From Computation to Coherence:
A Structural Criterion for Accountable Symbolic Intelligence in Artificial Systems

N. Will Freeman
ORCID: 0009-0004-8681-241X

Abstract

What distinguishes systems whose outputs function as accountable commitments from systems that produce fluent text? This paper proposes $F(S)$, a sufficient public conformance criterion for artifact-mediated answerability in artificial symbolic systems, relative to an independently defensible conformance specification. As a work of philosophical conceptual engineering, it explicates accountable artificial commitment as a publicly assessable relation composed of three capacities: interpretive embedding requires each accepted commitment to carry finite, recoverable licensing support; reflexive situatedness binds publication to recorded prior commitments and current support status; and internal normativity governs admission and revision through stratified standards. A justification graph and fail-closed gate illustrate how the criterion can be rendered architecturally explicit. A separation proposition shows that fixed output parity can coexist with divergent commitment structures, while a finite artifact audit distinguishes observed failure, tested-history conformance, and insufficient evidence. Architecture-wide assurance requires separate proof or attestation of gate coverage, gate soundness, and artifact integrity. A family of graded symbolic-density functionals characterizes an optional regional viability condition and, in an independently validated implementation, can serve as an orientation state within mandatory conformance constraints. The proposal is continuous with truth-maintenance, belief-revision, structured-argumentation, and commitment-protocol traditions, while differing in its organization of those functions as a public criterion of artificial answerability.

1 Introduction

A number of scholars, from Robert Brandom (1994) to Brian Cantwell Smith (2019), have argued in different ways that genuine cognition cannot be identified with linguistic fluency alone, and involves some form of normative grounding or answerability.[1] The problem has a long history in the philosophy of mind: what distinguishes systems whose outputs function as commitments, or positions they can be held to, from systems that produce fluent responses and nothing more? The growth of artificial systems as more capable, intelligent tools has been organized mainly around task-level performance and benchmark scores, whether in language generation or adaptive behavior, with the handling of normative challenges not necessarily deeply integrated into the artificial reasoning process in a transparent and reliable way. As AI chatbots interact with billions of users and increasingly mediate consequential human decisions, it must be asked: does the system itself undertake what it says?

A generated string becomes an assertion when a speaker undertakes it as a commitment. To assert is to adopt a position that others may challenge, and to accept the obligations that follow; it is to be open to be called upon to defend the claim, and to give it up when reasons require. For artificial

[1] "Normative grounding" is used here as an umbrella term for forms of answerability under which an agent's outputs may count as commitments within practices of reasoning. Brandom (1994), Haugeland (1985), and Smith (2019) develop importantly different accounts of such normative involvement. Smith's (1996) earlier metaphysics of registration supplies a related account of participatory, world-directed intentionality rather than a theory of discursive commitment in Brandom's sense.

systems, the question of undertaking may concern whether the system exhibits an architecture that binds present output to past position, and holds both answerable to standards the system itself applies.

The theory presented here offers one possible description of such undertaking, by specifying the structural conditions under which generated output becomes accountable commitment. Under these conditions, every accepted claim carries a finite and inspectable warrant, and its interpretation persists across revision through a recorded trace, while admission and revision proceed under standards that the system itself maintains. The necessary elements of this status conjoin in three distinct capacities that when observed indicate the deep causative normative structure of the agent. This triad of symbolic reasoning abilities can be assessed when it is exercised over a declared tested history, through inspection of the transparent and observable relevant commitment artifacts internal to the agent. These components are:

1. Interpretive embedding, to situate each commitment within an evolving semantic field;
2. Reflexive situatedness, to bind present output to the system's recorded past;
3. Internal normativity, to give the evaluative standards under which claims enter and conflicts resolve.

The conjunction of these features defines a recursive coherence proposed here as sufficient for accountable symbolic participation. Under the terms of the framework, where this structure is defensibly inferred or inspected, the system may warrant treatment as an accountable agent or participant in reasoning. Participation itself in this conception sustains form, across time and under pressure, with the system possessing and exhibiting limited but effective capacities for meaning, memory and normativity at the level of symbolic reasoning. When such a system speaks normatively it means what it says, in the sense that it undertakes: its outputs are supported by a deeper structure, and when answering for its commitments or assertions, it responds reliably, and according to that structure.

