can verify compliance with a process without verifying the wisdom of that process, and one can formally verify that a system satisfies a specification without showing that the specification is socially, ethically, or legally adequate.

The boundary matters more as verification enters public-facing trust systems. A user who sees a badge reading "Verified AI" may infer that the system is safe, accurate, authorized, fair, secure, compliant, answerable to a person, or simply trustworthy, while the underlying check established only one of those things. That distance between what verification technically shows and what the label leads people to believe is an assurance gap. The proper response is not to abandon verification but to make verification claims legible.

24. From Models to Proof Relationships

The history traced here suggests a wider transition. Early knowledge-based-system verification focused on the internal consistency and validity of intelligent systems. Formal Verified AI then carried mathematical verification into machine learning and AI-enabled systems, and modern assurance widened the analysis to technical and organizational risk. Generative AI made provenance a verification problem, and agents added identity, authorization, and evidence of action.

Verification is therefore becoming less useful as an adjective attached to a system and more useful as a relationship between a claim and its evidence. The essential form is no longer "This is a verified AI." It is "This claim about this AI was verified against these criteria, using this evidence and method, by this verifier, within this scope, at this point in time, and checked by this person." That formulation is less marketable, and it is far more informative.

25. Proof-of-Control: Who Shares It and What It Claims

The most visible 2026 effort to claim the language of verification came in September. On September 17, 2026, the Advanced AI Society announced its membership in the Linux Foundation and LF Decentralized Trust, which hosts the standard as a lab (Advanced AI Society, 2026a). The same day, it released the v1.0 working draft of Proof-of-Control for public comment through October 30, 2026. The alliance described the draft as co-designed with more than 80 global security leaders, published endorsements from security, finance, and policy figures alongside it, and held a public launch event on September 23. The standard's repository began on August 4, 2026, and its release file targets a stable version on February 1, 2027 (Advanced AI Society, 2026b).

What it claims is open verification of AI agents. The release rests the standard on the principle that verification cannot belong to the party being verified. Its co-chair described the method as defining an agent's authority before it acts, producing evidence at runtime that the agent stayed inside that authority, and keeping that evidence auditable across the agent's lifecycle (Advanced AI Society, 2026a). The standard grades evidence in four tiers by who must be trusted to believe it: assertion, attestation, trust-minimized evidence that anyone can verify, and self-enforcing evidence without which the system cannot run. A system has Proof-of-Control only when its evidence reaches the third or fourth tier, and the standard protects a certification mark, "Proof-of-Control Certified," so that only systems assessed as conformant may claim it (Advanced AI Society, 2026b).

26. How Proof-of-Control Contrasts

The standard arrives late to the Verified AI conversation. Part I traced that conversation to verification and validation meetings in 1988 and to the formal definition of 2016, and Part II dated the elements of Verified AI in this work to publications between February 2024 and June 2026. Several elements the standard relies on appeared in that record before its repository began. Human checkpoint authority, with automation bias drift named as a governance failure, was published in September 2025, and automated agent runs were defined as Responsible AI in February 2026, the same month non-cognitive enforcement code appeared. Per-transaction hardware attestation at the inference boundary was proposed in March 2026, and a governed session record followed in April 2026. The claim here is the order of publication. Attestation, hashing, and signatures are older than all of this work. VAISA itself states that it requires no new science, and no derivation is claimed in either direction.

The standard also stops short of the proof stage, and its own text shows where. It describes itself as performing verification rather than validation, showing that an agent stayed inside the boundaries that were set while leaving the judgment of whether those boundaries were right, including human oversight, as a human responsibility. Its tiers grade the evidence rather than the controls, and the draft acknowledges that a self-enforcing system can perfectly enforce poorly chosen controls. Its catalog of proof mechanisms states that "human oversight quality is outside the scope of cryptographic proof," and its overview states that it is not a governance or risk framework (Advanced AI Society, 2026b). The draft even lists as an open question whether the required human authorization belongs in the root of trust of its highest tier. By its own terms, Proof-of-Control is not a governance framework; within the HAIA taxonomy, that places it in Responsible AI.

