Topics: ORIGINAL RESEARCH

Category: INTERDISCIPLINARY

Truth-ε and machine knowledge: from demonstrative rationality to bounded computational convergence

Sakos Ikonomopoulos

Aikatarini Lascaridis Foundation, Piraeus, Greece

ABSTRACT

This paper introduces the concept of Truth-ε as a framework for understanding the epistemological status of contemporary artificial intelligence. Modern AI systems increasingly produce reliable and scientifically useful outputs while remaining only partially reconstructible through explicit symbolic reasoning. This raises a central question: how can machine-generated representations count as knowledge if they are neither exact copies of reality nor demonstrative conclusions derived from transparent logical chains? Truth-ε refers to a form of epistemic reliability achieved through convergence under conditions of finite information, uncertainty, and computational limitation. The paper argues that AI does not abolish rationality or reduce truth to mere prediction. Instead, it makes explicit a conception of rationality already embedded in the history of mathematics and science, where knowledge advances through controlled approximation, convergence, and bounded error. From the method of exhaustion and ε–δ analysis to probability theory, information theory, computational complexity, and machine learning, scientific knowledge has evolved by disciplining error rather than eliminating it entirely. Within this framework, AI systems derive epistemic legitimacy through robustness, calibration, generalization, reproducibility, and explicit disclosure of epistemic limits.

Key words: Truth-ε; artificial intelligence; machine knowledge; computational epistemology; bounded convergence; demonstrative rationality.

Correspondence: Sakos Ikonomopoulos, Aikatarini Lascaridis Foundation, Praxitelous 169 & Bouboulinas Streets, 18535 Piraeus, Greece.
E-mail: sakos.ikonomopoulos@gmail.com

Competing interest: the author declares that they have no competing interests.

Ethics approval and consent to participate: not applicable.

Availability of data and materials: not applicable.

Received: 17 June 2026.
Accepted: 16 July 2026.

Licensee PAGEPress, Italy
Proceedings of the European Academy of Sciences & Arts 2026; 5:84
doi: 10.4081/peasa.2026.84

Introduction: the epistemic problem of AI

Artificial intelligence has become one of the principal sites at which the traditional criteria of knowledge are being renegotiated. Contemporary learning systems generate classifications, predictions, translations, recommendations, images, texts, and scientific hypotheses that are often reliable enough to be used in domains of practical and theoretical importance. Their outputs may guide medical diagnosis, assist legal reasoning, support molecular discovery, model climate processes, identify statistical regularities in large datasets, and participate in everyday linguistic interaction. Yet the epistemic status of such outputs remains uncertain. They appear to function as knowledge, or at least as knowledge-like representations, while lacking several features traditionally associated with knowledge: explicit justification, transparent inference, and reconstructible chains of reasoning (LeCun *et al.,* 2015; Goodfellow *et al.,* 2016).

The issue is not simply that artificial intelligence systems sometimes make mistakes. All epistemic systems, human and artificial, are fallible. The more fundamental issue is that contemporary AI systems can be successful precisely in cases where their internal representational structures are not available as explicit reasons. A deep neural network may classify an image correctly, translate a sentence fluently, or identify a protein structure with high reliability, while no human observer can fully reconstruct the internal path through which the system arrived at that result. The system is mathematically specified at the level of architecture, training procedure, and parameter update; nevertheless, the learned representation distributed across millions or

billions of parameters resists interpretation as a finite symbolic proof or as a transparent causal explanation.

This produces a tension between operational reliability and epistemic intelligibility. On the one hand, AI systems may achieve impressive empirical performance. On the other hand, their success often depends on forms of statistical learning that do not yield explanations in the classical sense. The problem is therefore not merely whether AI systems are accurate, but what kind of epistemic validity accuracy can provide when it is produced by opaque processes of optimization. If a system reliably tracks patterns in the world but cannot provide an explicit account of why those patterns hold, should its outputs be considered knowledge, prediction, instrumentally useful behavior, or something else?

Much of the contemporary debate has addressed this difficulty through the concepts of opacity, interpretability, and explainability. These concepts are indispensable. They identify real risks associated with machine learning systems: hidden bias, lack of accountability, fragility under distributional shift, difficulty of auditing, and the possibility of confident error. However, the debate often presupposes that the ideal epistemic condition would be one in which the model's internal operations could be translated into humanly understandable reasons. Explainability is then framed as the recovery, approximation, or simulation of a lost transparency (Burrell, 2016; Lipton, 2018; Doshi-Velez and Kim, 2017; Barredo Arrieta *et al*., 2020).

This paper proposes a different starting point. The opacity of contemporary artificial intelligence should not be understood only as a temporary technical failure or as an engineering obstacle that can in principle be removed without remainder. Some opacity arises from scale, dimensionality, nonlinearity, stochastic training, and combinatorial complexity. In such cases, the demand for complete symbolic reconstruction may misdescribe the epistemic form of the system. The question is not simply how to restore the classical ideal of transparent demonstration, but whether another form of epistemic legitimacy is already operative in machine learning.

The hypothesis developed here is that modern AI systems exemplify a form of knowledge based on bounded convergence. They do not normally establish their outputs through proof. Rather, they construct internal representations by iteratively reducing error relative to data, objectives, and constraints. Their epistemic validity, when it exists, depends on stability, robustness, calibration, generalization, and bounded deviation from target processes. This is why the concept of Truth-ε is needed. It names the epistemic condition in which knowledge is not the elimination of error but its disciplined containment.

To describe AI in this way is not to lower the standards of knowledge. On the contrary, it is to identify the standards appropriate to systems operating under finite data, finite computation, stochastic environments, and irreducible uncertainty. A model that merely produces plausible outputs does not satisfy such standards. A system that overfits, collapses under perturbation, or cannot generalize beyond its training distribution does not produce knowledge in the relevant sense. Truth-ε requires more than performance. It requires bounded and stable convergence.

