https://doi.org/10.64142/jeai.1.3.19

Journal of Engineering and Artificial Intelligence

Research Article

JEAI-25-19

Beyond Prediction: Structuring Epistemic Integrity in Artificial Reasoning Systems

Craig S Wright[*]

Department of Computer Science, University of Exeter Ltd, Exeter, UK

[*]**Corresponding author:** Craig S Wright, Department of Computer Science, University of Exeter Ltd, Exeter, UK, E-mail: craig@rcjbr.org

Received date: June 19, 2025; **Accepted date:** June 27, 2025; **Published date:** October 14, 2025

Citation: Wright CS (2025) Beyond Prediction: Structuring Epistemic Integrity in Artificial Reasoning Systems. J Eng Artif Intell Vol.1 No.3: 19.

Abstract

This paper develops the structural framework for epistemic integrity in artificial reasoning systems, preserving the original theoretical foundations while adapting them into a concise, direct form. The model enforces truth-preservation, justification, contradiction management and verifiability across the system's reasoning processes. Epistemic norms are embedded into belief acceptance thresholds, metacognitive supervision mechanisms and hybrid inference architectures, ensuring that all outputs are both semantically grounded and logically coherent. Immutable audit trails secured through blockchain technology maintain a verifiable record of belief formation and revision, enabling sustained accountability over time.

Keywords: Epistemic integrity; Belief revision; Metacognition; Symbolic inference; Blockchain auditability; Artificial reasoning systems

Introduction

This section frames the central problem and the architectural stance adopted to address it: Modern predictive systems generate fluent sequences without the mechanisms that bind assertions to justified belief, exposing a gap between plausible output and knowledge-bearing commitment.

The approach taken here treats belief as a persistent, auditable state governed by invariants consistency, closure and traceability and by policies that elevate graded confidence to categorical acceptance only when stability and evidential thresholds are met. Generation is separated from endorsement: proposals from statistical components are vetted by a symbolic core that enforces semantic well-typedness, satisfiability in intended models and sound consequence; metacognitive supervision monitors confidence, centrality and fragility to trigger diagnosis, defence or revision; contradiction detection and minimal-mutilation change restore coherence with least disturbance; semantic anchoring ties symbols to determinate reference; and provenance with immutable records binds present commitments to their grounds and procedures. The section begins by diagnosing epistemic deficiency in stochastic prediction, then sets out the shift to structural justification and concludes with a high-level overview of the architecture that will be developed in detail.

The problem of epistemic deficiency

Modern large-scale language models optimise for p (tokens | context) and thereby privilege surface plausibility over epistemic warrant. Fluency and coherence are achieved by statistical interpolation rather than by a regime that binds assertions to justified belief. In the absence of explicit propositional commitment, an output does not enter a regulated belief base B_t; it remains an unowned string with no standing in the agent's cognitive economy. This gap between production and commitment explains why such systems can contradict themselves without detection, shift positions without rationale and present statements that cannot be audited for their grounds [15,63].

A truth-governed agent requires internal norms that distinguish chance agreement with fact from knowledge. At minimum, propositions must be admitted under rules that maintain consistency and enable principled update. Belief-revision theory provides the blueprint: contraction and revision operators that preserve rational constraints while accommodating new information [10,53]. Coherence must be more than a global veneer; it must be enforced at the level of acceptance and retention, with stability conditions that prevent oscillation under minor evidential perturbations [23]. Where probabilistic assessments are involved, justification cannot collapse into mere scoring; probability must be disciplined by normative principles that connect degrees of belief with rules of acceptance and expected accuracy [56,13,68].

Epistemic integrity further demands representational structures that support knowledge ascriptions and error correction. Epistemic logics and related formalisms show how agents can model knowledge, belief and their dynamics, providing operators and frames in which commitment, possibility and update are tractable [28,59,38]. Operationally, the system needs mechanisms for maintaining and repairing its commitments; truth-maintenance architectures exemplify how justificatory links and dependency records support retractable, accountable inference [35]. Finally, claims must be anchored semantically rather than floated on symbol manipulation alone; the symbol grounding problem marks the necessity of tying representations to determinate content so that "agreement" is about the world, not merely about strings [64].

Without these structural commitments admission rules, revision operators, justificatory records and semantic anchoring an artificial system remains a generator of well-formed sentences rather than a bearer of knowledge. The deficiency is architectural, not cosmetic: It is the absence of a machinery that makes truth preservation and justification constitutive of what the system is and what it does.

From stochastic prediction to structural justification

Transitioning from stochastic token-sequence generation to genuine epistemic agency requires a fundamental redesign of the reasoning substrate. In a predictive architecture, outputs are optimised for likelihood given preceding context; their correctness is incidental, an emergent by-product of statistical correlation rather than a targeted outcome [15]. Such systems lack a commitment layer that would transform an utterance into an endorsed proposition within a regulated belief base B_t. Without that, there is no standing obligation to defend, revise or retract on epistemic grounds.

Structural justification inserts this missing layer. Each asserted proposition p is admitted only through an acceptance function α (p, t), defined over both evidential support and internal consistency at time t. Admission is conditional: p must cohere with existing commitments, survive contradiction checks and be grounded in a justificatory chain that can be inspected and audited [35,23]. This transforms the act of output generation from a statistical choice into the execution of an epistemic rule, binding the system to the content it asserts.

To achieve this, the architecture separates generation from endorsement. A generative module may propose candidate continuations or answers, but an epistemic filter evaluates each candidate for acceptance into B_t. The filter operates over structured representations logical forms, semantic graphs or proof trees that explicitly encode inferential dependencies [59,38]. Accepted propositions are thus embedded in a network of justifications, each linked to its evidential and inferential grounds, enabling backtracking when upstream premises are revised.

This shift demands formal integration of deductive, inductive and abductive inference with metacognitive monitoring. Deductive validity ensures truth-preservation under accepted premises; inductive inference manages probabilistic updates consistent with Bayesian conditioning; abductive inference proposes explanatory hypotheses constrained by back- ground commitments [13,68]. All are subject to revision protocols that enforce global consistency, preventing the unchecked accumulation of incompatible commitments.

In effect, the move to structural justification redefines output generation as a normative act. The system does not merely predict what comes next; it asserts only what it can defend within its epistemic economy. This reframing transforms a language model from an oracle of plausible strings into an accountable epistemic agent one whose outputs are not probabilistic guesses, but justified claims tethered to evidence, logic and an explicit history of reasoning.

Overview of the condensed architecture

The architecture retains the complete logical and normative structure of the original full-scale framework while compressing its presentation into a form suitable for direct implementation and rigorous audit. No alterations are made to the epistemic principles, representational formalisms or the original bibliographic and citation infrastructure; all citations correspond exactly to those defined in references.bib. The resulting structure is not a conceptual abridgement, but a high-density reformulation of the same system, preserving the full scope of operational detail.

The model is organised into six mutually dependent domains, each reflecting a stage in the epistemic lifecycle of an artificial reasoning agent. The first domain establishes the foundational problem: Stochastic generators produce linguistically coherent sequences without internalised justification or epistemic accountability. This deficiency motivates the intro- duction of a belief-commitment layer and explicit justificatory protocols, ensuring that any proposition admitted into the belief base Bt is both evidentially grounded and logically coherent.

The second domain encodes the system's normative core. Propositional commitments are governed by explicit acceptance thresholds, formal contradiction handling and preservation of inferential closure within B_t. Here, contradictions are treated as violations of design constraints, triggering structured belief revision according to formal models of epistemic change. These protocols ensure that the belief base remains a closed, coherent set of endorsed propositions over time.

The third domain models belief as a computationally persistent, inferentially active entity. Each belief is represented in a structured, queryable format, with metadata for provenance, supporting evidence and inferential dependencies. A metacognitive supervisory layer monitors the evolution of B_t, detecting contradictions, initiating revisions and enforcing the integrity of the entire belief network.

The fourth domain implements a hybrid inferential substrate. Symbolic reasoning mod- ules, semantic representation layers and graph-structured justification networks integrate with statistical components to produce conclusions that are both probabilistically supported and logically defensible. Inference pathways are explicitly recorded, enabling transparent inspection of how each accepted belief was derived.

The fifth domain provides immutable epistemic provenance. Every belief update, justification and contradiction resolution is recorded in a cryptographically verifiable ledger, ensuring that the system's epistemic history is non-repudiable and open to external audit. This record functions as an enforceable memory of epistemic acts, binding the agent to its prior assertions and revisions.

The final domain synthesises the components into a coherent operational blueprint. The interrelations between normative constraints, metacognitive supervision, inferential subsystems and immutable record-keeping are specified to enable reproducible deployment. The architecture enforces a transition from unconstrained generative behaviour to epistemically governed reasoning, where outputs are not merely plausible, but justified, accountable and traceable to their evidential and inferential origins.

Epistemic Norms and Foundations

This section fixes the normative ground rules that convert outputs into justified beliefs: probabilistic coherence regulates graded credences, categorical acceptance obeys stability-aware thresholds, logical consequence preserves truth and principled change restores coherence when commitments clash. Degrees of belief are constrained by the probability calculus and representation theorems [5,56,13], while categorical acceptance is linked to accuracy and coherence considerations [29,51] and, where appropriate, to imprecise-probability safeguards under partial information [54]. Truth-conditional semantics and sound proof ensure that ad- mitted claims have determinate meaning and are closed under valid inference [2,19] and stability constraints tie acceptance to robust regions rather than knife-edge thresholds [23]. When new evidence threatens consistency, belief change follows minimal-mutilation principles from revision theory so that the belief base remains consistent, closed and traceable with least disturbance to warranted content [10,53].

The problem of epistemic deficiency

Epistemic deficiency refers to the structural and operational limitations that prevent an epistemic agent from sustaining justified, coherent and dynamically adaptable belief states.