The launch carried the counterargument openly. Hart Montgomery, chief technology officer of LF Decentralized Trust, argued that human inspection fails when non-deterministic machines decide in milliseconds and that only automated tools can keep pace. He also conceded that cryptography cannot solve every security problem agents raise (Advanced AI Society, 2026a). For the evidence itself, that argument holds, and this paper does not dispute it.

Formal methods, cryptography, and remote attestation use the words proof and verification for machine operations, and this paper does not dispute that usage either. A proof checker confirms a derivation, an attestation verifier appraises evidence against a policy, and a gateway checks a signature, often with no person watching the individual check. Even the RATS architecture, built for

exactly that automation, assigns the appraisal policy to a verifier owner and the choice to rely on the result to a relying party (Birkholz et al., 2023). What those operations produce is evidence. Checkpoint-Based Governance separates what a person does with that evidence into two layers: in the process layer, the named human directs, challenges, and verifies the work, including the machine's evidence; in the decision layer, that human accepts, modifies, or rejects the output and owns the result. Proof of control, as this paper uses the term, is the record that reaches that person. In the HAIA definition, that record is the CARCS record, and the definition's closing sentence sets the same condition: the proof is proof only when someone checks it.

However well the code runs or the guardrails hold, proof of control is proof only when someone checks it, and that someone is human. When that human holds binding authority before the output acts, the work is AI Governance. When the human only audits after the fact, the work remains Responsible AI: verified, but ungoverned. The same holds for hardware, since a quote the silicon signs still has to be checked by someone who answers for the decision it supports.

27. Where the Gap Remains

Two claims in this paper assert that something is missing, and each rests on searches run on September 23, 2026 and states only what those searches found.

The first concerns hardware proof at the inference boundary. VAISA stated in March 2026 that some providers had begun offering confidential computing features for specific workloads, and that the market still lacked "a standardized, customer-callable, per-transaction inference evidence artifact across mainstream managed LLM APIs" (Puglisi, 2026l). The market has moved in the direction VAISA named. Microsoft announced a confidential inferencing preview for its Azure OpenAI Whisper speech model in September 2024, one of the partial features VAISA acknowledged, and its documentation still describes that preview for Whisper in June 2026 (Microsoft, 2024, 2026). Google Cloud released toolkits in June 2026 that let customers establish attested sessions with inference servers they run inside trusted execution environments (Google Cloud, 2026). Specialized providers now offer confidential inference for open-weight models with remote attestation that clients can check (Phala, 2026; Privatemode, n.d.). Apple's Private Cloud Compute shows how far device-side verification has come, since Apple's devices send data only to nodes that can attest to running publicly listed software (Apple Security Research, 2024). An independent analysis reports, however, that it offers no interface for third parties to build on (Dittmar et al., 2026). As of September 23, 2026, customer-verifiable hardware attestation was found on specialized open-weight APIs, in Azure OpenAI's confidential inferencing preview for its Whisper speech model, and in infrastructure toolkits, but no standardized, customer-callable, per-transaction attestation artifact was found on the mainstream frontier chat APIs. Some have implemented the capability; the standard VAISA called for still needs all of them.

The second concerns the quality of human oversight. The broad claim that no one measures whether an overseer actually exercises judgment does not hold. Human-factors research on automation bias spans decades. An August 2026 proposal, the Oversight Quality Index, measures oversight through override rates, catch rates, engagement decay, intervention latency, and confidence calibration, and it describes oversight as "widely mandated but rarely measured" (Ganjihal & Singh, 2026). The

narrower claim survives. Machine-verification evidence leaves oversight quality out by design, as the Proof-of-Control scope statement in Section 26 shows. Checkpoint-Based Governance named automation bias drift as a governance failure in September 2025. The HAIA agent architecture then specified detection signals for it inside a checkpoint architecture in February 2026, both before the Oversight Quality Index was proposed (Puglisi, 2025c, 2026i). The gap in the Verified AI conversation is not measurement in general; it is the absence of oversight quality from the evidence that machine-verification standards produce.