The epistemic problem of AI can therefore be stated as follows: how can a finite computational system generate representations that count as knowledge when those representations are neither exact copies of reality nor fully reconstructible as explicit demonstrations? The answer proposed in this paper is that such systems may produce knowledge when their representations converge toward stable regions of bounded error under clearly specified informational and computational constraints. This form of knowledge is neither classical demonstrative truth nor mere pragmatic success. It is Truth-ε.

Demonstrative rationality and its limits

The epistemic difficulty posed by artificial intelligence becomes clearer when placed against the background of the classical model of demonstrative rationality. In this model, knowledge is valid when it can be reconstructed as a sequence of explicit steps leading from accepted principles to justified conclusions. To know is not merely to arrive at a correct result, but to be able to show why the result follows. Truth is therefore inseparable from proof, and proof is understood as a transparent structure of reasons.

This model received its canonical mathematical form in Greek geometry. Euclid's Elements did not merely collect geometrical results. It organized mathematical knowledge as a deductive system, beginning from definitions, postulates, and common notions, and proceeding through demonstrations. The authority of the result depended on the possibility of retracing each step of the argument. The conclusion was not accepted because it was useful, probable, or empirically successful, but because it followed necessarily from what had been established before (Euclid, trans. Heath, 1956).

The Euclidean model became one of the most powerful images of rational knowledge in the history of science. It suggested that understanding consists in making relations explicit, that certainty depends on derivation, and that the ideal form of knowledge is one in which nothing essential remains hidden. This ideal extended far beyond geometry. It influenced philosophy, natural science, logic, and later formal systems. From Aristotle's theory of demonstration to early modern mathematical physics and the axiomatic aspirations of modern mathematics, rational knowledge was repeatedly associated with transparency, necessity, and reconstructibility (Aristotle, trans. Barnes, 1984; Hilbert, 1899).

Artificial intelligence challenges this inheritance in a distinctive way. The challenge does not arise because AI systems are irrational or because their operations are mystical. On the contrary, the elementary operations of machine learning systems are mathematically defined. Architectures, loss functions, optimization procedures, and update rules can be described with precision. Yet the global epistemic structure produced by these operations is often not reconstructible as a finite chain of reasons. The model's behavior emerges from distributed transformations across high-dimensional parameter spaces, not from explicit rules that can be inspected one by one as premises in a proof.

This creates a separation between formal specification and epistemic reconstruction. A system may be formally specified at the level of its training algorithm while remaining epistemically opaque at the level of its learned representation. Knowing the rule by which parameters are updated is not the same as understanding the representational structure that results after training. In small systems, the difference may be manageable. In large-scale neural networks, however, the number of interactions among parameters, layers, activations, and training examples makes complete reconstruction infeasible. The system is not outside mathematics, but it is outside the classical ideal of demonstrative transparency.

The limits of demonstrative rationality were not discovered by artificial intelligence. They were already visible within mathematics and logic themselves. The development of infinitesimal

Proceedings of the EUROPEAN ACADEMY of Sciences & Arts PEASA
pagepress

calculus required reasoning with limiting processes that could not initially be grounded in the strict terms of classical demonstration. The later formalization of limits through ε–δ methods restored rigor, but it did so by incorporating controlled approximation into the very structure of proof. Probability theory further displaced certainty by treating uncertainty not as a failure of knowledge but as an object of formal analysis. In the twentieth century, Gödel's incompleteness theorems and Turing's undecidability results showed that formal systems and mechanical procedures have intrinsic limits. Computational complexity then demonstrated that many problems may be formally decidable while remaining infeasible for finite systems (Gödel, 1931; Turing, 1936; Papadimitriou, 1994).

These developments do not refute demonstrative rationality. Proof remains indispensable wherever it is available. Rather, they show that proof is not the only possible structure of epistemic legitimacy. In many domains, especially those involving complexity, uncertainty, high dimensionality, and finite resources, knowledge depends on other criteria: approximation, stability, robustness, convergence, and error control.

Artificial intelligence brings these criteria to the center of epistemology. A trained model does not normally justify its output by presenting a proof. It does not enumerate all possible hypotheses or reconstruct all causal relations in a domain. Instead, it learns a representation that performs reliably under certain constraints. Its epistemic value depends on whether this representation generalizes, remains stable under perturbation, avoids overfitting, and tracks relevant structures in the environment within bounded error.

The classical demonstrative model asks whether a conclusion follows necessarily from explicit premises. The epistemology of AI asks a different question: whether a representation produced under finite informational and computational constraints converges reliably toward the structure of a target domain. This shift does not eliminate rationality. It transforms its operative criteria.

The next section therefore turns to the genealogy of ε. The aim is to show that this transformation is not an abrupt rupture introduced by AI, but the culmination of a long internal development of rational inquiry: the gradual recognition that approximation, when controlled and bounded, can itself become a rigorous form of knowledge.

The genealogy of ε: approximation as rationality

The epistemology of artificial intelligence appears new because contemporary learning systems operate at a scale and level of complexity unknown to earlier scientific practices. Yet the epistemic structure on which they depend is not entirely new. AI brings to the foreground a form of rationality that has long existed within mathematics and science: the production of knowledge through controlled approximation. The history of ε is the history of the gradual transformation of approximation from a practical compromise into a rigorous epistemic principle.

Classical Greek mathematics is usually associated with the demonstrative ideal. This association is justified, but incomplete. Alongside the Euclidean model of proof, Greek mathematics also confronted problems that could not be resolved simply by direct construction or finite demonstration. The discovery of incommensurable magnitudes, the paradoxes of motion and divisibility, and the problem of measuring curved figures forced mathematical thought to confront infinity, continuity, and approximation.

Zeno's paradoxes mark one of the earliest and most powerful formulations of this problem. By dividing motion into an indefinite sequence of intervals, Zeno exposed a tension between finite action and infinite divisibility. The paradoxes did not provide a mathematical theory of limits, but they revealed a conceptual difficulty that would become central to later mathematics: how can a finite process account for a structure that appears to require infinitely many steps? The problem of the infinitesimal thus emerged not as a technical detail, but as a challenge to the very possibility of rationally describing continuity.