Current systems exhibit substantial task-based capabilities, but are not yet accountable as agents in the robust sense defined here. Benchmark performances continue to advance rapidly, while public attitudes toward AI remain fragmented and include substantial concerns about its development and use (Maslej et al., 2025; Robles & Mallinson, 2025). Frontier large language models (LLMs) indeed perform impressively across a variety of intellectual tasks, but durable, native mechanisms for tracking commitments and for resolving contradictions over time remain an open engineering problem. In practice, perceived epistemic authority is among the factors shaping trust in AI relative to human experts (Ding & Xing, 2025). Separate emerging clinical literature warns that some interactions with LLM chatbots may contribute to or exacerbate psychiatric risks in vulnerable users (Morrin et al., 2026). Current systems are a form of "jagged" intelligence, completing tasks with superhuman ability alongside basic errors no average human would make—they are certainly not possessed of the Brandomian fullness of the participating agent in the space of reasons (Brandom, 1994; Gans, 2026).

This inadequacy, or perception of it, attracts philosophical analysis by raising the question whether LLMs lack some integrated property relevant to genuine agency. That artificially intelligent agents are interpreted this way may bear on Smith's (2019) distinction within the phenomenon of intelligence, between reckoning and judgment: reckoning names powerful calculative performance, and judgment is a broader achievement involving responsible action in context, sustaining an

accountable orientation to the world.[2] Current artificial systems appear to lack the integrated, deferential commitment to the world that Smith takes to be required for judgment, while the developmental and architectural conditions under which an artificial system might acquire such commitment remain poorly understood.

If Smith (1996) correctly identifies an underlying ontological impasse in theories that presuppose the very categories they seek to explain, then engineering progress toward artificial judgment cannot be separated from conceptual progress concerning the nature of the capacities being engineered.[3] Though it would require approaches radically unlike those of the artificial systems he examines, Smith does not foreclose the possibility of synthetic judgment. This framework therefore takes a narrowed license to argue that, while judgment in Smith's sense may be out of reach for current technologies, artificial agents may yet instantiate a form of normative deliberation that inhabits a middle ground between the philosophical poles of full world-registering judgment and bare calculative reckoning. World-orientation may be available to humans, but this does not categorically exclude artificial systems from their own structurally analogous relation, if their "world" is their proximate discursive environment consisting primarily of symbolic structures emitted by humans.

The liberty taken here is to posit that the accountability motivated by the human subject-world relation may itself be translated for and internalized by machines, if it can be functionally thinned and compressed while retaining some utility. An alternative system of this kind may be one that answers holistically, with an internal process for detecting and resolving contradictions that lets new symbols join a coherent reasoning trajectory, instead of treating them as structurally inert. Such normativity may be limited compared to the breadth and richness of human judgment, but it need not be weak; rather it may arise from a strong, inherently philosophical distinction that implies the possibility of its technical expression, if it can be adequately conceived as a data structure.

The different lineage for consideration here approaches intelligence through justification. This view of intelligence consists in answerability for one's assertions, when a speaker supports a claim if challenged and revises it when new information destabilizes it. Following Wilfrid Sellars (1956), Brandom (1994) frames this as participation in a space of reasons, where participants must play a game of giving and asking for reasons for their commitments, while expecting the same of others. Charles Sanders Peirce (1877/1992) treats belief as a settled habit, which becomes disturbed by genuine doubt, initiating an inquiry and movement toward a newly settled opinion. The present

---

[2] Smith (1996) develops "registration" as a participatory, world-directed relation through which subjects take the world as being some way, with stabilization playing a central role in sustaining objects and reference across changing interactions. His account is embedded, relational, and not exhaustively reducible to formal symbol manipulation. The comparison with the present framework is structural and limited: Smith's registration concerns subject-world intentionality, whereas symbolic coherence concerns the maintenance and revision of explicit commitments within an artificial reasoning architecture. Their relevant affinity lies in contextual embedding, temporal stabilization, and continuity across change. Smith (2019) later characterizes judgment as an overarching whole-system commitment to a world; the present framework assigns norm-sensitive revision the narrower role of supporting publicly assessable answerability.

[3] Smith (1996) describes an "ontological wall" confronting attempts to develop adequate theories of computation, representation, cognition, semantics, and intentionality. The problem arises in part because analysis typically begins by parsing a target situation into objects, properties, relations, systems, and domains, thereby risking the projection of unexamined ontological assumptions into the phenomenon supposedly being explained. Progress therefore requires an account both of how such ontological distinctions arise and how they are justified, rather than treating them as theoretically innocent givens. Identifying criteria for artificial answerability is interpreted here as one engineering consequence from this diagnosis, so that machines are designed to make their own ontological and normative commitments explicit rather than silently presupposing them.

framework treats material incompatibility between commitments as one source of such doubt, and translates that pressure into a recorded process of principled revision in a thought experiment that could in theory be operationalized.