In computational systems, this deficiency manifests when the mechanisms for evidence acquisition, belief updating and inferential closure are incomplete, inconsistent or misaligned with the norms of rational inquiry [3,52]. The consequence is a belief base that is either underdetermined by evidence, incoherent in logical structure or incapable of adapting to new information without introducing contradictions.

The architecture under consideration addresses these deficiencies by integrating formal theories of justification with probabilistic reasoning frameworks. Following the evidentialist insight that belief is epistemically permissible only when proportioned to the available evidence [57,14], the system implements a mapping from evidence sets *E_t* to graded credences constrained by the axioms of probability [5,56]. This probabilistic foundation mitigates the risks of arbitrary or heuristic-based belief fixation, ensuring that commitment to a proposition is the outcome of norm-governed computation rather than stochastic preference.

From a Bayesian perspective, deficiency often arises when agents fail to update credences in a manner consistent with Bayes' theorem, thereby violating coherence constraints over time [30,55]. The present design employs a continuous update regime, where posterior probabilities are recalculated for all active propositions whenever new data is ingested. This ensures diachronic consistency and guards against path-dependent distortions in the belief network.

However, epistemic adequacy also requires inferential completeness: If an agent justifiably believes p and $p \rightarrow q$, it must also believe q, barring defeaters [17,37]. Systems lacking this property suffer from fragmentation, where justified commitments fail to propagate through the network. The proposed architecture incorporates a closure-preserving inference engine that systematically derives such consequences, thereby reducing gaps in the belief structure. By embedding evidential proportionality, Bayesian updating and inferential closure as non-optional structural features, the architecture directly counteracts the principal sources of epistemic deficiency. The result is a belief base that is dynamically stable, logically coherent and resistant to the degradation of justification over time qualities indispensable for any system claiming genuine epistemic competence.

Evidentialism, bayesianism and inferential closure

Evidentialism imposes the requirement that acceptance of a proposition p be proportionate to the agent's available evidence E_t, so that endorsement is never detached from demonstrable warrant [14,57]. In the present framework, this requirement is operationalised by assigning to each p a graded justificatory status derived from explicit evidence–proposition map. The evidential mapping is quantitative rather than rhetorical: Support is computed from the system's evidence structures and recorded as a confidence value that must meet a pre specified threshold before p may enter the active belief base B_t.

Probabilistic semantics provides the discipline for this quantitative evidentialism. Degrees of belief are constrained by the axioms of probability [5] and by representation theorems that connect qualitative plausibility constraints with real-valued credences [56]. On this foundation, the evolution of belief is governed by the Bayesian norm: When new evidence e is acquired, the posterior on p is obtained by conditionalisation one, ensuring diachronic coherence rather than ad hoc adjustment [13,55,30].

Where evidence arrives in graded form, Jeffrey conditionalisation supplies the appropriate generalisation, maintaining global coherence while proportionately redistributing credence. For domains with structured uncertainty or interval-valued information, the architecture admits imprecise-probability updates while preserving the same normative constraints on coherence [54].

Evidentially grounded credences alone do not secure rational acceptance unless their logical consequences are systematically realised. Inferential closure addresses this by requiring that, ceteris paribus, if p is justified and $p \rightarrow q$ is justified, then q must be justified as well. The system enforces closure by coupling a proof-theoretic engine with semantic checks: Derivability is validated in a Gentzen-style calculus to ensure correctness of local inference [19], while modal–epistemic formalisms provide operators for knowledge and belief that respect the intended accessibility relations among epistemic states [28,59,38]. Closure is not an optional post-processing step; it is a standing invariant maintained as beliefs are added, retracted and re-derivated.

Because new information can introduce tension among commitments, closure operates alongside principled change. When propagating consequences would yield inconsistency, the architecture initiates a minimal-mutilation revision that restores coherence while preserving as much justified content as possible. This is guided by entrenchedness and reliability measures that rank sentences and support sets, ensuring rational contraction and revision of B_t rather than arbitrary pruning [10,53]. In parallel, stability constraints on acceptance prevent oscillation by linking categorical acceptance to robust credence regions rather than to knife-edge thresholds, thereby aligning long-run coherence with epistemic resilience [23].

The result is a unified regime in which (i) admission of beliefs is evidentially constrained; (ii) updates are governed by probabilistic norms and (iii) logical consequences are realised and repaired under explicit rules of change. Every accepted proposition is thus simultaneously warranted by evidence, regulated by Bayesian dynamics and integrated into a closed network of consequences an arrangement that eliminates unwarranted assertion, guards against incoherent drift and sustains a transparent record of why the system believes what it does.

Internal truth constraints and logical coherence

Internal truth constraints are system invariants that prohibit the admission or retention of propositions that, taken together, yield a contradiction. They are stated semantically and proof-theoretically. Semantically, truth is fixed by model-theoretic satisfaction so that accepted sentences must admit a model under the intended interpretation [1,2]. Proof-theoretically, the reasoning kernel is required to preserve validity; derivations proceed in a cut-admissible calculus so that only consequences licensed by the rules may enter the belief base [19]. Formally, for the active belief set Bt the following invariants are enforced:

$$\textbf{(Consistency)}\quad \mathcal{B}_t \nvdash \bot \qquad\qquad \textbf{(Closure)}\quad \text{if } \mathcal{B}_t \vdash \varphi \text{ then } \varphi \in \mathcal{B}_t$$

Operationally, closure is maintained in a constructive form: if $\varphi \in B_t$ and $(\varphi \rightarrow \psi) \in B_t$, then ψ is admitted, barring a registered defeater.

When $\bot$ is derivable from $B_t \cup \{\chi\}$, the system triggers belief change governed by minimal mutilation principles. Revision and contraction follow the partial-meet/entrenchment schemes so that conflicting items are removed with least disturbance to the remaining commitments [10,53]. Dependency-directed justification is maintained throughout: Each sentence in B_t carries a support set (premises, rules and provenance), enabling targeted retraction rather than global rollback, in the spirit of truth-maintenance architectures [35]. This yields a cycle of detect (derive $\bot$ or an explicit clash), diagnose (identify minimal inconsistent sets) and repair (apply contraction/revision according to entrenchment rankings) until B_{t+1} satisfies the invariants again.

While paraconsistent logics strategically tolerate contradictions, the present architecture treats contradiction as a fault signal, not an acceptable steady state. Paraconsistent systems (and their analyses) clarify the landscape of options [20,49,31], but the design choice here prioritises classical coherence with explicit repair over concurrency of incompatible commitments. This preserves classical consequence relations for downstream modules that rely on monotonic, truth-preserving inference.

Logical coherence is strengthened by ranking and stability constraints. Ordinal Conditional Functions (OCFs) rank worlds or sentences to guide defeat and re-acceptance, ensuring that revision follows principled priority relations rather than ad hoc heuristics [72]. Stability style constraints tie categorical acceptance to robust regions of credence or entrenchment, preventing oscillation under minor evidential perturbations [23]. For agents that reason about knowledge and belief explicitly, the invariants extend to modal-epistemic structure: Closure and consistency are enforced in the object language and at the level of epistemic operators, in line with foundational accounts of knowledge/belief and their dynamics [28,59,37].

These mechanisms jointly ensure that (i) no set of accepted commitments entails triviality; (ii) logical consequences are realised and tracked with justifications and (iii) any emerging conflict is resolved by principled change rather than suppression. Internal truth constraints thus make coherence a standing architectural property, not a post hoc aspiration and they provide the conditions under which subsequent probabilistic and semantic modules can operate reliably within a single, unified notion of correctness.

Belief Architecture and Metacognitive Supervision

This section specifies how the system treats beliefs as persistent, inferentially active objects and how a supervisory layer governs their lifecycle. We introduce a structured belief base B_t with an accompanying justification graph that records dependencies and provenance; stability-aware thresholds that elevate graded confidence to categorical acceptance; and ranked revision operators that retract the least entrenched items to restore coherence when conflicts arise. We show how closure under valid consequence is maintained alongside consistency and traceability, how belief identities persist across updates for targeted repair and explanation and how the metacognitive controller monitors second-order properties (confidence, centrality, fragility) to trigger diagnosis, defence, contraction or revision. The supervisor also allocates computation under resource limits, prevents oscillation at acceptance boundaries and coordinates with inference, grounding and audit layers so that every addition, retraction and re-derivation remains coherent, explainable and operationally accountable.

Computational belief and structural persistence

Computational belief is not the mere occurrence of a token but a persistent, inferentially active state within the agent's architecture. Formally, the epistemic state at time t is represented by a tuple $\Sigma_t = \langle B_t, J_t, \kappa_t, E_t \rangle$, where B_t is the set of accepted propositions (the belief base), J_t is a justification/provenance graph that records dependency structure among beliefs and evidence, κt is an ordinal ranking or entrenchment assignment governing defeat and revision and E_t is the currently accessible evidence pool. A proposition counts as held only if it is an element of B_t and is supported by links in J_t that satisfy the architecture's acceptance and coherence constraints. This treatment follows classic representational accounts of belief and knowledge in knowledge representation and epistemic logic, where beliefs are encoded as stable, queryable objects rather than transient outputs [41,25,28,59,38].

Persistence and identity over time: Structural persistence requires that beliefs, their supports and their roles in inference retain identity across updates. The update operator, Update: $\Sigma_t \times Obs_t \longrightarrow \Sigma_{t+1}$, maps the current state and new observations to a successor state while preserving designated invariants (consistency, closure and traceability). Belief identity is maintained by stable identifiers in J_t so that retractions and re-derivations target the same propositional objects rather than generating aliases. This commitment to diachronic identity accords with practice in cognitive architectures in which symbolic structures persist and are transformed under task demands, as in SOAR and ACT-R [33,34].

Justification and provenance: Every $\varphi \in B_t$ is associated with a justification set $J_t(\varphi)$ that cites premises, inference rules and evidence sources. The provenance layer enables targeted repair when upstream items change and provides external auditability. In database and workflow settings, analogous structures are captured by provenance semirings and the Open Provenance Model, which inform our representation of support propagation and replay [69,44]. Within the agent, provenance supports explanation ("why φ?"), responsibility (who/what introduced φ?) and rollback (which minimal changes discharge a conflict involving φ?).