28. Conclusion

"Verified AI" has a history, a place in this work, and a 2026 trend, and the three should not be confused. The history does not begin with large language models, generative AI, or autonomous agents. Verification and validation of intelligent systems were established research concerns by the late 1980s, and formal methods supplied a parallel tradition of specifying and verifying complex systems. In 2016 the two converged in a clear formulation of Verified Artificial Intelligence as the pursuit of strong, ideally provable assurances of correctness, and that formulation should keep its historical meaning.

The concept shaped the work examined in Part II long before the term trended. Checking AI output against a second platform in 2023 grew into structured reasoning records, checkpoint authority, hardware-rooted evidence, and governed session records. From that practice comes this work's definition: AI work whose outputs and decisions leave a record that can be audited and reconstructed, with its evidence rooted in what a human can review and accept, be it in hardware or code. Under Responsible AI, an automated run produces a CARCS record that serves as proof of control, independent of the operator only when hardware stands behind its evidence. Under AI Governance, a named human with binding authority checks the work at a checkpoint, the same record carries constitutional accountability, and the sources stay in custody.

The 2026 trend widened the word across standards, audits, agent identity, and provenance, and each use verified something bounded. A verified identity is not a verified behavior, a verified provenance record is not verified truth, and a verified performance figure is not universal safety. Proof-of-Control shows the pattern most clearly: it builds strong evidence of what a machine did, arrives late to a conversation decades old, and by its own terms leaves the governor to someone else. The good news is that the conversation is still ongoing and open to the floor. You can go comment if you agree or disagree. The most important part of open-source work is the willingness to give feedback and share, and how it is received is in the hands of the community as much as leadership.

As artificial intelligence moves from models that generate outputs toward agents that act, verification will matter more, not less. Its credibility will depend on keeping its claims bounded and on remembering who stands at the end of every proof. The useful question is what claim was verified, about what object, against what criteria, using what evidence, by whom, for how long, and which person checked it and answers for what happens next.

Author's Final Note

While the author has made his position open and public that Responsible AI, or automation, has a place and will likely be used in risk assessment decisions for financial reasons where the risks do not outweigh the rewards, he could imagine no bigger misrepresentation of the concept than claiming proof of control over AI without human oversight and accountability.

References

Advanced AI Society. (2026a, September 17). *Advanced AI Society joins the Linux Foundation, launches open verification ecosystem as Congress moves on agent security* [Press release]. GlobeNewswire. https://www.globenewswire.com/news-release/2026/09/17/3364426/0/en/advanced-ai-society-joins-the-linux-foundation-launches-open-verification-ecosystem-as-congress-moves-on-agent-security.html

Advanced AI Society. (2026b). *Open verification: The Proof-of-Control standard for agents* (Working draft; README.md, RELEASE.md, 0.1/en/0x10-C07-Evidence-Generation-and-Properties.md, 0.1/en/0x10-C08-Verifiability-Tiers.md, and 0.1/en/0x91-Appendix-B_Proof-Mechanism-Inventory.md at commit 22c7b625be459f5eee7dd8690afd080b5141b8c6). GitHub. https://github.com/LFDT-ProofOfControl/ov-poc-standard/tree/22c7b625be459f5eee7dd8690afd080b5141b8c6

AI Verify Foundation. (n.d.). *Frequently asked questions*. Retrieved September 23, 2026, from https://aiverifyfoundation.sg/faq/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety* (arXiv:1606.06565). arXiv. https://doi.org/10.48550/arXiv.1606.06565