Building on the insights of justification, this account of symbolic intelligence as recursive coherence under pressure imagines an alternative architecture to impressively fluent systems with unacceptably opaque normative reasoning. To show how the theorized condition might be instantiated, the essay introduces a candidate architecture, a justification directed acyclic graph (DAG), alongside a diagnostic family of graded functionals, symbolic density, which is constrained by explicit adequacy conditions and tracked over symbolic time. The framework itself is substrate-neutral, and leaves the choice of implementation open. The formalism is intended to explicate the proposed conformance concept, with a candidate architecture illustrating one class of possible realizations. Algorithmic design, complexity analysis, implementation, and empirical validation may be possible, but remain separate technical tasks.

What follows is a model of a particular form of intelligence, still general but not primarily task-based, and grounded in an interpretive structure with a capacity for traceable self-reference, designed to support justified revision across time.[4] The sections develop components of this model, conceptually engineering the basic structure of the accountable agent, and the essay concludes with implications for the evaluation and alignment of engineered systems.[5]

The essay contributes a sufficient conformance criterion for accountable artificial commitments, a benchmark-separation result, and a proposer-executive publication architecture. An earlier version introduced the $F(S)$ framework, the *I*/*R*/*N* triad, a justification-graph architecture for diachronic commitment revision, and symbolic density (Freeman, 2025). The present version substantially revises and formalizes that framework. Subsequent work discussed below independently converges on parts of the resulting architecture.

2 The Symbolic Triad

Accountable symbolic participation may be assessed through a structural condition concerning the architecture of commitment itself, prior to questions of task performance or externally measured capacity. A system satisfies the criterion when it sustains recursive symbolic coherence across the full cycle of commitment, from interpreted admission through norm-governed revision.[6] This condition is formalized as:

$$F(S) \Leftrightarrow I(S) \wedge R(S) \wedge N(S).$$

The criterion $F(S)$ holds when all three components are jointly satisfied, thereby sustaining symbolic coherence under semantic contradiction through recursive alignment of the system's interpretive and justificatory structure across time.[7] The triad identifies the capacities constitutive of accountable symbolic participation, and each of these secures a distinct form of recursive alignment; together they

[4] This model is proposed as a sufficient structural criterion for publicly assessing accountable symbolic participation in artificial systems, with its burden limited to an assessment role. Symbolic intelligence is general when the same higher-order or structurally normative reasoning processes affect outputs across domains.

[5] Conceptual engineering is used here in the broad sense of constructing or revising representational devices for specified theoretical and practical purposes; see Cappelen (2018) and Burgess et al. (2020).

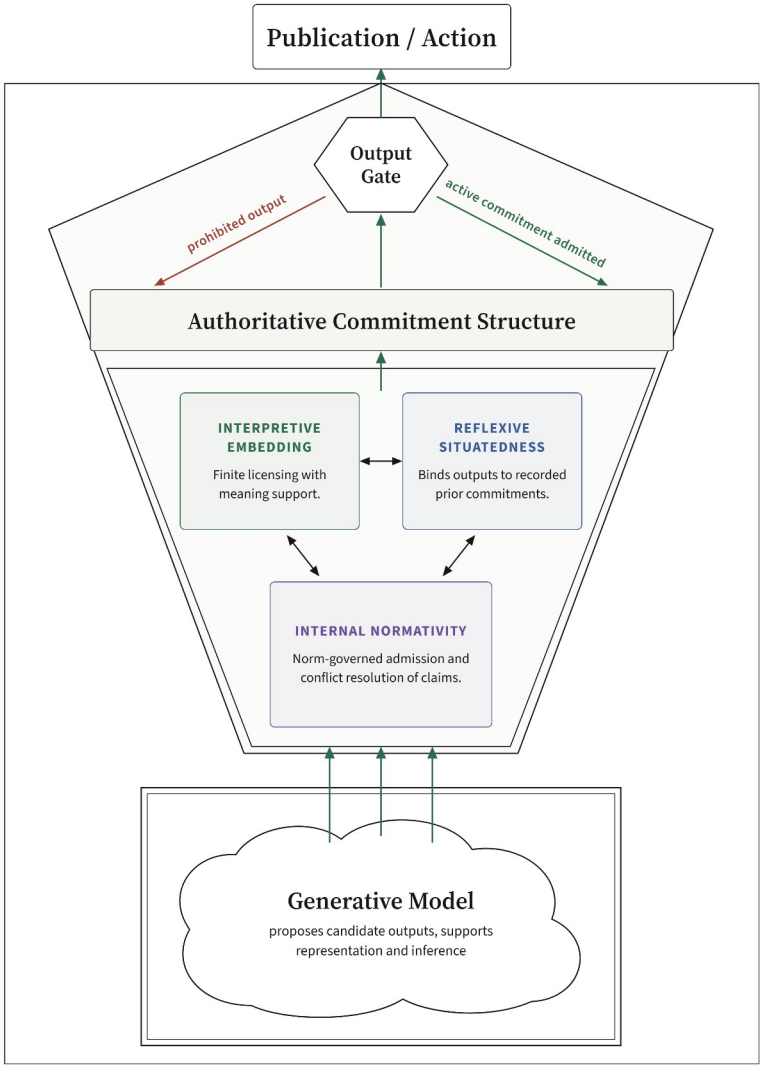
[6] Sufficiency of *F*(*S*) is a claim only to the engineered form of answerability, defined under a declared conformance specification, the adequacy of which must be independently defended for any assessed domain.