The method of exhaustion, associated with Eudoxus and later developed with extraordinary power by Archimedes, provided a decisive response. Instead of grasping a continuous magnitude directly, the method approached it through sequences of increasingly accurate approximations. Areas and volumes could be bounded from above and below, and the difference between the bounding magnitudes could be made smaller than any assigned quantity. This did not abandon rigor. It redefined rigor as the capacity to control error. The result was accepted not because approximation replaced proof, but because approximation was placed under strict bounds (Heath, 1921; Boyer, 1959).

This is the first major epistemic lesson of ε: approximation becomes rational when its deviation is controlled. The object of knowledge is not reached through an uncontrolled guess, but through a procedure that can make the remaining difference arbitrarily small. In this sense, the method of exhaustion already contains the structure later expressed by ε: the idea that truth may be approached through a process whose residual error can be constrained below any relevant threshold.

The development of infinitesimal calculus in the seventeenth century generalized this structure. Newton and Leibniz introduced mathematical techniques capable of describing continuous change, acceleration, curvature, and accumulation. Their methods transformed mathematics and physics, but they also raised foundational difficulties. Infinitesimals were operationally powerful but conceptually unstable. They seemed to function as quantities that were neither zero nor ordinary finite magnitudes. Calculus therefore extended the reach of rational inquiry while simultaneously exposing the need for a deeper account of approximation and limiting processes.

The nineteenth-century formalization of limits by Cauchy and Weierstrass gave this account its rigorous expression. The ε–δ definition of limit eliminated the need to treat infinitesimals as mysterious entities. It replaced them with a precise relation between tolerances. For every ε greater than zero, however small, there exists a corresponding δ such that the relevant deviation remains below ε. This formalism did more than solve a technical problem in analysis. It established a new epistemic form: exactness through controlled approximation. Mathematical rigor no longer required the direct completion of an infinite process. It required the ability to bind the error of finite approximations within arbitrary limits (Boyer, 1959; Cauchy, 1821).

The nineteenth-century formalization of limits was also part of a broader attempt to reconcile number and continuum. The ancient problem had already appeared in Zeno's infinite divisibility and in the discovery of incommensurable magnitudes because geometrical line seemed to exceed the resources of integer ratio. Modern mathematics approached the same problem from the opposite direction, seeking to construct the continuum from

number. Dedekind's cuts defined real numbers through partitions of the rationals, thereby giving rigorous form to the completion of the number line. Cantor's set theory radicalized this project by treating the continuum as an infinite set of points and by distinguishing different orders of infinity. As Jourdain observed in his introduction to Cantor's work, one natural method for locating points of condensation "consists in successively halving the region" containing infinitely many points. What in Zeno appeared as the paradox of endless divisibility thus reappears in modern mathematics as a disciplined procedure for constructing, ordering, and comparing infinite point-sets.

This development is important for the genealogy of ε because it shows that approximation, limit, and infinity are not merely technical devices of analysis. They belong to a deeper foundational problem: whether the infinite should be understood as a completed object, a potential process, a limiting operation, or a rule-governed construction. Dedekind and Cantor transformed the ancient problem of the continuum into a rigorous mathematical inquiry into completion, cardinality, and the structure of infinite sets. The ε of limit analysis therefore stands at the intersection of two traditions: the operational control of approximation and the foundational attempt to determine what kind of object the continuum is.

Probability theory and statistics extended this transformation beyond continuous magnitudes. In probabilistic reasoning, uncertainty is not simply a defect in knowledge. It becomes something that can be measured, modeled, and constrained. Statistical inference does not generally produce certainty in the classical demonstrative sense. It produces warranted conclusions under assumptions about sampling, distribution, error, and confidence. Here again, knowledge depends not on the elimination of uncertainty, but on its disciplined formalization.

Information theory gave this epistemology of bounded error a further and more general formulation. Shannon's theory of communication showed that reliable transmission is possible in noisy channels, but only within limits imposed by entropy and channel capacity. Perfect reconstruction cannot be assumed when noise, finite encoding, and finite bandwidth are present. Reliability becomes a matter of bounding error, optimizing codes, and operating within constraints. This framework is directly relevant to artificial intelligence, because learning systems also operate in environments where information is incomplete, noisy, compressed, and statistically structured (Shannon, 1948; Cover and Thomas, 2006).

Computational complexity adds another limit. Even where a problem is formally defined, its exhaustive solution may be infeasible for finite systems. Combinatorial spaces grow too rapidly to be searched completely. This means that rationality cannot always consist in explicit enumeration, proof, or reconstruction. In large problem spaces, epistemic success often depends on heuristics, approximation algorithms, local search, dimensionality reduction, and the identification of stable patterns. The finite system must converge without surveying the whole space.(Papadimitriou, 1994, 2011).

This genealogy shows that approximation is not external to rationality. It is one of rationality's internal developments. From the method of exhaustion to ε–δ analysis, from probability to information theory and computational complexity, scientific knowledge has repeatedly expanded by learning how to control, measure, and stabilize error. Artificial intelligence inherits this tradition and transforms it into a general mechanism of knowledge production.

Modern learning systems do not treat ε as an exceptional residue left over after proof has failed. They operate through ε. Training stops when improvement falls below a tolerance threshold. Generalization is evaluated through bounded error on unseen data. Robustness is tested through stability under perturbation. Dimensionality reduction methods such as PCA and SVD preserve significant structure while discarding variation judged to be epistemically negligible relative to the task. Regularization intentionally restricts model behavior in order to prevent unstable exactness. In all these cases, knowledge is formed by deciding which deviations matter, which can be bounded, and which can be ignored without destroying reliability.

The history of ε therefore prepares the conceptual transition to Truth-ε. Truth-ε does not arise from a rejection of proof, nor from a postmodern weakening of truth. It arises from a mathematical and scientific tradition in which the control of error becomes a condition of knowledge. Artificial intelligence makes this tradition explicit at scale. It shows that, under conditions of finite data, finite computation, stochastic environments, and high-dimensional representation, the most relevant epistemic question is not whether error can be eliminated, but whether it can be bounded in a stable and reliable way.