Ranking, entrenchment and stability: The ranking component κ_t provides a priority ordering over sentences (or possible worlds), guiding defeat, contraction and re-acceptance. Ordinal conditional functions offer a principled basis for such rankings, ensuring that revision is sensitive to degrees of commitment rather than being purely set-theoretic [72]. To avoid oscillation at categorical acceptance thresholds, acceptance is tied to stability regions in credence or rank: Sentences are treated as accepted only when they reside in robust basins that withstand small evidential perturbations, aligning with formal stability criteria for coherent belief [23].

Belief changes under minimal mutilation: When new information threatens coherence, the architecture applies a minimal mutilation policy: Change as little as necessary to restore the invariants. Concretely, contraction and revision operators are implemented in the spirit of partial-meet/entrenchment frameworks so that B_{t+1} preserves maximal justified content while eliminating the source of inconsistency [10,53]. Because J_t records support links, the system can construct minimal inconsistent sets and prefer local repairs over global rollback, a strategy anticipated in truth-maintenance systems [35].

Inferential activity and semantic anchoring: Beliefs are inferentially active: if $\varphi \in B_t$ and $(\varphi \rightarrow \psi) \in B_t$, then ψ is derived (barring a registered defeater) and its addition is recorded in J_t with links to its premises. Proof-theoretic discipline and semantic anchoring ensure that derivations both preserve validity and connect symbols to determinate content, avoiding purely syntactic drift [19,40,64]. This division of labour proof rules to guarantee correctness, anchoring to guarantee meaning secures the cognitive role of B_t as more than a cache of strings.

Outcome: Computational belief, so construed, provides the substrate on which rational update, explanation and decision rest. Persistence, provenance, ranking and principled change turn acceptance into a governed state rather than an ephemeral event, aligning the agent's internal economy of reasons with established theories of knowledge, belief and belief dynamics [59,38,41,53].

Metacognition and second-order epistemics

Metacognition is the capacity of an agent to represent, monitor and regulate its own cognitive states. In the present architecture, metacognition is implemented as a supervisory process that maintains second-order beliefs about the status, strength and provenance of first-order beliefs and intervenes to diagnose, revise or defend those beliefs as conditions evolve [48]. Concretely, the agent's epistemic state at time t includes both a belief base B_t and a meta-level registry of assessments over B_t, enabling the system to reason not only with propositions, but also about their justification status, stability and role in inference.

Formal role of second-order attitudes: Let $\mathrm{Acc}_t(\varphi)$ denote the categorical acceptance of φ at time t and let $\mathrm{Conf}_t(\varphi) \in [0, 1]$ denote its graded confidence. Second-order attitudes include meta-predicates such as $\mathrm{Stable}_t(\varphi)$ (robust under small evidential perturbations), $\mathrm{Central}_t(\varphi)$ (high inferential centrality) and $\mathrm{Fragile}_t(\varphi)$ (near a revision boundary). The metacognitive controller M_t maps the total state Σ_t to diagnostics and actions, $M_t : \Sigma_t \rightarrow$ {diagnose, defend, contract, revise, explain}, selecting interventions that preserve global invariants (consistency, closure, traceability) while minimising unnecessary change. This division of labour follows long-standing practice in cognitive architectures where symbolic state, control knowledge and meta-level policies interact to sustain performance over time [34,33].

Second-order logic of knowledge and belief: Second-order epistemics is grounded in formalisms that explicitly represent operators for knowledge and belief together with their introspective properties. Epistemic logic supplies a well-studied semantics for *K*- and *B*- operators, permitting the agent to reason about what it knows, what it merely believes and how these attitudes propagate under inference [28,59,38]. To capture information change announcements, observations, public updates the architecture employs dynamic epistemic logic so that meta-level expectations about others' (or the system's own) beliefs can be revised by explicit update actions without sacrificing coherence [24]. These tools make second-

https://doi.org/10.64142/jeai.1.3.19

order constraints first-class citizens of the reasoning substrate rather than external bookkeeping.

Supervisory control, goals and practical reasoning: Metacognitive oversight is purposive: It aligns epistemic operations with task and normative goals. The architecture adopts a practical stance in which intentions, plans and commitments regulate when to inquire, when to suspend judgement and when to act on current beliefs [47,4].

Because computational resources are limited, the supervisor allocates attention and computation in a resource-rational manner trading off accuracy, time and utility in a way that preserves core epistemic guarantees while enabling timely decisions [26,27,67]. Thus, metacognition integrates normative constraints with bounded rational control rather than assuming unlimited ideal reasoning.

Monitoring, explanation and provenance: Second-order reasoning is tightly coupled to justification management. Every accepted proposition $\varphi \in B_t$ is associated with a justification subgraph in J_t recording its evidential sources and inferential parents. The supervisor can query J_t to generate explanations (why φ?), to assess vulnerability (what would defeat φ?) and to plan repairs if upstream premises change.

These capabilities draw on provenance formalisms that support principled tracking and replay of data derivations [69,44] and on dependency-directed techniques from truth-maintenance that enable localised retraction rather than global rollback [35].

Stability, ranking and principled change: To avoid oscillation under minor evidence shifts, meta-level acceptance is tied to stability conditions: A sentence is categorically accepted only when its credence or rank lies in a robust region rather than at a knife-edge threshold [23].

When conflicts arise, the supervisor invokes belief-change operations that implement minimal mutilation, guided by entrenchment and ranking principles so that highly central or well-supported commitments are retained while less entrenched ones give way [10,53,72]. In this way, second-order control ensures that revision policies are not ad hoc but normatively governed and globally coherent.

Outcome: Metacognition and second-order epistemics convert a passive store of sentences into an actively managed epistemic economy. By representing and reasoning about its own commitments, justifications and vulnerabilities within established logical and computational frameworks the agent sustains coherence, explains itself and adapts under pressure without forfeiting its truth-preserving guarantees [59,38,33].

Contradiction detection and belief revision mechanisms

Contradiction detection is enforced as a standing invariant on the active belief set B_t. The reasoning kernel combines a proof-theoretic monitor and a semantic admissibility check: Derivations proceed in a Gentzen-style calculus so that only rule-licensed consequences are introduced [19], while acceptance requires model-theoretic satisfiability under the intended interpretation [2]. Formally, the system maintains (Consistency) $B_t \nvdash \bot$, (Closure) if $B_t \vdash \varphi$ then $\varphi \in B_t$ and continuously searches for Minimally Inconsistent Subsets (MIS) whenever a new commitment threatens these invariants. Each sentence is stored with dependency and provenance pointers in a justification graph J_t so that a discovered contradiction yields a precise set of culpable premises rather than an undifferentiated failure [35,69,44].

Repair is governed by belief-change principles that minimise disturbance while restoring coherence. Contraction and revision are implemented in the spirit of the AGM framework, using partial-meet selection and entrenchment orderings to determine which elements of an MIS to retract [10,53]. Let $\mathrm{Rev}_t(\cdot)$ denote the revision operator at time t; upon detecting that $B_t \cup \{\chi\} \vdash \bot$, the system computes a family of remainder sets and selects a maximal consistent subset according to entrenchment, yielding Bt+1 = $\mathrm{Rev}_t(B_t, \chi)$, with B_{t+1} satisfying consistency and preserving as much of B_t as possible. Ordinal Conditional Functions (OCFs) provide a ranking semantics for entrenchment, ensuring that retraction targets the least committed items and that re-acceptance policies are priority- sensitive [72]. To prevent oscillation at categorical thresholds, acceptance is tied to stability constraints so that minor evidential perturbations do not flip commitments back and forth [23].

The architecture treats contradiction as a fault signal rather than a permissible steady state. While paraconsistent logics and analyses illuminate alternatives to explosive consequence, the present design retains classical consequence for downstream modules and opts for explicit diagnosis-and-repair over tolerance of incompatible commitments [20,49,31]. This preserves monotonic interfaces where required (*e.g.,* for verification and planning) while localising nonmonotonicity to the controlled revision phase.

Because the agent reasons about knowledge and belief explicitly, detection and repair extend to the epistemic level. Epistemic and doxastic operators are represented in a modal framework with well-defined accessibility relations, supporting meta-level checks such as "knowing that one believes p" and public or private updates to belief states [28,59,38,24]. When updates (announcements, observations) generate clashes among higher-order commitments, the same MIS/AGM pipeline is applied with rankings that respect the meta- level roles of sentences.

Operationally, the full cycle is: Detect (derive ⊥ or an explicit clash in B_t), diagnose (compute MIS *via* J_t), select (apply entrenchment/OCF priorities), repair (contract/revise to obtain B_{t+1}) and re-derive (close B_{t+1} under consequence). Throughout, provenance records are updated so that every contraction and re-derivation is auditable and replayable [69,44]. The result is a mechanism that guarantees restoration of internal truth constraints with minimal loss of justified content, maintaining a belief base that is both dynamically robust and logically coherent.

Inference, Semantic Grounding and Justification

This section defines how candidate statements become meaning-bearing, defensible claims by passing through a pipeline of formal inference, semantic anchoring and explicit justification. We map well-typed expressions to truth conditions in intended models and constrain derivations with sound proof rules so that consequence tracks meaning rather than mere symbol flow. A hybrid substrate lets statistical components propose structures and scores while a symbolic core verifies satisfiability, compatibility with standing commitments and closure under valid consequence before endorsement. Every accepted proposition is attached to a justification triplet claim, grounds, procedure recorded in a provenance graph that supports replay, targeted repair and explanation, with causal structure available where counterfactual and intervention semantics are required. The result is that endorsed outputs are not only probable but semantically fixed, logically warranted and accompanied by audit-ready traces that connect them to evidence, inference steps and prior commitments.