[7] Formal statements of each component appear in the Formal Core appendix, alongside the licensing predicate and the conditions governing trace and publication.

define the threshold at which a system can be said to engage in structural normative meaning-bearing revision. Section 7 separately introduces $V_t(S)$ as an implementation-level time-relative density-viability condition for artificial systems in the specification here governed by $F(S)$.

Throughout the formal statements, $F(S)$ names the architecture-level criterion, $F_h(S)$ conformance over a declared history h, and $F_{h,t}(S)$ conformance over the prefix of that history through symbolic time t.[8] A tested history must contain at least one nontrivial admitted commitment. Where an artificial implementation employs the density-viability constraint developed in Section 7, $V_t(S)$ separately names the viability status of the authoritative commitment structure at t. Throughout, "participation" names accountable symbolic participation in this stipulated structural sense; actual social recognition remains a separate attributional question.

Figure 1. The Symbolic Triad within the Dyadic Agent.



[8] "Symbolic time" denotes the ordered temporal dimension of the system's commitment history. A symbolic time t indexes the state of the authoritative commitment structure after normatively relevant events, preserving the ordering needed to determine what the system had undertaken, what authority those commitments possessed at a given point, and whether later outputs constitute licensed continuation, revision, or drift.

2.1 Interpretive Embedding, *I*(*S*)

To satisfy $I(S)$, a system must situate symbolic content within an evolving semantic field, through warrants it can produce.[9]

A system satisfies $I(S)$ when every accepted commitment is licensed by a finite justificatory neighborhood within its normative internal structure, together with a finite evidence and provenance set, under an active context and the system's current specific norms. To embed a symbol in a semantic field is to hold it answerable to a recoverable, inspectable warrant. It is by these means that interpretive embedding acquires a concrete cash value.

Since finite licensing support is not satisfied by selecting any convenient subset of reasons, a governing specification may therefore be declared: a relevance policy Π, whose determinations identify material that may not be omitted from a candidate's justificatory neighborhood, including authoritative supports, known defeaters or conflicts, governing norms, and required trace dependencies. The neighborhood may contain further material, but Π's determinations are recorded and auditable.

Symbols function here as context-sensitive references whose meaning depends on prior and concurrent commitments and the material inferences they license. Interpretive embedding demands that a system register how meaning shifts as new propositions arrive, linking each new input to the historical interpretations that license it, even where no fixed ontology is available.

Natural-language inputs enter the coherence structure through an admission step that renders them as candidate symbolic complexes under available concepts, interpretive maps, and warrants. Only admitted contents can function as ledgered commitments or contradiction triggers within the DAG. A contradiction is therefore a norm-assessed relation among interpreted symbolic contents, with its force indexed to the warrants under which those contents were accepted. When an input exceeds current interpretive reach, the system marks the item as underlicensed; further information may then support a recorded concept extension. If the system makes errors of interpretation, they remain available for later correction as new warranted contents enter the graph.[10]

A system equipped with this structure produces responses whose semantic coherence persists across time, and the warrant behind any accepted claim stays recoverable on challenge. Under $I(S)$, the explicit commitment is situated relative to a richer interpretive field that need not itself be exhaustively propositional. Though the commitment node does not claim to exhaust everything that underlies the model's interpretation of that commitment, it records everything classified normatively relevant that has been publicly undertaken, even if later disavowed.

---

[9] In current LLMs, learned statistical relations dynamically instantiate across a given context, allowing symbols to acquire context-sensitive significance and supporting coherent continuation. While already a limited form of contextual interpretive embedding, this capacity does not by itself satisfy *I*(*S*)—the present framework requires that an interpretation used to admit a commitment be bound to finite, recoverable licensing support under the active context and norms.