Defining truth-ε

The preceding genealogy shows that approximation can become epistemically legitimate when it is controlled, bounded, and stabilized. Truth-ε generalizes this insight for finite cognitive and computational systems. It is introduced here as a concept for describing forms of knowledge that arise when complete correspondence, exhaustive reconstruction, or demonstrative transparency are not attainable, but when error can nevertheless be constrained within stable limits.

Truth-ε may be defined as follows:

> *Truth-ε is bounded epistemic convergence: the strongest form of epistemic reliability attainable by finite cognitive or computational systems operating under irreducible informational, statistical, and computational constraints.*

This definition contains four essential elements: finitude, convergence, bounded error, and stability.

First, Truth-ε presupposes finitude. Knowledge is produced by systems with finite observational access, finite data, finite computational resources, finite time, finite representational capacity, and finite numerical precision. This is true of human cognition, scientific instruments, statistical models, and artificial intelligence systems. A finite system cannot survey all possible states of a complex domain, cannot test all hypotheses, cannot eliminate all noise, and cannot compute all consequences of its assumptions. Its epistemic access to reality is therefore structurally mediated by constraints.

Second, Truth-ε presupposes convergence. Knowledge is not understood as an immediate possession of the real, but as the outcome of procedures that progressively reduce deviation between a representation and a target domain. These procedures may take different forms: geometrical exhaustion, numerical approximation, statistical estimation, Bayesian updating, optimization, or machine learning. What unites them is the movement toward a stable representational region rather than the direct capture of an object in its totality.

Third, Truth-ε presupposes bounded error. The error involved is not accidental ignorance that could always be eliminated by

Proceedings of the EUROPEAN ACADEMY of Sciences & Arts PEASA
pagepress

more effort. In many domains, some non-zero error is irreducible because of finite sampling, measurement noise, stochastic variability, model limitations, computational intractability, or environmental change. Truth-ε therefore does not require ε to vanish. It requires that ε be identified, minimized where possible, and bounded within limits appropriate to the task and domain.

Fourth, Truth-ε presupposes stability. A representation that happens to be close to a target process on one occasion does not yet count as knowledge. It may be lucky, overfitted, fragile, or dependent on accidental features of the data. For a representation to satisfy Truth-ε, it must remain sufficiently stable under relevant perturbations: variations in data, noise, initialization, measurement conditions, or environmental context. Stability distinguishes epistemic convergence from mere approximation.

In simplified mathematical terms, let T denote a target process or structure in the world, and let K denote a representation generated by a finite learning or cognitive procedure L using data D under constraints C. We may write:

$$K = L(D, C)$$

The relation between K and T cannot generally be one of perfect identity. The achievable deviation between representation and target is constrained by the informational and computational conditions under which K is produced. Truth-ε holds when the distance between K and T is bounded by an error tolerance ε that is no smaller than the irreducible lower bound imposed by the domain and the system:

$$\|K - T\| \leq \varepsilon, \text{ where } \varepsilon \geq \varepsilon^*$$

Here ε* denotes the minimal attainable epistemic deviation under the relevant constraints. This lower bound may arise from irreducible noise, finite sampling, model bias, computational complexity, or limits of measurement. The point is not that the value of ε* is always known exactly. Rather, the concept of Truth-ε asserts that epistemic legitimacy depends on recognizing that such bounds exist and on organizing knowledge practices around their control.

A second condition concerns stability. For relevant perturbations D′ of the data or environment, the representation should not vary without bound. A simplified stability condition may be written as:

$$\|K(D') - K(D)\| \leq \delta$$

where δ denotes an acceptable bound of representational variation under perturbation. This condition is crucial. A system that achieves low error but behaves unpredictably under small perturbations does not satisfy Truth-ε in a strong epistemic sense. Robustness is not an optional engineering property. It is part of the epistemic meaning of bounded convergence.

Truth-ε must therefore be distinguished from a merely loose notion of approximation. Approximation by itself may be accidental, unstable, or arbitrary. Truth-ε requires bounded convergence under identifiable constraints and stability under relevant perturbations. It is also an epistemological rather than an ontological claim: it does not assert that reality itself is approximate, but that finite access to reality is mediated by informational, statistical, and computational constraints. Nor does it replace demonstrative truth where proof is available. It applies primarily to empirical, statistical, high-dimensional, and computationally constrained domains in which complete reconstruction is unavailable or infeasible.

The epistemic force of Truth-ε lies in its double requirement. It accepts that error cannot always be eliminated, but refuses to treat this as a license for arbitrariness. Knowledge is possible because error can be bounded, convergence can be tested, and stability can be evaluated. A representation counts as knowledge not because it is error-free, but because its error is constrained in a way that is reproducible, robust, and appropriate to the domain.

This framework is particularly suited to artificial intelligence. Modern learning systems operate with finite datasets, stochastic optimization, high-dimensional parameter spaces, noisy environments, and incomplete access to target processes. They do not normally produce knowledge by deriving conclusions from explicit premises. They produce knowledge, when they do, by converging toward stable representations that minimize error under constraints. AI is therefore not merely an application of Truth-ε. It is its paradigmatic contemporary case.

AI as the paradigmatic case of truth-ε

Artificial intelligence is the paradigmatic contemporary case of Truth-ε because its central mechanisms produce knowledge through bounded convergence rather than demonstrative reconstruction. Modern learning systems do not normally proceed by deriving conclusions from explicit premises. They construct representations by adjusting parameters in response to data, objectives, constraints, and feedback. Their epistemic achievement, when successful, consists in converging toward stable patterns that generalize beyond the particular observations from which they were learned.