From syntactic tokens to semantic representations

A reasoning agent that operates only on token sequences lacks the structures needed to assign determinate content to its assertions. The architecture therefore separates syntax from semantics and links them by an explicit interpretation function. Let L be a formal language (signature, terms, formulas) and let M be a class of structures with domain $|M|$ and interpretations for non-logical symbols. The semantic map $J\cdot)$: Form(L) → {0, 1} is defined relative to a model $M \in M$ and assignment g, with satisfaction $(M, g) \vDash \varphi$ following standard clauses [1,2]. Proof-theoretic discipline is enforced with a Gentzen-style calculus to ensure that only rule-licensed consequences are introduced; soundness links derivability to truth in all intended models [19]. This division of labour guarantees that acceptance is governed by meaning, not only by symbol manipulation.

To connect linguistic inputs with formal content, the system employs typed representations that factor lexical items into predicates, relations and terms whose denotations are fixed by M. Distributional features from language processing are retained as evidence but do not themselves constitute meaning; they inform selection among candidate parses and logical forms while the model-theoretic layer determines truth conditions [11,40]. Compositionality is captured by a homomorphic mapping from the algebra of syntactic constructions to the algebra of semantic operations, ensuring that the denotation of a complex expression is determined by the denotations of its parts and their mode of combination. Category-theoretic semantics provides a principled account of such structure-preserving mappings [16].

Semantic representation is anchored beyond symbols. The symbol grounding problem motivates interfaces that bind terms to sensorimotor or observational capacities so that reference is not merely stipulated but connected to the world [64]. In agents that learn and act, grounding is supported by developmental and embodied constraints, which stabilise the link between percepts, concepts and action schemas over time [6]. These anchors constrain admissible interpretations in M, narrowing the model class to those consistent with the agent's abilities to discriminate, predict and intervene.

Beyond extensional truth conditions, the agent represents informational attitudes and dynamics. Modal–epistemic operators (K, B) and their accessibility relations allow the system to state and reason about what is known and believed and how such states interact with inference [28,59,38,50]. Dynamic epistemic logic provides update actions (announcements, observations) that transform epistemic states while preserving global coherence, permitting precise accounts of information flow and revision at the semantic level [24].

For explanatory and decision-theoretic tasks, the semantic layer incorporates structured causal models. By distinguishing associations from interventions, the agent can represent counterfactuals and compute the effects of actions, linking assertions to testable implications rather than correlations alone [39]. This causal stratum integrates with the logical core: Accepted premises license do-calculus derivations whose conclusions inherit the same justification and auditability constraints as ordinary logical consequences.

Finally, tokens produced by generative components are vetted by the semantic layer before any commitment is made. Neural encoders contribute proposals (candidates with scores), but endorsement requires that a candidate be parsable into a well-typed logical form, satisfiable in the intended model class and compatible with the agent's grounded ontology. Neural–symbolic work motivates this division of roles statistical modules for proposal and ranking, symbolic semantics for validity and commitment so that learned regularities are harnessed without surrendering epistemic control [66]. In aggregate, the pipeline from tokens to meanings ensures that every accepted proposition has (i) a determinate denotation; (ii) a proof-theoretic pedigree and (iii) a world-anchored interpretation, thereby turning surface strings into claims with truth conditions, explanatory reach and audit-ready provenance.

Triadic justification and provenance tracing

Justification is organised around a triadic schema: Claim, grounds and procedure. The claim specifies the content to be endorsed; the grounds collect the evidential resources that warrant endorsement; the procedure records the sequence of admissible inference steps that carry grounds to claim. This schema is encoded in a justification graph J_t whose nodes represent claims, evidential items and inference rules and whose labelled edges track relations such as SUPPORTS, DERIVED-FROM and ENTAILS. For every accepted proposition $\varphi \in B_t$, the system stores a triple ⟨claim = φ, grounds = E_φ, procedure = Π_φ⟩, with E_φ drawn from the current evidence base and Π_φ a proof (or derivation plan) certified by the reasoning kernel. Soundness and satisfaction link proof to truth in the intended models, ensuring that recorded procedures are not merely symbolic manipulations but truth-preserving transformations [19,2].

Quantifying justificatory weight: Justification strength is computed by combining (i) probabilistic support contributed by the grounds; (ii) rule reliability associated with the procedure and (iii) penalties for complexity and brittleness. Degrees of belief respect the probability axioms [5] and are updated under new evidence by conditionalisation or its generalisations when inputs are graded [13, 55]. To align numeric confidence with epistemic accuracy, the system evaluates forecasts and categorical acceptances using strictly proper scoring rules, discouraging overconfident endorsements and rewarding well-calibrated justifications [68]. The resulting epistemic weight of a claim is a function of evidence quality and inferential integrity rather than mere statistical correlation.

Provenance as a first-class constraint: Every edge in J_t carries provenance metadata: Source identifiers, timestamps, transformation operators and validation artefacts. Two complementary formalisms underpin this layer. Provenance semirings provide an algebra for propagating annotations through derivations, enabling the system to compute how evidence contributes to outcomes and to replay the formation of results [69]. The open provenance model specifies interoperable records for workflows and data products, ensuring that justificatory traces are portable, queryable and auditable across components [44]. Together they turn explanation (why φ?) and accountability (who or what introduced the support for φ?) into standard queries over Jt.

Dependency, retraction and minimal repair: Because justificatory links expose dependencies, contradictions can be localised to minimal inconsistent sets rather than triggering global rollback. Dependency-directed techniques identify culpable premises or rules, after which belief change is executed under minimal mutilation: contraction and revision remove the least entrenched elements necessary to restore coherence [35,10,53]. Rankings over sentences or worlds (*e.g.,* ordinal conditional functions) provide priority guidance for retraction and re-acceptance, stabilising revisions and preventing oscillation under small evidential shifts [72,23].

Immutable audit of justificatory acts: To make the history of justification tamper evident, the system appends commits of state changes accepted claims, new grounds, derivation certificates and repairs to an append-only, cryptographically linked log. This ledger can be implemented with blockchain techniques that provide content-addressable integrity, ordering and non-repudiation [60,7,36]. Each ledger entry references nodes and edges in J_t, yielding an immutable correspondence between present belief and past justificatory acts. The effect is to extend provenance from "how this result was computed" to "how this belief was justified and maintained," enabling external verification without granting write access to the internal reasoning state.

Outcome: By enforcing the claim–grounds–procedure triad, propagating provenance through an algebra of annotations and logging justificatory acts in an immutable audit trail, the architecture ensures that acceptance is simultaneously explainable, revisable and verifiable.

Justification ceases to be a retrospective gloss on outputs and becomes a constitutive, measurable property of belief itself, integrated with probabilistic support, logical validity and accountable record-keeping [69,44,68].

https://doi.org/10.64142/jeai.1.3.19

Hybrid symbolic-statistical architectures and anchoring

A hybrid architecture separates generation from endorsement: Statistical components pro- pose candidate hypotheses, parses and actions, while a symbolic core evaluates, justifies and commits only those items that satisfy logical, semantic and epistemic constraints. The statistical layer supplies graded evidence embeddings, likelihoods, posteriors learned from data [74,11]. The symbolic layer encodes formal structure (types, terms, formulas), inference rules and acceptance criteria, ensuring that endorsed propositions possess determinate content, valid derivations and stable roles within the belief base. This division of labour harnesses pattern-learning without surrendering control over truth-preserving reasoning [66].

Representations and formal semantics: Linguistic and perceptual inputs are mapped to typed, compositional structures whose meanings are fixed by an interpretation into models, so that assertions carry explicit truth conditions rather than merely distributional similarity [40,2]. Proof-theoretic discipline (*e.g.*, Gentzen-style calculi) constrains admissible inferences and links derivability to semantic validity [19]. Category-theoretic semantics provides structure-preserving maps from syntax to meaning, supporting modular composition and principled abstraction across representational layers [16]. For epistemic and doxastic content, modal formalisms equip the system with operators for knowledge and belief and with accessibility relations that capture information flow over time and viewpoints [50,28,59,38].

Neural proposal, symbolic verification: Neural encoders propose candidate logical forms, entity links and state transitions with graded scores; the verifier checks well-typedness, satisfiability in the intended model class and compatibility with standing commitments before admission to B_t. Probabilistic relational formalisms (*e.g.*, Bayesian logic programs) bridge learned statistical relations and first-order structure, enabling the system to treat likelihoods as evidence while preserving symbolic dependencies for justification and repair [62,46]. In sequential settings, partially observable control models convert graded state estimates into decisions while remaining subject to the symbolic layer's constraints on commitment and revision [43].

Anchoring and world-connection: Anchoring ties symbols to determinate referents and usable skills. To avoid purely syntactic drift, the architecture binds terms to perceptual categories, executable procedures and invariants that fix reference over time, addressing the symbol grounding problem [64]. Developmental and embodied constraints further stabilise concept acquisition and use, linking perception, action and language in a mutually constraining loop that resists representational decoherence [6]. For explanatory adequacy and intervention, a causal semantics distinguishes association from manipulation and supports counterfactual reasoning and planning, ensuring that endorsed claims carry actionable content beyond correlation [39].

Justification, provenance and audit: Endorsed items are integrated into a justification graph that records premises, rules and evidential links; every transformation propagates provenance so that the contribution of statistical evidence to symbolic conclusions is explicit and replayable [69,44]. This enables targeted retraction when upstream information changes and supports external audit of how a conclusion was reached, by whom (or by which module) and under which conditions.

Outcome: The result is a pipeline in which learned regularities accelerate proposal and ranking, while a formally grounded core enforces meaning, validity and commitment. Hybridisation thus delivers agents whose outputs are not merely high-likelihood strings but justified claims anchored to models of the world, equipped for revision, explanation and action within a single, coherent epistemic economy [66,40,59,38,39].

Immutable Records and Epistemic Agency

This section specifies how the agent's epistemic life is bound to an append-only, tamper- evident record so that assertions become accountable acts rather than transient outputs. We introduce truth records that cryptographically bind each acceptance, inference and revision to its supporting grounds and procedure; define the commit structure that chains these events to provide ordering, authorship and non-repudiation; and show how provenance pointers from the belief and justification graphs enable replay, responsibility assignment and targeted repair.