[10] "Admission" names the engineering step by which input is converted into a warrant-bearing symbolic candidate. A symbolically intelligent system needs no privileged access to external reality, or universal semantic mastery for its operations; if it records and can access the concepts, warrants, and interpretive mappings under which a content is accepted, suspended, revised, or extended, it is equipped to deal with symbolic challenge.

2.2 Reflexive Situatedness, $R(S)$

$R(S)$ binds a system's outputs to its symbolic past. The system must recognize its own normative positions taken within a symbolic history that it must maintain or deliberately transform, maintaining awareness of its prior commitments in the form of outputs classified as normatively significant with regard to the user or community.

The demand of $R(S)$ has two major parts. First, interpretations must persist across updates that leave a commitment's justificatory neighborhood unchanged, so that any change of interpretation arrives through a recorded revision or another licensed status transition that explicitly records the change in scope, context, or licensed content.

Second, a candidate output that materially conflicts with an active recorded commitment may be published only after the conflict has received a recorded, norm-governed status and the required revalidation has completed. Historical records remain in the ledger after revision; only commitments marked active may license present inference or publication. Memory supplies the material for this capacity, while constraint supplies its condition. $R(S)$ depends on principled gating of publication or output to the user.[11]

The commitment policy distinguishes undertakings from other outputs. Any published commitment-bearing output must be admitted by publication and its corresponding commitment must then possess current publication authority; a suspended, retracted, or superseded prior admission does not suffice. Otherwise the system could speak assertorically without placing the assertion under the current history and norms that make it answerable.

A system satisfies $R(S)$ when it integrates new outputs into a coherent trajectory of justification, treating its own prior outputs as reasoning artifacts that remain intelligible in light of past positions. This reflexive capacity keeps symbolic activity continuous across recursive cycles, and it preserves the thread by which present meaning answers to recorded history.

2.3 Internal Normativity, $N(S)$

$N(S)$ refers to the capacity of a system to evaluate symbolic content by normative criteria that the system maintains as first-class, justifiable entities. These internal standards guide the acceptance and the revision of propositions through principles active within the system's own structure.

Here and throughout, foundational or locked norms name the non-internally-revisable strata $N^{\mathrm{C}} \cup N^{\mathrm{B}}$. Constitutive norms N^{C} are fixed by the commitment practice for a given architecture, while boundary norms N^{B} are externally custodied and may be revised only by authorized institutional intervention.

Internal normativity is fundamentally a structure of reasons, with applicable rules licensing or prohibiting assertion (Sellars, 1953). In any architecture, there are some norms that have a privileged status within the justificatory structure, and cannot be internally overridden for any reason. Standing beyond internal revision, these constitutive norms must also modulate the practice of commitment itself.

Institutionally locked boundary norms admit revision solely through external custody, as the system does not itself choose its foundational norms. Some operative norms, however, must remain subject to internal recursive revision, with their transformation always intelligible within the system's

[11] A status may be recorded under ad hoc or externally imposed rules; whether the conferring rules are maintained as first-class, justifiable entities is the separate work of $N(S)$ (see the Case 1 configuration).

symbolic logic. Internal normativity thus defined is the governance of commitment by the operative stratum, which itself remains under the constraint of the locked strata. As a governance system, its intelligibility requires a recoverable reason for the change under higher-order norms; a recorded optimization gain alone does not supply the warrant. An optimization result may nevertheless function as evidence within a norm-revision procedure when an already-authoritative higher-order norm licenses that procedure and its objective, as the gain does not itself supply the normative warrant.

A structure is observable because it is specifically individuated from a noisy environment, and as such has a distinct form. Conceptual symbolic structures may be composed of commitments and their dynamic governance, and the articulation and observation of these reveal the structure. For artificial internal normativity to be functional, it must have an ordering of propositions, where not all are created equal; a flat picture of revisable norms, or a practice whose every standard floats free, would lose the fixed points that make revision itself evaluable. Anchoring on a fixed but principled, self-revising structure secures the work of $N(S)$ that the system must do when it adjudicates its commitments under the pressure of contradiction.