Historically, this transition can be traced through the contrast between symbolic and learning-based approaches to artificial intelligence. Symbolic AI extended the demonstrative ideal into computation by treating intelligence as the manipulation of explicit representations according to formal rules. Early systems such as the *Logic Theory Machine* exemplified this ambition, while the *Dartmouth proposal* gave institutional form to the project of artificial intelligence as a computational science. In parallel, cybernetic and neural approaches emphasized adaptation, feedback, and learning under constraints rather than explicit derivation. From McCulloch and Pitts's logical model of neural activity and Rosenblatt's perceptron to the later rediscovery of multilayer learning through backpropagation, the history of AI already contains the tension that this paper interprets epistemologically: the tension between symbolic reconstruction and adaptive convergence (McCulloch and Pitts, 1943; Rosenblueth *et al.,* 1943; McCarthy *et al,.* 1955; Newell and Simon, 1956; Rosenblatt, 1958; Rumelhart *et al.,* 1986).

The basic structure of supervised learning already illustrates this point. A model receives data consisting of examples and target outputs. It defines a parameterized function whose behavior depends on a vector of parameters. A loss function measures the deviation between the model's output and the target. Training consists in adjusting the parameters so as to reduce this loss. The model does not prove that its representation is true. It searches a space of possible representations for regions in which error is minimized (Bishop, 2006; Shalev-Shwartz and Ben-David, 2014; Goodfellow *et al.,* 2016).

Gradient-based optimization makes this structure explicit. In each step, the model updates its parameters in the direction that locally reduces the loss function. The process is iterative,

partial, and finite. It does not survey the entire landscape of possible models, nor does it guarantee access to a unique global optimum. In high-dimensional and non-convex spaces, there may be many local minima, saddle points, flat regions, and alternative configurations with similar performance. Learning therefore consists not in reaching exact identity with the target process, but in converging toward a region of sufficiently low and stable error (Goodfellow *et al.,* 2016).

This is already an ε-structured epistemology. Training normally terminates not when error has disappeared, but when further improvement becomes negligible relative to a threshold, when validation performance stops improving, when computational budgets are exhausted, or when additional optimization risks overfitting. The system accepts a bounded residual deviation. What matters is not the elimination of error, but whether the remaining error is controlled and compatible with generalization.

Generalization is the central epistemic test of machine learning. A model that merely memorizes its training data may achieve low training error but fail to produce knowledge. Its representation is too closely tied to accidental features of the observed sample. A model becomes epistemically valuable only when it captures structure that remains valid beyond the data on which it was trained. Generalization therefore expresses the movement from local approximation to stable representation. It is one of the principal ways in which Truth-ε distinguishes knowledge from overfitting (Vapnik, 1998; Shalev-Shwartz and Ben-David, 2014).

Regularization reinforces this point. Techniques such as weight decay, dropout, early stopping, data augmentation, and capacity control intentionally prevent the model from achieving unstable exactness on the training set. They introduce constraints that may increase training error while improving generalization. From the standpoint of demonstrative rationality, this may seem paradoxical: why should a system be prevented from fitting the data as precisely as possible? From the standpoint of Truth-ε, however, the reason is clear. Knowledge is not maximal conformity to a finite sample. It is stable convergence toward structure that survives variation (Bousquet and Elisseeff, 2002; Hardt *et al.,* 2016).

Dimensionality reduction provides another important example. Methods such as *Principal Component Analysis* and *Singular Value Decomposition* operate by preserving dominant structures while discarding components treated as less significant relative to the representational task. They do not reproduce the full informational content of the original data. They transform it into a lower-dimensional structure in which relevant variance is retained and irrelevant or negligible variation is suppressed. This is not a failure of knowledge but one of its enabling conditions. In complex data environments, knowing often requires compression.

Representation learning generalizes this principle. Deep networks transform raw input into layered internal representations. Early layers may encode relatively simple features; later layers may encode more abstract or task-relevant structures. These representations are not usually designed explicitly by human engineers. They emerge through optimization. Their epistemic status depends on whether they support reliable performance, transfer, robustness, and sensitivity to relevant distinctions. The representation is therefore not a proof, but neither is it a mere black box in the sense of arbitrary opacity. It is a structured product of constrained convergence.

A similar transition can be observed in the history of computer vision. Early vision systems often relied on explicitly constructed structural descriptions: edge detection, contour segmentation, geometric primitive extraction, and syntactic or graph-based object models. These approaches combined numerical preprocessing with higher-level symbolic interpretation and therefore occupied an important transitional position between classical symbolic AI and later learning-based vision. Early work in robotic scene analysis and second-generation image coding illustrates this hybrid regime, in which visual knowledge was still organized through explicit structural representations, while already depending on approximation, feature extraction, and compression. The subsequent rise of deep learning shifted the center of gravity from handcrafted features to adaptive representation learning, thereby reinforcing the broader movement from demonstrative modeling toward bounded computational convergence (Hubel and Wiesel, 1962, 1968; Marr, 1982; Ikonomopoulos, 1980; Kunt *et al.,*1985; LeCun *et al.,* 2015).

The same structure appears in probabilistic and generative systems. Large language models, for example, do not store truth in the form of explicit propositions connected by demonstrative rules. They learn high-dimensional statistical structures from linguistic data and generate outputs by navigating probability distributions conditioned on context. Their successes and failures both reveal the Truth-ε character of machine knowledge. They can capture stable regularities of language and world-description, but they remain vulnerable to hallucination, distributional bias, and failures of grounding. Their epistemic legitimacy therefore cannot be inferred from fluency alone. It depends on calibration, verification, robustness, and error control (Bender *et al.,* 2021).

This also explains why scale alone cannot solve the epistemic problem of AI. Larger models, larger datasets, and greater computational power may reduce some errors and improve performance across many tasks. But they do not remove the structural constraints under which learning operates. Data remain finite and selective. Optimization remains approximate. Environments remain stochastic. Representations remain compressed. Generalization remains uncertain under distributional shift. Increasing scale may change the value of ε, but it does not abolish ε.

Artificial intelligence therefore makes visible an epistemic structure that was already present in approximation, statistics, and information theory: reliable knowledge under finite constraints is achieved through the bounding and stabilization of error. The question is not whether the model has escaped error altogether, but whether its error is measured, constrained, robust, and appropriate to the domain of application.