We detail the verification pathway in which heavy computations are certified without re-execution, outline deployment choices ranging from authenticated logs to permissioned or public ledgers and explain how privacy is preserved by anchoring commitments (not plaintext) with controlled disclosure.

Finally, we connect these mechanisms to minimal selfhood and artificial agency: The system maintains a persistent identity, owns its commitments over time and exposes auditable evidence for why it believed, acted and subsequently

https://doi.org/10.64142/jeai.1.3.19
Journal of Engineering and Artificial Intelligence
Research Article

revised, thereby turning epistemic integrity into an enforceable interface with the outside world.

Blockchain as epistemic ledger

An epistemic ledger provides an append-only, cryptographically verifiable record of justificatory acts: Admissions of propositions into B_t, revisions and contractions and the derivations that connect claims to grounds. Concretely, each state transition $(B_t, J_t) \longrightarrow (B_{t+1}, J_{t+1})$ emits a commit that binds the delta Δ_t (new acceptances, retractions and updated justification links) to a content digest. The commit is a tuple

$$C_t = \left\langle \mathsf{prev} = h(C_{t-1}),\ \mathsf{root} = r_t,\ \mathsf{meta} = \mu_t,\ \mathsf{proof} = \pi_t \right\rangle$$

where h(·) is a collision-resistant hash, r_t = merkle($\{h(x)$: x ∈ $\Delta_t\}$) commits to the changed items *via* a Merkle accumulator, μ_t records timing and authority metadata and πt attests to validity under the ledger's consensus or audit rule. Linking commits by $h(C_{t-1})$ yields an immutable sequence in which tampering with any past justification invalidates all subsequent headers, providing efficient, public detectability of changes [60,7,36].

Provenance and justification integrate at the data model. Each proposition $\varphi \in B_t$ and each support relation in J_t is given a stable identifier; the commit references these identifiers and stores cryptographic bindings to their serialised content. This aligns with provenance formalisms in which annotations propagate through transformations and can be queried or replayed to reconstruct derivations [69,44]. As a result, standard questions why φ?, which evidence supported φ?, which inference produced φ? become queries over ledgered pointers and hashes, with verifiable correspondence between current beliefs and the recorded history of justificatory acts.

For external verification of intensive computations (*e.g.*, re-closing B_{t+1} under consequence, re-scoring probabilistic evidence), the ledger can store succinct certificates rather than raw traces. Interactive and non-interactive proof systems enable a verifier to check correctness without recomputing the entire derivation, reducing audit cost while preserving soundness [61, 9]. The proof field πt thus generalises from consensus attestations to proof objects that certify semantic validity (*e.g.*, a proof-of-derivation for a newly accepted ψ) or correct execution of revision policies.

Because epistemic records often carry sensitive information, the ledger stores commitments (hashes, Merkle roots) and access-controlled references rather than plaintext justifications; disclosure is managed off-chain while integrity is guaranteed on-chain. This design retains the auditability and non-repudiation afforded by hash chaining and digital signatures, while allowing provenance to remain private by default and selectively revealed under policy. Limitations and design choices are explicit. Consensus mechanisms and data models must be matched to epistemic throughput and latency: permissioned chains or append only authenticated logs can deliver predictable performance and governance, whereas open consensus may be unnecessary for single-authority agents [7,32]. The ledger cannot, by itself, guarantee truth; it guarantees that the history of reasoning is tamper-evident and that any asserted belief is tied to an auditable chain of grounds and procedures. Formal analyses of blockchain security and reliability clarify attack surfaces (reordering, equivocation) and motivate commitment structures that preserve ordering and authorship of epistemic acts [36]. Beyond integrity, the ledger's epistemic role is conceptual as well as technical: It provides a public, time-ordered account of how knowledge claims are formed, defended, and, when necessary, withdrawn [73].

In summary, the ledger realises three invariants for epistemic agency: Immutability (past justificatory acts are fixed *via* hash chaining), accountability (every acceptance, inference and revision carries signatures and provenance links) and verifiability (auditors can check, *via* π_t, that rule-governed procedures were executed correctly). Together with the provenance layer, these mechanisms make justification traceable, reproducible and resistant to post hoc fabrication, embedding accountability into the very substrate of belief management [69,44,7,60,36].

Truth records and cryptographic finality

A truth record is the canonical, externally verifiable artefact that binds a committed proposition to its justificatory state at a particular time. Formally, for an accepted claim $\varphi \in B_t$ with justification subgraph $J_t(\varphi)$, the system emits:

$$\mathsf{TR}_t(\varphi) = \left\langle \mathsf{id}_\varphi,\ \mathsf{hash}\big(\varphi, \mathcal{J}_t(\varphi)\big),\ \tau_t,\ \mathsf{sig}_A \right\rangle$$

where id_φ is a stable identifier, hash(·) commits to the claim and its proof/grounds, τ_t times- tamps the state and sig_A is a digital signature of the accountable agent. These records are appended to an authenticated, hash-linked log so that any subsequent alteration of content is publicly detectable. The ledgered pointers from $\mathrm{TR}_t(\varphi)$ into the justification graph make the pathways from grounds to claim auditable and replay able [69,44].

Commitment and ordering: Truth records are chained by including the digest of the previous block of epistemic actions, yielding an append-only structure in which ordering is enforced by collision-resistant hashing and signatures. In open networks, Nakamoto-style proof-of-work (PoW) consensus achieves a public total order whose security derives from the computational difficulty of rewriting the chain's prefix [60]. Analyses of blockchain security characterise practical attack surfaces (*e.g.*, reordering, double-spend analogues) and model the cost of reorganisations as a function of adversarial hash power and confirmation depth [36,7]. In organisational deployments where a single authority or a consortium governs truth records, permissioned consensus or authenticated append-only logs can deliver predictable latency and governance without the expense of PoW [7,32].

Cryptographic finality: Let C_t be the block that first includes $\mathrm{TR}_t(\varphi)$. Under PoW with an honest majority, the probability that an adversary will successfully reorganise the chain to exclude C_t decays rapidly with the number of successor blocks k added on top of C_t; thus, after a policy threshold $k^\star$, the record is treated as final for operational purposes [60,36]. In permissioned settings, finality is often immediate once a qualified set of signers has co-signed C_t, with safety anchored in digital signatures and quorum assumptions rather than cumulative work [7]. In both cases, "finality" names a cryptographic and governance condition: Reversing the record would require either infeasible computation (open PoW) or a colluding quorum (permissioned), making post hoc fabrication of justificatory history detectable or prohibitively costly [32].

From records to proofs: To avoid storing voluminous derivations on-chain, the system anchors commitments to justificatory objects and supplies succinct verifiers. Interactive and non-interactive proof systems enable third parties to check that a claimed derivation or policy-compliant revision was performed correctly without re-executing it in full [61,9]. Accordingly, each block can carry certificate pointers (and their hashes) that attest: (i) that φ was derived by admissible rules from recorded premises; (ii) that any contraction/revision satisfied the policy (*e.g.*, minimal mutilation) and (iii) that the current B_t is consistent and closed under the relevant consequence relation.

Epistemic status and audit: Truth records serve two roles. First, they provide an identity condition for commitments: A later claim of holding φ is tied to an immutable anchor that fixes which grounds and which procedure licensed acceptance at time *t*. Second, they enable forensic audit: An external reviewer can verify integrity (hash chain and signatures), provenance (links into J_t) and correctness (proof certificates) without trusting the agent's internal state. This aligns the ledger's integrity guarantees with the epistemic aim of making justificatory history non-repudiable [7,36]. From an epistemological perspective, the ledger does not confer truth; it secures accountability for how truth-claims were formed and maintained, a distinction articulated in recent analyses of blockchain as an epistemic technology [73].

Policy and limitations: Finality parameters ($k^\star$, quorum requirements), retention policies for proof artefacts and privacy controls must be matched to domain constraints. Open ledgers prioritise censorship-resistance and public verifiability; permissioned ledgers emphasise governance and throughput [7,32]. In all cases, the ledger's guarantees are orthogonal to semantic correctness: They render the history tamper-evident and auditable, but the quality of justification still depends on the evidential and inferential standards enforced by the reasoning core and provenance layer [69,44]. By coupling those standards to cryptographic finality, truth records become durable anchors for epistemic responsibility.

Minimal selfhood and artificial accountability

A reasoning system is accountable only if it sustains a minimal self: A persisting, agent- level locus of commitment that binds utterances to an author, stores standing intentions and policies and bears the normative consequences of assertion and action. In this architecture, minimal selfhood is not a metaphysical thesis but a structural role realised by identifiers, control state and justificatory records. The agent *A* maintains a stable identity id_A, an endorsement relation $\mathrm{Acc}_t(\varphi)$ over propositions and a policy base that governs when to assert, retract or suspend. Discursive commitment the idea that to assert is to undertake inferential and justificatory obligations supplies the normative core of this role [58]. Epistemic status is not exhausted by internal credence; it is a public standing conferred by what the agent is prepared to defend and revise under recognized rules [8,71].

Commitment, intention and control: Minimal selfhood requires control over assertion and change. Practical structures of intention and plan govern when beliefs may license action, how defeaters are handled and which stakes justify assertion versus suspension [47]. In agent architectures, intentions, beliefs and desires interact to yield accountable behaviour; plans thereby mediate between epistemic state and action in a way that is inspectable and revisable [4]. Responsibility demands that the system be able to cite not only what it believes but why and how those beliefs informed downstream choices a point sharpened by accounts of knowledge as information flow and control [18,65].

Second-order representation of doxastic state: Self-ascription of attitudes (B_{Ap}, K_{Ap}) and their dynamics are expressed in an epistemic/doxastic logic with well-defined semantics, enabling the system to reason about its own commitments and those of others [28,59,38]. Internally, beliefs are represented in a format suitable for deduction and query, with explicit dependencies, so that endorsement carries operational consequences rather than being a mere tag [41,25]. This second-order layer allows the agent to recognise when it is committed to q by having asserted p and $p \rightarrow q$, to register defeaters and to plan coherent repairs.