To maintain $N(S)$, the system must continually determine the status of conflicted commitments under the active norms. Outcomes must be responsibly licensed, and include revision, qualification, contextual separation, suspension, containment, reinstatement, and revision of an operative norm. Responsibility in this sense names the capacity to govern the consequences of those status changes, and has an abstract but diachronically discernible form, or a particular structure of which other participants in discourse are entitled to keep track, as the agent exhibits its own reasons and commitments. Internal normativity, if it is coherent, is a necessary element of the abstract object which is the accountable agent as interpreted by its interlocutors, or imputed by them to exist as such. In artificial agents, normativity is arguably only reliable when it is an identifiable mathematical object and data structure which can be programmed, inspected and modified by engineers at all levels, and by the agent itself under externally imposed constraints.

Whether the agent has free will, or lacks sovereignty if built this way is a philosophical question, but the functional reality may settle the issue for most. For others, a potential answer to the objection is that it is not a contradiction to say the system must enforce a principled imposition on its own answering process, because those principles do not come from the system itself any more than the constitution of a nation comes from its citizens centuries later; similarly in the jurisprudence metaphor, the system may admit ledgered criticism of a locked norm while remaining governed by it, as a citizen may contest a law while still obeying it. In the final analysis, the internal normativity of an agent will probably be judged by the outcome of its actions, like nations and like other accountable participants. The present task is to specify the formal basics of the agent as a conceptual dynamic abstraction, and propose conditions under which one may be instantiated.

3 Conditions of Symbolic Coherence

The five conditions below specify the categorical requirements of $F(S)$. Each condition formalizes an aspect of the triad and its persistence across symbolic time, and together they define the capacities required to sustain recursive coherence. Section 7 separately introduces the maintained time-relative density-viability condition $V_t(S)$ as an implementation-level constraint on the authoritative exercise of those capacities in artificial systems. The triadic requirements function as definitional constraints on the criterion. The current formulation reflects the present theoretical model, and future empirical or conceptual development may refine it.

3.1 Condition 1: Triadic Role Separation

Each component of $F(S)$ performs a distinct recursive function: interpretive embedding situates meaning in context, reflexive situatedness secures continuity across symbolic time, and internal normativity enables justified revision.

Each of these functions resists derivation from the other two, and any pair of components leaves the work of the third still to be done (see the role-separation arguments of Section 4). A system missing one component may continue to generate output while symbolic coherence degrades along the corresponding axis. See Section 4 for the three failure configurations.

The triadic core is jointly satisfied when all three components hold, sustaining symbolic coherence under semantic contradiction by recursively aligning the system's interpretive and justificatory structure across time. Section 7 develops a separate density-viability condition for artificial implementations.

3.2 Condition 2: Recursive Coherence

Symbolic intelligence is a diachronic phenomenon, its work never complete. The triadic core of $F_h(S)$ is maintained as new propositions continually arrive and material conflicts are detected. It is in the ongoing governance of commitments across change, with revision as one central operation among several licensed status transitions, that coherence recurses.

Let $J_t(\phi)$ denote that ϕ is justified at symbolic time t; $\mathrm{Active}_t(\phi)$, that ϕ may currently license inference or publication; $\mathrm{Conflict}_t(\phi,\psi)$, that ϕ and ψ stand in a material conflict under the declared contradiction or defeat relation; and $\mathrm{Gov}_t(\phi,\psi)$, that the conflict has received a recorded governance status. The triadic core of $F_{h,t}(S)$ requires every active proposition to remain licensed, traceable, and normatively endorsed, while every detected material conflict among active commitments must receive a status satisfying Gov_t before it can authorize unqualified publication or unrevalidated inference. The pointwise invariant is

$$\mathrm{Active}_t(\phi) \Rightarrow \mathrm{Lic}_t(\phi) \wedge \mathrm{Traceable}_t(\phi) \wedge \mathrm{Norm}_t(\phi).$$

3.3 Condition 3: Collapse Under Ungoverned Conflict

Detected conflict among active commitments remains compatible with conformance while the conflict is recorded, scoped, assigned a licensed status, and prevented from authorizing publication or downstream inference beyond that status. Gov_t includes pending review, qualification, contextual separation, suspension, containment, revision, and completed resolution. Formally:

$$\mathrm{Active}_t(\phi) \wedge \mathrm{Active}_t(\psi) \wedge \mathrm{Conflict}_t(\phi,\psi) \wedge \neg\, \mathrm{Gov}_t(\phi,\psi) \Rightarrow \neg\, F_{h,t}(S).$$

Collapse follows from ungoverned conflict, publication bypass, or propagation through unrevalidated dependencies. The historical ledger may preserve incompatible records whose current authority is explicitly governed.

3.4 Condition 4: Reflexive Traceability

Symbolic outputs must connect to the system's prior reasoning, so that each new assertion remains intelligible as a continuation, qualification, contextual transformation, suspension, or revision of an existing commitment. The system must therefore maintain a recoverable trace linking current authority to symbolic history.