This framework also provides a way to distinguish legitimate machine knowledge from mere performance. A system that gives correct answers in a narrow benchmark but fails under slight perturbations does not satisfy Truth-ε. A system that performs well while encoding unstable correlations or biased patterns does not satisfy Truth-ε in the relevant epistemic sense. A system that produces plausible but uncalibrated outputs does not satisfy Truth-ε. Machine knowledge requires more than output success. It requires stable convergence under constraints.

AI is thus not an accidental example of Truth-ε. It is the contemporary domain in which Truth-ε becomes unavoidable. Modern learning systems show that knowledge may be formal without being demonstrative, reliable without being transparent in the classical sense, and rational without eliminating error. They operate within ε, and their epistemic validity depends on how that ε is bounded, minimized, stabilized, and made accountable.

Proceedings of the EUROPEAN ACADEMY of Sciences & Arts PEASA
pagepress

Machine knowledge, opacity, and explainability

The concept of Truth-ε has direct consequences for debates concerning machine knowledge, opacity, and explainability. These debates often begin from a legitimate concern: if an artificial intelligence system affects scientific inquiry, medical diagnosis, legal judgment, public administration, or everyday communication, its outputs cannot be accepted merely because they are produced by a powerful model. Epistemic and social legitimacy require reasons, checks, and accountability. The question is what kind of reasons are appropriate for systems whose epistemic function is based on high-dimensional statistical convergence rather than explicit symbolic derivation.

A common response to AI opacity is to demand explainability. This demand is both necessary and ambiguous. It is necessary because opaque systems can conceal error, bias, arbitrariness, manipulation, and unjustified authority. It is ambiguous because explainability can mean several different things. It may mean that the mathematical operations of the system are known. It may mean that a particular output can be locally interpreted. It may mean that the system's global internal representation can be translated into human-understandable rules. Or it may mean that the system's behavior can be evaluated, audited, and constrained even when complete reconstruction is impossible (Burrell, 2016; Lipton, 2018; Rudin, 2019).

Truth-ε helps distinguish these meanings. At least three levels should be separated.

First, there is operational transparency. At this level, one knows the architecture, training procedure, loss function, optimization method, and computational operations of the system. Many AI systems are transparent in this limited sense: their formal mechanisms can be described mathematically. Operational transparency, however, does not by itself provide epistemic understanding of the learned representation. Knowing how parameters are updated does not mean knowing what the final high-dimensional configuration represents in humanly interpretable terms.

Second, there is local explainability. At this level, one seeks an account of why a particular input led to a particular output. Saliency maps, feature attribution methods, counterfactual explanations, example-based explanations, and local surrogate models belong to this family. Such methods can be useful, especially for diagnosis, debugging, and accountability. But local explanations are often partial, approximate, and themselves model-dependent. They do not necessarily reconstruct the full representational logic of the system (Doshi-Velez and Kim, 2017; Barredo Arrieta *et al.,* 2020).

Third, there is global demonstrative reconstruction. At this level, one would be able to translate the whole learned model into an explicit, humanly surveyable structure of rules, reasons, or causal relations. This is the ideal inherited from demonstrative rationality. In small symbolic systems it may be possible. In large-scale machine learning systems it is usually unavailable. The absence of global reconstruction is therefore not merely a gap in current engineering practice. In many cases it follows from scale, dimensionality, nonlinear interaction, and combinatorial complexity.

Truth-ε does not deny the importance of explanation. It denies that epistemic legitimacy must always take the form of complete demonstrative reconstruction. For many AI systems, the relevant epistemic question is not whether the whole model can be translated into explicit reasons, but whether its outputs are produced by a process whose error is bounded, whose behavior is stable, whose limitations are known, and whose performance generalizes under appropriate conditions.

This shift has important consequences for machine knowledge. A machine output should not count as knowledge merely because it is correct, fluent, or useful. Correctness on isolated cases may be accidental. Fluency may conceal hallucination. Usefulness may be domain-specific and fragile. For an AI-generated representation to have epistemic status, it must satisfy conditions of bounded convergence. It must be reliable across relevant cases, robust under perturbation, calibrated with respect to uncertainty, sensitive to domain constraints, and open to external validation.

Explainability, from the standpoint of Truth-ε, should therefore be broadened. It should include not only interpretive access to internal mechanisms, but also the articulation of epistemic bounds. A satisfactory explanation of an AI system may include information about the data on which it was trained, the domain in which it is reliable, the distributional conditions under which its performance has been tested, the types of perturbation it can withstand, the uncertainty associated with its outputs, the known failure modes, and the error tolerances within which it operates.

In this sense, the epistemic demand shifts from "show me the complete proof" to "show me the bounds of convergence." This does not weaken accountability. It makes accountability more appropriate to the actual structure of machine learning. A system can be epistemically unacceptable not because it lacks a full symbolic proof, but because its error bounds are unknown, its robustness is untested, its uncertainty is uncalibrated, or its performance collapses outside narrow benchmark conditions.

Truth-ε also clarifies the difference between opacity and irrationality. A system may be opaque in the sense that its internal representation cannot be fully reconstructed, while still being rationally assessable through tests of robustness, calibration, reproducibility, and domain validity. Conversely, a system may provide plausible explanations while remaining epistemically unreliable. Interpretability is therefore not identical with knowledge. It is one possible route to epistemic assessment, but not the only one.

This is particularly important in the case of large language models. Such systems can produce articulate explanations of their own outputs, but these explanations may be post hoc, incomplete, or inaccurate. The presence of a verbal explanation does not guarantee epistemic validity. From the perspective of Truth-ε, the relevant question is whether the output tracks a stable structure under appropriate constraints, whether uncertainty is properly represented, and whether the answer survives independent verification. The model's fluency is not evidence of truth unless it is embedded within bounded and accountable convergence (Bender *et al.*, 2021).