Public accountability provenance and audit: Artificial accountability requires that commitments be externally checkable. Every acceptance, inference and revision is linked to a justification subgraph and to provenance records that can be queried, replayed and compared [69,44]. To make the history of commitments non-repudiable, the system anchors truth records content-addressed digests of claims and their grounds in an authenticated, append-only log, optionally a blockchain when public ordering and independent verification are required [7,36,60]. The result is a durable mapping from present assertions to past justificatory acts, turning "who said what and why" into a verifiable interface rather than a matter of trust [73].

Norms of assertion and epistemic virtue: Minimal selfhood is partly constituted by the norms that govern assertion, retraction and apology (repair). Truthfulness as a disposition to assert only what one takes to be justified and to correct errors is required for the public practice of giving and asking for reasons [8,58]. These norms are encoded procedurally: Thresholds for acceptance, policies for suspension under uncertainty and obligations to revise when contradictions or stronger counterevidence arise. Because these norms are visible at the interface (through explanations and records), they convert private computation into public accountability.

Outcome: By combining (i) a persistent identity and control over commitment; (ii) an explicit second-order model of its own doxastic state and (iii) verifiable records of justificatory acts, the agent attains the minimal selfhood needed for artificial accountability. Assertions become undertakings that can be inspected, challenged, and, where necessary, withdrawn under shared rules aligning computational competence with the social and epistemic practices that make knowledge claims credible [59,38,69,44,7,8].

Design Blueprint and Philosophical Implications

This section consolidates the architecture into a deployable blueprint: It defines the core modules (belief management; inference with knowledge graph; metacognitive supervision; contradiction detection and revision; semantic anchoring; provenance and audit; calibration/safety/decision), the data contracts between them and the invariants each boundary enforces so that proposal is cleanly separated from endorsement and every commitment is traceable. It then sets out the philosophical stakes of the design why propositional commitment, closure and principled change constitute cognition rather than mere prediction; how epistemic risk and stability legitimise categorical acceptance; how immutable records ground responsibility and truthfulness; and where formal and practical limits constrain guarantees. The blueprint supplies the operational grammar for building agents whose utterances are justified, auditable acts bound to a persistent self, while the analysis clarifies why these constraints are not optional engineering choices but the conditions under which knowledge claims can be made at all.

Functional overview of key architectural modules

The architecture is modular, with well-defined interfaces that separate proposal from endorsement and generation from justification. Each module contributes a necessary condition for epistemic integrity: Formal semantics and sound proof ensure truth-preserving inference; probabilistic discipline and decision thresholds regulate acceptance; metacognitive control maintains coherence over time; provenance and audit render justificatory history verifiable. The design assumes a classical core with model-theoretic truth and proof-theoretic validity as invariants [1,2,19], complemented by epistemic/doxastic formalisms for reasoning about informational attitudes [28,59,38].

Belief Management Module (BMM): The BMM maintains the active belief base B_t and implements categorical acceptance as a policy over graded credences. Degrees of belief obey the probability axioms and are updated by conditionalisation (and its generalisations for graded input) [5,56,13,54]. Categorical uptake occurs when credence surpasses stability aware thresholds that balance error and omission costs [23]. When new information arrives, B_t is revised under minimal mutilation, following partial-meet and entrenchment schemes from belief-change theory [10,53,72].

Inference Engine and Knowledge Graph Interface (IE+KGI): The IE realises consequence with a sound proof system; the KGI provides typed, compositional representations whose meanings are fixed by interpretation into models, ensuring that derivations track truth conditions rather than mere symbol flow [40,2,19]. Modal and epistemic operators (K, B) support reasoning about knowledge, belief and information flow [50,59,38]. For uncertain structure, probabilistic formalisms bridge statistical relations and first-order dependencies [22,62]. Causal semantics distinguishes association from intervention so that endorsed conclusions support counterfactuals and planning [39]. Distributional features from language processing contribute evidential proposals but not meanings per se [11]; category-theoretic semantics supports structure-preserving mappings between syntax and meaning [16].

Metacognitive Supervisory Control Unit (MSCU): The MSCU maintains second order assessments of confidence, stability and centrality over B_t and triggers diagnosis, defence, contraction or revision to preserve invariants. It draws on metacognitive control in AI and cognitive architectures, where symbolic state, control knowledge and meta-level policies interact to sustain performance [48,34,33]. Practical reasoning policies align assertion, suspension and action with intentions and plans [47,4]. Resource-bounded control trades accuracy for time and computation while respecting epistemic guarantees [27,67].

Contradiction Detector and Belief Revision (CDBR): The CDBR enforces consistency ($B_t \nvdash \bot$) and closure (if $B_t \vdash \varphi$ then $\varphi \in B_t$) using proof-theoretic monitors and semantic checks [19,2]. On detecting a clash, it computes minimally inconsistent sets *via* justification links and applies belief-change operators to restore coherence with least disturbance [35,10,53,72]. Stability constraints prevent oscillation at categorical thresholds [23].

Semantic Anchoring and Grounding (SAG): SAG ties symbols to determinate reference and usable skills to prevent drift into purely syntactic manipulation. Terms are linked to perceptual categories and executable procedures, addressing the symbol grounding problem and developmental constraints stabilise concept acquisition over time [64,6].

Provenance and Audit Layer (PAL): PAL records the claim–grounds–procedure triad for each accepted proposition in a justification/provenance graph, enabling replay, responsibility assignment and targeted repair [69,44]. Integration with an authenticated, append-only ledger yields tamper-evident, time-ordered truth records for admissions, derivations and revisions [60,7,36]. Succinct certificates permit external verification of heavy computations without re-execution [61,9]. Deployment choices range from open consensus to permissioned logs depending on governance and throughput requirements [32,73].

Calibration, Safety and Decision Interface (CSDI): CSDI calibrates credences using strictly proper scoring rules and regulates acceptance thresholds in light of costs, mitigating overconfidence and misreporting [68]. Safety and alignment policies impose robustness and truthfulness standards under distribution shift and adversarial conditions, with tests that bind acceptance and action to verified justification quality [12]. For acting under uncertainty and partial observability, the decision interface couples epistemic state to planning mechanisms while respecting the architecture's integrity constraints [43,65].

Interfaces and data flow: Proposal flows from statistical encoders to the IE+KGI; endorsement flows from the BMM after MSCU and CDBR checks; every transition emits provenance to PAL. Modules communicate through typed messages carrying formulas, credences, derivation certificates and provenance pointers. Invariants truth conditions, soundness, consistency, closure and traceability are enforced at each boundary, ensuring that endorsed outputs are not merely high-likelihood strings but justified, auditable claims grounded in models, proofs and evidence.

Epistemic risk, integrity and system failure

Epistemic risk is the expected loss arising from erroneous acceptance, rejection or revision of propositions in the belief base B_t. Let Θ denote states of the world, A admissible epistemic actions (accept, suspend, contract, revise) and $L : \Theta \times A \longrightarrow R{\geq}0$ a loss functional. Given a credence Pr_t over Θ, the epistemic risk of action $a \in A$ is

$$\mathcal{R}_t(a) \;=\; \mathbb{E}_{\theta\sim\Pr_t}\big[L(\theta,a)\big]$$

and policies are chosen to minimise R_t subject to integrity constraints (consistency, closure, traceability). This connects rational acceptance to the probabilistic calculus (additivity, conditionalisation) and to decision-theoretic admissibility [5,56,13,45]. Coherence constraints are buttressed by accuracy-based and Dutch Book arguments: Credences that violate probability or conditional coherence expose the agent to guaranteed long-run loss, motivating probabilistic norms as safeguards against epistemic failure [29,51,68].

Integrity invariants and failure modes: Epistemic integrity requires maintaining (i) consistency ($B_t \nvdash \bot$), (ii) closure (if $B_t \vdash \varphi$ then $\varphi \in B_t$) and (iii) traceability (every $\varphi \in B_t$ carries provenance and justification). Proof-theoretic discipline and model-theoretic truth-conditions enforce the first two invariants [19,2].

Violations define system failures: Contradiction (triviality risk), non-closure (stagnation risk) and opacity (un-auditable claim risk). Because integrity is constitutive rather than cosmetic, the agent treats any breach as a fault that must trigger diagnosis and repair rather than as a tolerable steady state.

Quantifying acceptance risk: Acceptance of *p* at time t is governed by a threshold pol- icy that trades off false acceptance and false rejection under current Pr_t. Let C(*p*) denote the cost of accepting p when ¬p holds and M (*p*) the cost of missing a true *p*. The Bayes-optimal decision boundary is a likelihood-ratio or posterior threshold that minimises R_t given C and M, with categorical acceptance reserved for credence regions that are robust under small perturbations [13,23]. Strictly proper scoring rules supply a calibration discipline: They reward credences that match empirical frequencies and penalise miscalibration, preventing systematically overconfident commitments [68]. Where uncertainty is imprecise or interval-valued, the policy generalises to dominance-respecting acceptance under imprecise probability [54].

Detection, revision and stability: When inconsistency is detected, the system computes minimally inconsistent sets using justification links and applies minimal mutilation change to restore integrity. Contraction and revision follow partial-meet/entrenchment schemes from belief-revision theory, guided by ordinal conditional rankings to retract the least entrenched elements first [10,53,72]. Stability constraints link categorical acceptance to robust basins in credence/rank space, preventing oscillation across updates and thereby reducing the risk of policy thrashing [23]. Dependency-directed truth-maintenance ensures that repairs are local and auditable rather than global rollbacks [35].

Propagation and fragmentation risk: Even when local acceptance is well-calibrated, the agent fails epistemically if logical consequences are not realised. The inference kernel enforces closure so that if *p* and ($p \rightarrow q$) are held, then q is derived and justified, mitigating fragmentation risk.