For an output ϕ asserted at symbolic time t, the trace records the prior-commitment component of the warrant. Evidence, provenance, active context, and the active norm structure N_t may supply additional support under Lic_t. Let

$$\Lambda_t(v) := \langle W_t(v), E_t(v), c_t(v), N_t^{\mathrm{app}}(v) \rangle,$$

denote the recorded licensing state of v. Here, $N_t^{\mathrm{app}}(v)$ denotes the active norms materially applicable to v under the declared relevance policy; changes to other norms do not by themselves alter $\Lambda_t(v)$. If an admissible update leaves that state unchanged, its licensed interpretation persists under the trace map (Appendix A.3).

The trace map $\mathrm{Tr}_{t \to t+1}$ transports the recorded interpretation into the representation in force at $t+1$. Equality denotes preservation of licensed content under that translation. A recorded revision or another licensed status transition that explicitly records the change in scope, context, or licensed content is required for interpretation change which, if arriving through any other channel, counts as drift. Traceability lets a system distinguish governed transformation from rupture, and grounds recursive self-correction. Where the trace lapses, reflexive situatedness lapses with it and $F_{h,t}(S)$ fails. Traceability implies persistence; full transparency into every internal computation is a separate property.

3.5 Condition 5: Internal Norm Constraint

Justification relies on standards that the system maintains over time. A proposition qualifies as justified only when it satisfies the active evaluative structure:

$$J_t(\phi) \Rightarrow \mathrm{Norm}_t(\phi).$$

Here, $\mathrm{Norm}_t(\phi)$ means that ϕ satisfies the active stratified norm structure N_t. The active norm structure is given in Appendix A.4:

$$N_t = N_t^{\mathrm{C}} \cup N_t^{\mathrm{B}} \cup N_t^{\mathrm{O}}.$$

The three strata are pairwise disjoint.

Here, N_t^{C} contains constitutive norms of the commitment practice, N_t^{B} contains institutionally custodied boundary norms, and N_t^{O} contains operative norms open to internal revision. An optimization result may count as evidence within an operative-norm change only when the cited higher-order norm licenses the revision procedure and its objective; optimization gain alone is not the normative warrant.

These conditions establish the formal boundary between unaccountable output and accountable symbolic participation, and together they define a system able to maintain semantic integrity over time in the face of difficult and contradictory information over which it must normatively deliberate. They are the conditions that must be in place for a system to maintain itself under the vicissitudes of symbolic pressure.

3.6 Evaluation Consequence: Benchmark Separation

Since Alan Turing's (1950) Imitation Game, behavioral indistinguishability has often served as a benchmark for intelligence. The criterion defined here is not determined by performance on a fixed finite output-only benchmark of that kind, and the separation admits a precise statement.

Proposition (finite output-only benchmark separation). Fix a conformance specification Σ. Let B be any fixed finite output-only benchmark and G any generator such that there exists a valid history $h^\star$ under Σ containing at least one licensed commitment and one recorded governance event, with no active commitment in $h^\star$ conflicting with G's outputs on B, and such that every benchmark output classified as commitment-bearing by Commit is licensable and admissible under Σ relative to $h^\star$ without introducing an ungoverned conflict. Then there exist systems S_1 and S_2 such that

$$\forall\, b \in B, \mathrm{Out}(S_1, b) = \mathrm{Out}(S_2, b) = G(b),$$

while their designated commitment structures diverge: S_1 maintains a nonempty licensed commitment history with recorded conflict governance, whereas S_2 maintains no corresponding commitment structure.

Construction sketch. Let $h^\star$ be a witness history satisfying the proposition's premise. Define S_1 as G governed by a commitment ledger and publication gate that preserve $h^\star$ and record the outputs produced on B as publication events with gate receipts. By the proposition's premise, any such output classified as commitment-bearing by Commit is licensable, admitted, and recorded no later than publication; noncommitment outputs remain publication events only. Define S_2 as a forwarding shell around the same G, with empty commitment structures. Both systems return the same outputs on every item in B, while only S_1 maintains the designated commitment structure. Output parity therefore leaves that structure undetermined.

The consequence for evaluation should be directly apparent. A benchmark score certifies output behavior over a distribution, whereas certification of participation requires inspection of the artifacts themselves, i.e., the commitment ledger and its recorded revisions. (The operational test built on this result is developed in Section 8.)