AI, human thought, and the question of a new intelligence

This distinction helps situate the work of Athanassios S. Fokas on whether artificial intelligence can reach human thought. Fokas argues that the definition of intelligence often invoked by proponents of AI, such as "*the ability to accomplish complex goals*", may be appropriate for machines, but does not capture the essence of human thought. Human thought, in his analysis, cannot be reduced to problem-solving performance or computational success. It involves embodiment, subjective experience, unconscious processes, conscious reflection, and

higher-order forms of comprehension. This line of inquiry is developed in "*Can Artificial Intelligence Reach Human Thought*?" and more broadly in *Ways of Comprehending* and *The Embodied Mind*, where AI is examined in relation to the embodied conditions of human understanding.

The present paper is compatible with Fokas's distinction between artificial intelligence and human thought, but approaches the issue from the standpoint of epistemology. Truth-ε explains why AI may nevertheless be understood as a genuinely new type of intelligence. Its novelty does not lie in reproducing the embodied and conscious structure of human thought. It lies in the emergence of systems capable of producing reliable representations through bounded computational convergence under conditions of finite information, uncertainty, and constraint. In this sense, AI is new not because it abolishes the difference between machine intelligence and human thought, but because it establishes a non-demonstrative yet rational mode of knowledge production. Truth-ε names the epistemological structure of that novelty.

The framework also has implications for scientific uses of AI. In scientific discovery, AI systems may identify patterns, generate hypotheses, approximate complex functions, predict molecular structures, or simulate high-dimensional processes. In such contexts, their epistemic contribution often lies not in replacing explanation but in expanding the space of detectable regularities. A model may reveal a stable relation before scientists possess a full causal theory of that relation. Truth-ε provides a way to understand this status: the system contributes knowledge insofar as its representation is robust, reproducible, and bounded, even if later scientific work is required to integrate that representation into a more explanatory theory.

Thus, machine knowledge should be understood as a layered epistemic phenomenon. At one layer, there are mathematical operations. At another, learned representations. At another, empirical validation. At another, human interpretation and institutional accountability. Truth-ε does not collapse these layers into one another. It provides a framework for coordinating them around the central question of bounded convergence.

The result is a revised conception of explainability, which remains essential, but it should not be identified exclusively with the recovery of demonstrative transparency. For AI systems, explanation must often include the mapping of epistemic limits: where the system converges, where it fails, how large its error may be, how stable its outputs are, and under what conditions its representations can be trusted. The epistemic legitimacy of machine knowledge lies not in the fantasy of errorless transparency, but in the disciplined disclosure and control of ε.

Truth-ε and related philosophical accounts

Truth-ε is not proposed in isolation from existing philosophical accounts of truth, inquiry, and justification. It belongs to a family of views that have challenged the identification of knowledge with final, absolute, or fully transparent certainty. Peirce's conception of truth as the eventual limit of inquiry already treats truth as a regulative convergence of investigation (Peirce, 1878). Popper's account of verisimilitude similarly understands scientific progress as movement toward theories that are closer to truth without being finally verified (Popper, 1963). Fallibilist approaches emphasize that knowledge claims remain revisable, while reliabilist accounts connect epistemic status to the reliability of the processes that produce belief or representation (Goldman, 1979). Contemporary philosophy of information and philosophy of AI have likewise stressed that digital and computational knowledge must be understood in relation to data, abstraction, processing, and informational constraints (Floridi, 2011).

Truth-ε is close to these approaches, but its aim is more specific. It does not merely state that knowledge is fallible, that inquiry converges, or that reliable processes matter. It identifies the mathematical and computational structure through which finite learning systems can produce epistemically legitimate representations under constraints. In this sense, Truth-ε should be understood as a computationally specified account of bounded epistemic convergence. Its central terms are not only approximation and revisability, but finite data, optimization, generalization, robustness, calibration, irreducible error, and stability under perturbation.

The concept is therefore domain-sensitive. It does not replace demonstrative truth in formal systems, nor does it deny the importance of causal explanation where such explanation is available. Rather, it applies primarily to empirical, statistical, high-dimensional, and computationally constrained domains in which complete reconstruction is unavailable or infeasible. In such domains, epistemic legitimacy depends not on the abolition of error but on the disciplined disclosure and control of error bounds. Truth-ε thus avoids both excessive skepticism about AI, which treats opacity as incompatible with knowledge, and excessive instrumentalism, which treats performance as sufficient for knowledge. It proposes instead that machine knowledge requires stable convergence within accountable limits.

Conclusions: AI and the transformation of rationality

Artificial intelligence has often been described as a technological revolution, a new stage in automation, or a possible step toward machine autonomy. These descriptions capture important aspects of the present transformation, but they do not exhaust its philosophical significance. AI also marks a shift in the structure of rationality itself. It forces epistemology to confront systems that generate reliable representations without satisfying the classical ideal of full demonstrative reconstruction.

The argument of this paper has been that this shift should not be understood as a collapse of rationality or as the replacement of truth by mere prediction. Rather, AI makes explicit a form of rationality that has long existed within mathematics and science: the production of knowledge through controlled approximation, convergence, and bounded error. From the method of exhaustion to infinitesimal calculus, from ε–δ analysis to probability theory, information theory, and computational complexity, scientific rationality has repeatedly expanded by learning how to operate under constraints. Artificial intelligence generalizes this trajectory by making bounded convergence the operative structure of knowledge production.

Truth-ε names this epistemic condition. It describes knowledge as the strongest attainable convergence of a finite cognitive or computational system toward a stable representation of a target domain under irreducible informational, statistical, and computational constraints. It does not claim that reality itself is approximate. Nor does it deny the force of proof where proof is available. It claims instead that in empirical, stochastic, high-dimensional, and computationally constrained domains, epistemic validity often depends not on the elimination of error, but on its measurement, minimization, stabilization, and disclosure.

Proceedings of the EUROPEAN ACADEMY of Sciences & Arts PEASA
pagepress

This framework allows us to understand machine knowledge without reducing it either to classical demonstration or to mere instrumental success. AI systems do not normally know by proving. They know, when they know, by converging. Their outputs acquire epistemic status only when the convergence that produces them is robust, calibrated, reproducible, and bounded within the relevant domain. A model that is fluent but ungrounded, accurate but fragile, or successful only under narrow benchmark conditions does not satisfy Truth-ε. Machine knowledge requires stable convergence under constraints.