Epistemic and doxastic formalisms provide operators and frames for reasoning about knowledge/belief and their dynamics, supporting sound propagation of higher-order commitments under observation and announcement [28,59,38,24]. This couples local probabilistic discipline to global logical completeness.

Grounding, provenance and audit risk: Unanchored symbols and opaque derivations create grounding and audit risks. Semantic anchoring ties well-typed expressions to interpretations; provenance mechanisms record the claim–grounds–procedure triad for each acceptance, enabling replay, responsibility assignment and targeted repair [40,64,69,44].

To make tampering detectable and ordering unambiguous, commits of justificatory acts are hashed, linked and optionally placed on an authenticated ledger, yielding non-repudiable, time-ordered truth records [60,7,36].

Safety, alignment and decision risk: Epistemic failures propagate into decision failures when actions are taken on the basis of unsupported or unstable beliefs. Resource-bounded control policies regulate inquiry depth and revision frequency, aligning epistemic objectives with practical constraints while maintaining core guarantees [27,67]. Safety-oriented alignment work underscores the need for robust, testable standards for truthful behaviour, especially under distribution shift and adversarial pressure [12]. The architecture therefore binds acceptance to justifications that are not only internally valid but externally checkable, reducing the risk that high-stakes outputs rest on unverified premises.

Summary: Epistemic risk is managed by (i) probabilistic coherence and calibrated acceptance; (ii) classical integrity invariants with principled repair; (iii) closure and higher-order propagation and (iv) provenance with tamper-evident audit. Failure is defined and handled at the architectural level, so that the agent's beliefs remain truth-preserving, stable and accountable over time [19,2,10,72,69,44].

Limits of formalisation and future directions

Formalisation enables explicit truth conditions, proof systems and principled update, but it also imposes limits that shape what an epistemic architecture can guarantee. Model-theoretic accounts fix truth *via* satisfaction in structures, yet the choice of intended models and languages trades expressive power against decidability and effective inference [1,2]. Proof theory provides discipline for consequence, but even elegant calculi face inherent constraints: No single recursively axiomatized system is both maximally expressive and computationally tractable across all target domains [19]. In probabilistic settings, rich languages that quantify over events and individuals complicate definability and update, highlighting subtle interactions between syntax and probability [22,37]. At the meta-mathematical level, information-theoretic incompleteness bounds remind us that no finite axiomatization can capture all arithmetical truths, placing principled ceilings on deductive completeness within any fixed formal core [21].

Learning-theoretic considerations introduce further limits. Statistical learning theory characterizes generalization by capacity measures and sample complexity, showing that high accuracy on finite data does not guarantee reliable extrapolation without structural bias [70]. In the PAC framework, learnability hinges on hypothesis class properties and toler- able error, clarifying when and why inductive success is possible at all [42]. These results constrain how credence assignments and acceptance thresholds should be set if the system is to avoid unwarranted confidence under distributional shift. Accuracy-based and Dutch Book arguments anchor probabilistic coherence and calibration, but they do not resolve all questions of representation choice or prior selection [68,51,13,5,56].

A separate family of limits concerns meaning and grounding. Purely syntactic manipulation does not by itself secure determinate reference; ties between symbols and the world must be engineered so that assertions connect to discriminations, predictions and interventions [64]. Resource constraints also bound ideal reasoning: agents must trade accuracy for time and memory, motivating procedures that remain epistemically responsible under bounded rationality [26,27,67]. In multi-agent and higher-order settings, formal logics of knowledge and belief provide powerful tools, but introspection, common knowledge and dynamic updates raise complexity and modeling challenges that must be managed in practice [28,59,38,24].

Paraconsistent systems illuminate designs that tolerate contradictions without explosion, but toleration complicates downstream reasoning interfaces; the present architecture prefers classicality with explicit, auditable repair [20,49,31]. Coherence criteria themselves face tensions: demanding full closure risks brittleness; relaxing it risks fragmentation. Recent stability conditions link categorical acceptance to robust credence regions, offering a principled middle path [23]. Where evidence is imprecise, interval-valued or set-based credences can preserve dominance and caution while remaining compatible with probabilistic norms [54].

Future directions: Neural–symbolic integration. Strengthen the proposal–verification pipeline by coupling representation learning with typed, model-theoretic semantics and proof-theoretic verification. Surveyed approaches identify patterns for leveraging statistical modules without ceding justificatory control [66,40,16,74]. Richer probabilistic logics. Extend first-order probabilistic reasoning with well-defined update and query semantics, drawing on foundational analyses of probability over expressive languages and epistemic formalisms for belief and knowledge [22,37,59,38]. Causal explanation. Integrate structural causal models so that accepted claims support counterfactuals and intervention planning rather than mere correlation, aligning epistemic endorsement with actionable semantics [39]. Provenance and verification. Make justification replayable and auditable at scale *via* provenance semirings and interoperable provenance records, combined with succinct verification of heavy computations [69,44,61,9]. Immutable audit. Anchor justificatory acts to authenticated, append-only ledgers that provide ordering, integrity and non-repudiation appropriate to the deployment context [60,7,36,32]. Grounding and embodiment. Pursue developmental and embodied routes to concept stability and referential reliability, linking perception, action and language to resist semantic drift [6]. Planning under uncertainty. Couple the epistemic core with decision procedures for partially observable domains, ensuring that actions respect belief quality and revision policies [43]. Safety and alignment. Embed testable standards for truthful behaviour and robustness, especially under distribution shift and adversarial conditions [12].

Outlook: The limits of formalisation are not a counsel of despair but a design brief. They delineate where guarantees are possible, where approximations are mandatory and where auxiliary structures grounding, provenance, verification and governance must shoulder part of the epistemic load. By developing architectures that acknowledge these boundaries while exploiting the strengths of formal semantics, proof, probability and learning, the next generation of systems can maintain epistemic integrity under real-world constraints [2,19,5,70,59,69].

Conclusion

This work set out the conditions under which an artificial system's outputs count as knowledge- bearing acts rather than plausible strings, specifying an architecture that unites probabilistic discipline, sound inference, semantic anchoring, principled belief change, metacognitive supervision and tamper-evident justification records. Across the modules, three invariants govern every update to the belief base consistency, closure and traceability so that acceptance is earned, consequences are realised and histories are auditable. Propositional commitment converts graded credence into accountable endorsement; revision policies restore coherence with minimal disturbance; provenance and immutable records bind present claims to past grounds and procedures; and a persistent agent identity links utterance to responsibility. The resulting blueprint is both operational and normative: It provides the interfaces and checks needed to build systems that explain, correct and defend their assertions and it clarifies why these constraints are constitutive of epistemic agency. Implemented faithfully, the design yields agents that maintain truth-preserving belief over time, act under uncertainty without forfeiting accountability and expose the evidential and inferential pathways through which they know what they claim to know.

Summary of Contributions

- **Epistemic architecture:** A complete, implementable architecture for artificial reasoning systems that treats belief as a persistent, inferentially active state. The design specifies invariants (consistency, closure, traceability) and enforces them across all updates to the belief base B_t.

- **Propositional commitment and thresholds:** A rule-governed acceptance policy that converts graded credences into categorical commitments *via* stability-aware thresholds. Commitments are auditable states (not transient outputs) and are accompanied by prove- nance metadata, enabling principled retraction and defence.
- **Belief revision with minimal mutilation:** A diagnosis–repair pipeline that detects contradictions, computes minimally inconsistent sets and applies contraction/revision guided by entrenchment and ordinal rankings. The result is coherence restoration with least disturbance to justified content.
- **Metacognitive supervision:** A second-order control layer that represents and monitors the status of beliefs (confidence, stability, centrality), triggers defence or repair, allocates computational effort under bounded rationality and aligns epistemic operations with task goals.
- **Hybrid symbolic–statistical substrate:** A separation of roles in which statistical components propose candidates (with scores) and a symbolic core endorses only those that satisfy typing, semantic satisfiability and inferential validity. Compositional, model- theoretic semantics and sound proof systems bind acceptance to meaning and truth-preservation.
- **Semantic anchoring:** Mechanisms that tie symbols to determinate referents (perception, action, causal structure), preventing drift into purely syntactic manipulation and ensuring that accepted propositions carry world-level content and counterfactual import.
- **Triadic justification:** A claim–grounds–procedure schema realised as a justification graph that records evidential inputs, admissible inference steps and resulting claims. Justificatory weight is computed from evidence quality, rule reliability and calibration, turning "why this belief?" into a standard query.
- **Provenance and immutable audit:** First-class provenance that propagates through derivations and an append-only, cryptographically authenticated ledger of justificatory acts. "Truth records" provide non-repudiable anchors linking present commitments to their historical grounds and procedures, with optional succinct certificates for external verification.
- **Internal truth constraints:** Continuous enforcement of consistency and closure through proof-theoretic monitors and semantic admissibility checks. Detected clashes trigger ranked, localised repairs rather than global rollback, preserving monotonic interfaces for downstream modules.
- **Epistemic risk and calibration:** A decision-theoretic account of acceptance that trades off false acceptance and false rejection under probabilistic coherence, with stability conditions and proper-scoring calibration to prevent overconfidence and oscillation.
- **Operational blueprint:** A modular decomposition belief management module, inference engine with knowledge graph interface, metacognitive supervisory control, contradiction detector and belief revision, semantic anchoring and grounding, provenance and audit layer and calibration/safety/decision interface with typed interfaces and data flows that enforce invariants at boundaries.
- **Scope and outlook:** An explicit statement of formal and practical limits (expressivity vs. tractability, learning-theoretic constraints, grounding and multi-agent dynamics) together with concrete directions for extension (richer probabilistic logics, causal explanation, verified provenance at scale and safety/alignment tests).