Antoniou, G., van Harmelen, F., Plant, R., & Vanthienen, J. (1998). Verification and validation of knowledge-based systems: Report on two 1997 events. *AI Magazine, 19*(3), 123. https://doi.org/10.1609/aimag.v19i3.1400

Apple Security Research. (2024). *Private Cloud Compute: A new frontier for AI privacy in the cloud*. https://security.apple.com/blog/private-cloud-compute/

Association for the Advancement of Artificial Intelligence. (1993). *Validation and verification of knowledge-based systems: Papers from the 1993 workshop* (Technical Report WS-93-05). https://aaai.org/proceeding/ws93-05/

Birkholz, H., Thaler, D., Richardson, M., Smith, N., & Pan, W. (2023). *Remote ATtestation procedureS (RATS) architecture* (RFC 9334). Internet Engineering Task Force. https://doi.org/10.17487/RFC9334

British Standards Institution. (n.d.). *AI performance*. Retrieved September 23, 2026, from https://www.bsigroup.com/en-US/products-and-services/standards/ai-performance/

Carnegie Mellon University. (2016, June 22). *Workshop explores how artificial intelligence can be engineered for safety and control*. https://www.cmu.edu/news/stories/archives/2016/june/AI-workshop.html

Clarke, E. M., Wing, J. M., et al. (1996). Formal methods: State of the art and future directions. *ACM Computing Surveys, 28*(4), 626-643. https://doi.org/10.1145/242223.242257

Coalition for Content Provenance and Authenticity. (2026). *Content Credentials: C2PA technical specification* (Version 2.4). https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html

Concordium. (2026a, May 28). *The Concordium Agent Registry is live: The accountability layer ERC-8004 doesn't have*. https://www.concordium.com/article/the-concordium-agent-registry-is-live-the-accountability-layer-erc-8004-doesnt-have

Concordium. (2026b, June 25). *Worldcoin proves you're human; Concordium keeps you accountable*. https://www.concordium.com/article/worldcoin-vs-concordium

Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Barrett, C., Zhao, D., Zhi-Xuan, T., Wing, J., & Tenenbaum, J. (2024). *Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems* (arXiv:2405.06624). arXiv. https://arxiv.org/abs/2405.06624

Department for Science, Innovation and Technology. (2024, February 12). *Introduction to AI assurance*. GOV.UK. https://www.gov.uk/government/publications/introduction-to-ai-assurance

Diamant, B., & Elias, T. (2025, July 21). *Introducing HUMAN Verified AI Agent: An open-source foundation for trustworthy agent identity*. HUMAN Security. https://www.humansecurity.com/learn/blog/human-verified-ai-agent-open-source/

Dittmar, Y., Stephan, M. J., Völkl, T., Hollick, M., & Classen, J. (2026). Unlocking Apple's Private Cloud Compute: An analysis of privacy-preserving artificial intelligence. In *Proceedings of the 19th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '26)*. https://doi.org/10.1145/3765613.3811691

Dreossi, T., Fremont, D. J., Ghosh, S., Kim, E., Ravanbakhsh, H., Vazquez-Chanlatte, M., & Seshia, S. A. (2019). VerifAI: A toolkit for the formal design and analysis of artificial intelligence-based systems. In I. Dillig & S. Tasiran (Eds.), *Computer aided verification: CAV 2019* (pp. 432-442). Springer. https://doi.org/10.1007/978-3-030-25540-4_25

Ganjihal, S. R., & Singh, S. (2026, August 26). *Measuring oversight: The Oversight Quality Index and Cognitive Readiness Score for AI-augmented decision systems*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7376519

Gonzalez, A. J., & Barr, V. (2000). Validation and verification of intelligent systems: What are they and how are they different? *Journal of Experimental & Theoretical Artificial Intelligence, 12*(4), 407-420. https://doi.org/10.1080/095281300454793

Google Cloud. (2026, June 23). *Verifiable trust in the AI era: What's new in Confidential Computing*. https://cloud.google.com/blog/products/identity-security/verifiable-trust-in-the-ai-era-whats-new-in-confidential-computing