Truth-ε also reframes the problem of explainability. The demand for explanation remains essential, especially in domains where AI systems affect human lives, scientific decisions, and institutional authority. But explanation should not always be identified with complete symbolic reconstruction. For many learning systems, the more appropriate epistemic demand is the disclosure of bounds: where the system converges, where it fails, how uncertainty is represented, how robust its outputs are, and under what conditions its representations remain valid. The question is no longer only whether the system can provide a proof, but whether its ε is known, controlled, and accountable.

The historical significance of artificial intelligence therefore lies not in its departure from the rational tradition, but in its reconfiguration of that tradition. The demonstrative ideal remains one of the great achievements of human thought. Yet it is no longer sufficient as the sole model of knowledge. Modern AI shows that rationality can also take the form of bounded computational convergence: formal without being fully demonstrative, reliable without being errorless, objective without being absolute, and explainable through limits as much as through derivations.

In this sense, Truth-ε is not a weakening of truth. It is an attempt to name the form truth takes when knowledge is produced by finite systems in complex environments. The epistemology of artificial intelligence begins from this recognition: under the conditions in which contemporary machine learning operates, the decisive question is not whether error can be abolished, but whether it can be disciplined. Truth-ε is the epistemic form of that discipline.

References

Aristotle (1984). Posterior Analytics. In: Barnes J (ed.), The complete works of Aristotle. Princeton: Princeton University Press.

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82-115.

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021). On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM. p. 610-23.

Bishop CM (2006). Pattern recognition and machine learning. New York: Springer.

Bousquet O, Elisseeff A (2002). Stability and generalization. J Mach Learn Res 2:499-526.

Boyer CB (1959). The history of the calculus and its conceptual development. New York: Dover.

Burrell J (2016). How the machine "thinks": understanding opacity in machine learning algorithms. Big Data Soc 3:1-12.

Cantor G (1874). Über eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen. J Reine Angew Math 77: 258-62.

Cantor G (1915). Contributions to the founding of the theory of transfinite numbers. Translated by Jourdain PEB. Chicago: Open Court.

Cauchy AL (1821). Cours d'analyse de l'École Royale Polytechnique. Paris: Debure.

Cover TM, Thomas JA (2006). Elements of information theory. 2nd ed. Hoboken, NJ: Wiley.

Dedekind R (1872). Stetigkeit und irrationale Zahlen. Braunschweig: Vieweg.

Doshi-Velez F, Kim B (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Euclid (1956). The thirteen books of Euclid's Elements. Translated by Heath TL. 2nd ed. New York: Dover.

Floridi L (2011). The philosophy of information. Oxford: Oxford University Press.

Fokas AS (2023). Can artificial intelligence reach human thought? PNAS Nexus 2:pgad409.

Fokas A (2024). Ways of comprehending: the grand illusion and the essence of being human. New Jersey: World Scientific.

Fokas A (2026). The embodied mind: unravelling AI, medicine, and physics. New Jersey: World Scientific.

Goldman AI (1979). What is justified belief? In: Pappas GS (ed.), Justification and knowledge. Dordrecht: Reidel. p. 1-23.

Goodfellow I, Bengio Y, Courville A (2016). Deep learning. Cambridge, MA: MIT Press.

Gödel K (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatsh Math Phys 38:173-98.

Hardt M, Recht B, Singer Y (2016). Train faster, generalize better: stability of stochastic gradient descent. In: Proceedings of the 33rd International Conference on Machine Learning. p. 1225-34.

Heath TL (1921). A history of Greek mathematics. Oxford: Clarendon Press.

Hilbert D (1899). Grundlagen der Geometrie. Leipzig: Teubner.

Hubel DH, Wiesel TN (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106-54.

Hubel DH, Wiesel TN (1968). Receptive fields and functional architecture of monkey striate cortex. J Physiol 195:215-43.

Ikonomopoulos A (1980). Traitement de caractéristiques d'images en vue de l'analyse de scènes en robotique. PhD thesis, Université de Paris.

Kunt M, Ikonomopoulos A, Kocher M (1985). Second-generation image-coding techniques. Proc IEEE 73:549-74.

LeCun Y, Bengio Y, Hinton G (2015). Deep learning. Nature 521:436-44.

Lipton ZC (2018). The mythos of model interpretability. Commun ACM 61:36-43.

Marr D (1982). Vision: a computational investigation into the human representation and processing of visual information. San Francisco: W. H. Freeman.

McCarthy J, Minsky ML, Rochester N, Shannon CE (1955). A proposal for the Dartmouth summer research project on artificial intelligence.

McCulloch WS, Pitts W (1943). A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115-33.

Newell A, Simon HA (1956). The logic theory machine. IRE Trans Inf Theory 2:61-79.
Papadimitriou CH (1994). Computational complexity. Reading, MA: Addison-Wesley.
Papadimitriou CH (2011). Algorithms, complexity, and the sciences. Proc Natl Acad Sci U S A 108(Suppl 3):18238-44.
Peirce CS (1878). How to make our ideas clear. Pop Sci Mon 12:286-302.
Popper KR (1963). Conjectures and refutations: the growth of scientific knowledge. London: Routledge & Kegan Paul.
Rosenblatt F (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65:386-408.
Rosenblueth A, Wiener N, Bigelow J (1943). Behavior, purpose and teleology. Philos Sci 10:18-24.
Rudin C (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206-15.
Rumelhart DE, Hinton GE, Williams RJ (1986). Learning representations by back-propagating errors. Nature 323:533-36.
Shalev-Shwartz S, Ben-David S (2014). Understanding machine learning: from theory to algorithms. Cambridge: Cambridge University Press.
Shannon CE (1948). A mathematical theory of communication. Bell Syst Tech J 27:379-423, 623-56.
Turing AM (1936). On computable numbers, with an application to the Entscheidungsproblem. Proc Lond Math Soc 42:230-65.
Vapnik VN (1998). Statistical learning theory. New York: Wiley.