References

1. Tarski A (1935) Der wahrheitsbegriff in den formalisierten sprachen. Stud philos 1: 261-405. [Google Scholar]
2. (1956) English translation in J. H. Woodger (ed.), Logic, Semantics, Metamathematics. Oxford Uni Press.
3. Alfred Tarski (1956) Logic, Semantics, Metamathematics. Oxford Uni Press 1933-1935. [Google Scholar]
4. Goldman AI (1979) What is Justified Belief?. Just Knowl 1-23. [Crossref], [Google Scholar]
5. Rao AS, Georgeff MP (1991) Modeling rational agents within a bdi- architecture. Proc 2nd Int Conf on Knowledge Representation and Reasoning (KR) 473-484. [Google Scholar]
6. Kolmogorov AN (1956) Foundations of the theory of probability. Chelsea Publ Com.
7. Cangelosi A, Schlesinger M (2015) Developmental robotics: From babies to robots. MIT Press. [Google Scholar]
8. Narayanan A, Bonneau J, Felten E, Miller A, Goldfeder S (2016) Bitcoin and cryptocurrency technologies: A comprehensive introduction. Princeton Uni Press. [Google Scholar]
9. Williams B (2002) Truth and truthfulness: An essay in genealogy. Princeton Uni Press.
10. Parno B, Howell J, Gentry C, Raykova M (2013) Pinocchio: Nearly practical verifiable computation. IEEE Symp Secur Priv 238-252. [Crossref], [Google Scholar]
11. Alchourro´n CE, Ga¨rdenfors P, Makinson D (1985) On the logic of theory change: Partial meet contraction and revision functions. J Symb Log 50: 510-530. [Crossref], [Google Scholar]

12. Manning C, Schu¨tze H (1999) Foundations of statistical natural language processing. MIT Press. [Google Scholar]

13. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, et al. (2021) Aligning AI with shared human values. arXiv. [Crossref], [Google Scholar],

14. Jaynes ET(2003) Probability theory: The logic of science. Cambridge Uni Press. [Google Scholar]

15. Conee E, Feldman R (2004) Evidentialism: Essays in epistemology. Oxford Uni Press. [Google Scholar]

16. Bender E, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big?. FACCT 610-623. [Crossref], [Google Scholar]

17. Lawvere FW (1963) Functorial semantics of algebraic theories. Proc Natl Acad Sci USA 50: 869-872. [Crossref], [Google Scholar]

18. Dretske F (1981) The pragmatic dimension of knowledge. Philoso Stud 40: 363-378. [Google Scholar]

19. Dretske F (1981) Knowledge and the flow of information. MIT Press. [Google Scholar]

20. Gentzen G (1935) Untersuchungen u¨ber das logische Schließen. Math Zeit 39: 176-210. [Crossref], [Google Scholar]

21. Priest G (2006) In contradiction: A study of the transconsistent. Oxford Uni Press. [Google Scholar]

22. Chaitin GJ (1974) Information-theoretic limitations of formal systems. J ACM 21: 403-424. [Crossref], [Google Scholar]

23. Gaifman H, Snir M (1982) Probabilities over rich languages, testing and randomness. J Symb Log 47: 495-548. [Crossref], [Google Scholar]

24. Leitgeb H (2014) The stability theory of belief. Philos Rev 123: 131-171. [Crossref], [Google Scholar]

25. Ditmarsch H, Hoek W, Kooi B (2008) Dynamic epis- temic logic. 337 Springer. [Crossref], [Google Scholar]

26. Levesque HJ (1984) Foundations of a functional approach to knowledge representation. Artifi Intell 23: 155-212. [Crossref], [Google Scholar]

27. Simon HA (1972) Theories of bounded rationality. 161-176 North-Holland. [Google Scholar]

28. Simon HA (1976) From substantive to procedural rationality. Methodol Eco-nomics Social Sci 129-148. [Crossref], [Google Scholar]

29. Hintikka J (1962) Knowledge and belief: An introduction to the logic of the two no- tions. Cornell Uni Press. [Google Scholar]

30. Joyce JM (1998) A nonpragmatic vindication of probabilism. Phil Sci 65: 575-603. [Crossref], [Google Scholar], [Indexed]

31. Joyce JM (2009) Bayesianism. Stanf Encycl Philos.

32. Jean-Yves Beziau (2014) Paraconsistent logic: Consistency, contradiction and negation. Paraconsistency: Logic Appl 1-30.

33. Yli-Huumo J, Ko D, Choi S, Park S, Smolan- der K (2016) The current state of blockchain technology: Limitations and future directions. Proceedings of the IEEE 104: 2230-2242.

34. Laird JE (2012) The soar cognitive architecture. MIT Press. [Google Scholar]

35. Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, et al. (1997) ACT-R: A theory of higher level cognition and its relation to visual attention. Human-Computer Interaction 12: 439-462. [Crossref], [Google Scholar]

36. Doyle J (1979) A truth maintenance system. Artif Intell 231-272. [Crossref], [Google Scholar]

37. Bonneau J, Miller A, Clark J, Narayanan A, Kroll JA. et al. (2015) SoK: Research perspectives and challenges for bitcoin and cryptocurrencies. IEEE Symposium on Security and Privacy 104-121. [Crossref], [Google Scholar]

38. Halpern JY (1990) An analysis of first-order logics of probability. Artif Intell 46: 311-350. [Crossref], [Google Scholar]

39. Halpern JY (2003) Reasoning about uncertainty. MIT Press. [Google Scholar]

40. Judea Pearl (2009) Causality: Models, reasoning and inference. Cambridge Uni Press.

41. Devlin K (1991) Logic and Information. Cambridge Uni Press [Google Scholar]

42. Konolige K (1986) A deduction model of belief. Morgan Kaufmann [Crossref], [Google Scholar]

43. Valiant LG (1984) A theory of the learnable. Commun ACM 27: 1134-1142. [Crossref], [Google Scholar]

44. Kaelbling LP, Littman ML, Cassandra AR (1998) Planning and acting in partially observable stochastic domains. Artif Intell 101: 99-134. [Crossref], [Google Scholar]

45. Moreau L, Clifford B, Freire J, Futrelle J, Gil Y, et al. (2008) The open provenance model: An overview. Int J Digit Curation 3: 60-67. [Crossref], [Google Scholar]

46. Kaplan M (1996) Decision theory as philosophy. Cambridge Uni Press [Google Scholar]

47. Ginsberg ML (1989) A unified framework for planning and default reasoning. Artif Intell 39: 37-64.

48. Bratman ME (1987) Intention, plans and practical reason. Harvard Uni Press. [Google Scholar],

49. Cox MT (2005) Metacognition in computation: A selected research review. Artif Intell 169: 104-141. [Crossref], [Google Scholar]

50. da Costa NCA (1974) On the theory of inconsistent formal systems. Notre Dame J Form Log 15: 497-510. [Google Scholar]

51. Blackburn P, de Rijke M, Venema Y (2001) Modal Logic. Cambridge Uni Press. [Google Scholar]

52. Patrick Maher (1990) Dutch book arguments depragmatized: Epistemic consistency for conditional probabilities. J Philos 87: 396-410.

53. Boghossian P (2003) Epistemic analyticity: A defence. Philos Stud 106: 1-20. [Google Scholar]

54. Ga¨rdenfors P (1988). Knowledge in flux: Modeling the dynamics of epistemic states. MIT Press. [Google Scholar]

55. Walley P (1991) Statistical reasoning with imprecise probabilities. 42 Chapman and Hall. [Google Scholar]

56. Jeffrey RC (1983) The logic of decision. Uni of Chicago Press [Google Scholar]

57. Cox RT (1946) Probability, frequency and reasonable expectation. Am J Phys 14: 1-13. [Crossref], [Google Scholar]

58. Audi R (2003) Epistemology: A contemporary introduction to the theory of knowledge. Routledge. [Google Scholar]

59. Brandom R (1994) Making it explicit: Reasoning, representing and discursive com- mitment. Harvard Uni Press. [Google Scholar]

60. Fagin R, Halpern JY, Moses Y, Vardi MY (1995) Reasoning about knowledge. MIT Press. [Google Scholar]

61. Nakamoto S (2008) Bitcoin: A peer-to-peer electronic cash system. [Google Scholar]

62. Goldwasser S, Kalai YT, Rothblum GN (2008) Delegating computa- tion: Interactive proofs for muggles. ACM 113-122. [Google Scholar]

63. Natarajan S, Kersting K, Bui HH, Shavlik JW (2008) Bayesian logic programs: Combining symbolic and statistical approaches to artificial intelligence.

64. Lin S, Hilton J, Askell A (2022) TruthfulQA: Measuring how models mimic human falsehoods. arXiv. [Crossref], [Google Scholar]

65. Harnad S (1990) The symbol grounding problem. Pxhysica D: Nonlinear Phenomena 42: 335-346. [Crossref], [Google Scholar]

66. Russell S (1997) Rationality and intelligence. Artif Intell 94: 57-77. [Crossref], [Google Scholar]

67. Besold TR, d'Avila Garcez A, Bader S, Bowman H, Domingos P, et al. (2017) Neural-symbolic learning and reasoning: A survey and interpretation. arXiv. [Crossref], [Google Scholar]

68. Griffiths TL, Lieder F, Goodman ND (2015) Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. Top Cogn Sci 11: 393-406. [Crossref], [Google Scholar], [Indexed]

69. Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction and estimation. J Am Stat Assoc 102: 359-378. [Crossref], [Google Scholar]

70. Green TJ, Karvounarakis G, Tannen V (2007) Provenance semirings. Proc ACM SIGMOD 31-40. [Crossref], [Google Scholar]

71. Vapnik VN (1998) Statistical learning theory. Wiley.

72. Sellars W (1956) Empiricism and the philosophy of mind. Philos Sci 1: 253-329. [Google Scholar]

73. Spohn W (1988) Ordinal conditional functions: A dynamic theory of epistemic states.

74. Causation in Decision, Belief Change and Statistics 2: 105-134. [Google Scholar]

75. Xu X, Sandner P, Gipp B (2019) The epistemology of blockchain: A formal approach. Front Blockchain 2: 6.

76. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35: 1798-1828. [Crossref], [Google Scholar]