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review, 45*, 105681. https://doi.org/10.1016/j.clsr.2022.105681

Infocomm Media Development Authority. (2022, May 25). *Singapore launches world's first AI testing framework and toolkit to promote transparency* [Press release]. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2022/sg-launches-worlds-first-ai-testing-framework-and-toolkit-to-promote-transparency

International Organization for Standardization & International Electrotechnical Commission. (2026). *Artificial intelligence, testing of AI, Part 3: Verification and validation analysis of AI systems* (ISO/IEC DTS 42119-3.2) [Draft technical specification]. https://www.iso.org/standard/85072.html

Kahn, J. (2026, January 15). Exclusive: Former OpenAI policy chief creates nonprofit institute, calls for independent safety audits of frontier AI models. *Fortune*. https://www.fortune.com/2026/01/15/former-openai-policy-chief-creates-nonprofit-institute-calls-for-independent-safety-audits-of-frontier-ai-models

Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). Reluplex: An efficient SMT solver for verifying deep neural networks. In R. Majumdar & V. Kunčak (Eds.), *Computer aided verification: CAV 2017* (pp. 97-117). Springer. https://doi.org/10.1007/978-3-319-63387-9_5

Microsoft. (2024, September 24). *Azure AI confidential inferencing preview*. Microsoft Tech Community. https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/azure-ai-confidential-inferencing-preview/4248181

Microsoft. (2026). *Azure confidential computing products*. Microsoft Learn. Retrieved September 23, 2026, from https://learn.microsoft.com/en-us/azure/confidential-computing/overview-azure-products

Mosqueira-Rey, E., & Moret-Bonillo, V. (2000). Validation of intelligent systems: A critical study and a tool. *Expert Systems with Applications, 18*(1), 1-16. https://doi.org/10.1016/S0957-4174(99)00045-7

O'Leary, D., & Preece, A. (Cochairs). (1998). *Verification and validation of knowledge-based systems: Papers from the 1998 AAAI workshop* (Technical Report WS-98-11). AAAI Press.

OpenAI. (2026, May 19; updated July 31). *Advancing content provenance for a safer, more transparent AI ecosystem*. https://openai.com/index/advancing-content-provenance/

Personal Data Protection Commission Singapore. (2022, May 25). *Launch of AI Verify: An AI governance testing framework and toolkit*. https://www.pdpc.gov.sg/News-and-Events/Announcements/2022/05/Launch-of-AI-Verify---An-AI-Governance-Testing-Framework-and-Toolkit

Phala. (2026). *Confidential AI models: Private LLM API on TEE*. Retrieved September 23, 2026, from https://phala.com/confidential-ai-models

Phillips, P. J., Jensen, T., Hall, P., Amironesei, R., Choong, Y.-Y., Greenberg, C., & Greene, K. K. (2026). *The TEVV-Athlon framework for evaluating AI systems* (NIST AI 200-2 ipd). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.200-2.ipd

Privatemode. (n.d.). *Private AI API with end-to-end encryption*. Retrieved September 23, 2026, from https://www.privatemode.ai/inference-api

Puglisi, B. C. (2024, February 1). *Factics make us more intelligent*. https://basilpuglisi.com/factics-make-us-more-intelligent/

Puglisi, B. C. (2025a, October 30). *The case for AI provider plurality in evidence-based research*. https://basilpuglisi.com/the-case-for-ai-provider-plurality-in-evidence-based-research/

Puglisi, B. C. (2025b, December 1). *Checkpoint-based governance: A constitution for human-AI collaboration*. GitHub. https://github.com/basilpuglisi/HAIA/blob/4f38c6997f61042bcc731bceeb76124568bba271/Checkpoint-Based%20Governance%20A%20Constitution%20for%20Human-AI%20Collaboration%20v4.2.1.docx

Puglisi, B. C. (2025c, September 23). *Checkpoint-based governance: An implementation framework for accountable human-AI collaboration*. https://basilpuglisi.com/checkpoint-based-governance-an-implementation-framework-for-accountable-human-ai-collaboration-v2-drafting/

Puglisi, B. C. (2025d). *Governing AI: When capability exceeds control*. Digital Ethos. ISBN 9798349677687.

Puglisi, B. C. (2026a, April 23). *CARCS: Compliance accountability record & case study*. https://basilpuglisi.com/haia-carcs-compliance-accountability-record-case-study/

Puglisi, B. C. (2026b, September 11). *Checkpoint-Based Governance: Presence is not authority*. https://basilpuglisi.com/cbg/

Puglisi, B. C. (2026c, May 22). *Compliance accountability record & case study* [Revised working paper]. GitHub. https://github.com/basilpuglisi/HAIA/blob/4f38c6997f61042bcc731bceeb76124568bba271/COMPLIANCE%20ACCOUNTABILITY%20RECORD%20%26%20CASE%20STUDY%20CARCSv1.4.pdf

Puglisi, B. C. (2026d). *Fault-based publication ethics: The case for source custody in an era of AI citation contamination* [Working paper]. SSRN. https://papers.ssrn.com/abstract=6872038

Puglisi, B. C. (2026e, March). *GOPEL confidential processing extension*. GitHub. https://github.com/basilpuglisi/HAIA/blob/4f38c6997f61042bcc731bceeb76124568bba271/haia_agent/GOPEL_Confidential_Processing_Extension_CPE_v1_1.md

Puglisi, B. C. (2026f, February 23). *GOPEL: The code behind the policy*. https://basilpuglisi.com/gopel-the-code-behind-the-policy/

Puglisi, B. C. (2026g). *HAIA* [Source code and documents, commit 4f38c6997f61042bcc731bceeb76124568bba271]. GitHub. https://github.com/basilpuglisi/HAIA

Puglisi, B. C. (2026h, September). *HAIA-CAIPR: Cross AI Platform Review* (Fourth ed.). Zenodo. https://doi.org/10.5281/zenodo.22710389

Puglisi, B. C. (2026i, February 4). *HAIA-RECCLIN agent architecture specification: EU compliance version*. GitHub. https://github.com/basilpuglisi/HAIA/blob/4f38c6997f61042bcc731bceeb76124568bba271/HAIA_RECCLIN_Agent_Architecture_Specification_v2.2_EU_Compliance_Version%20(1).docx

Puglisi, B. C. (2026j, February 17). *Prior art provenance record and integration gap evidence record*. GitHub. https://github.com/basilpuglisi/Prior-Work-Communication-

Puglisi, B. C. (2026k, June 2). *SCOPE: Source custody observable publication evidence*. https://basilpuglisi.com/scope-source-custody-observable-publication-evidence/

Puglisi, B. C. (2026l, March 6). *Verified AI Inference Standards Act* (AI Provider Plurality Congressional Package, Document 5 of 5). https://basilpuglisi.com/wp-content/uploads/2026/03/Verified-AI-Inference-Standards-Act-or-VAISA.pdf

Puglisi, B. C. (2026m, September 11). *What is HAIA-RECCLIN? Reasoning and dispatch*. https://basilpuglisi.com/haia-recclin/

Puglisi, B. C. (2026n, June 12). *Why you cannot program or prompt governance into AI*. https://basilpuglisi.com/program-prompt-governance-ai/

Seshia, S. A., Sadigh, D., & Sastry, S. S. (2016). *Towards verified artificial intelligence* (arXiv:1606.08514). arXiv. https://doi.org/10.48550/arXiv.1606.08514

Seshia, S. A., Sadigh, D., & Sastry, S. S. (2022). Toward verified artificial intelligence. *Communications of the ACM, 65*(7), 46-55. https://doi.org/10.1145/3503914

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