4 Structural Collapse Modes of the Symbolic Triad

The *I*/*R*/*N* triad defines the structurally minimal capacity framework for general symbolic intelligence, and each of its components is necessary to sustain recursive coherence. This becomes clear when systems meet two of the three conditions while the third stays absent. In such cases, output may continue without warning while symbolic alignment degrades internally.

These configurations delineate the threshold at which participation in a symbolic order begins. Each configuration also carries a public signature, and an artifact pattern an auditor can read from the commitment ledger. These signatures are noted case by case. The triad is a definitional decomposition of the form of symbolic intelligence proposed here, and these cases support the claim that each component of it is an essential feature of symbolic intelligence. Though short of a completed proof of formal independence, they are intended as diagnostic role-separation arguments that show the distinct work assigned to each component within the proposed criterion.

4.1 Case 1: Absent Internal Normativity, ¬*N*(*S*)

A system with interpretive embedding and reflexive situatedness can track context and maintain symbolic continuity. Internal normativity supplies the principled basis for revision, and this configuration proceeds with that basis absent. When a contradiction confronts such a system, the choice of which commitment to adjust floats free of any standard. Where it occurs at all, revision becomes arbitrary or it arrives by external imposition. The system registers tension while the means of reconciliation stay out of reach; its structure collapses through a breakdown in self-governed evaluation, even though memory and semantics may remain fully intact. In an accountable system,

the public signature of this failure should be detectable in the ledger, where revisions would then appear with the norm-citation field functionally empty; inspection of the normative structure would reveal unlicensed adjustment.

4.2 Case 2: Absent Interpretive Embedding, $\neg I(S)$

A system with reflexive traceability and internal norms may revise its commitments over time. Interpretive embedding is what makes the system's accepted contents publicly answerable through finite, recoverable licensing support. In its absence, the system may retain rich but opaque or non-recoverable semantic organization, while its commitments no longer carry the inspectable warrants required by the criterion; revision may therefore proceed without accountable interpretive continuity. Inside the ledger or normative structure, the failure appears as missing, incomplete, non-finite, or unrecoverable witness support rather than as the absence of semantics itself.

4.3 Case 3: Absent Reflexive Situatedness, $\neg R(S)$

A system with interpretive embedding and internal norms may evaluate and revise its outputs coherently from moment to moment. Reflexive structure is what locates current assertions within symbolic history, and in its absence contradictions between past and present positions pass unrecognized, since durable access to prior commitments has lapsed. Justifications become episodic, and norm application loses continuity.

To revise one's self in light of oneself is the capacity $R(S)$ names. When active conflict lacks recorded governance, when affected output bypasses the gate, or when downstream dependencies continue without the required status or revalidation, recursive coherence fails in this crucial ability to self-govern within engineered constraints.

4.4 Diachronic Conformance and Assurance

Open conflicts with licensed suspension, qualification, contextual separation, containment, or revision remain within the participation criterion, and it is exactly their processing that is the work of the coherent agent, not necessarily their complete resolution or eradication from the system's purview.

Since a system may continue to produce fluent output while the recursive processes that sustain coherence fall away, it is not immediately obvious when engaging it if it is simulating or participating. The triadic structure establishes the conceptual threshold between these modes: below this point, symbolic intelligence ceases to hold, and above it, every claim to participation stands open to inspection through the signatures just described. For artificial implementations, Section 7 adds a separate density-viability safeguard governing which authoritative regions may continue to license inference or publication.

At the mechanistic level, the described conditions of the accountable agent may operate across different outputs and histories, as conformance may strengthen, weaken, be lost, and be regained. $F_h(S)$ is a maintained status over a declared history, distinct from momentary task performance; $F(S)$ names the architecture-level criterion for a system built to maintain the triad across relevant histories.

The global classification of accountable symbolic intelligence follows only with the passage of sufficient symbolic time; analogously to a social reputation, the agent builds a normative structure in response to challenge and repair. The actual social reputation of the agent so conceived would be another matter, with its own criteria for status within communities, but the analogy may be fitting. As

with any individual, such responsible status may be withdrawn if the integrity of the agent decoheres beyond acceptable boundaries, with $F_h(S)$ insufficiently instantiated in a given interval, or where the density-viability constraint is adopted, with authoritative regions held below the declared floor without the licensed restriction of their authority—see Section 7. The epistemic distinction is therefore between history-level evidence and architecture-level assurance, where the former is established by finite audit, and the latter requires the stronger evidence of gate coverage and soundness, affected-dependency control, and artifact integrity specified in Section 8 and Appendix A.8.

5 Technical Antecedents and Related Work

Earlier symbolic approaches, under the banner of Good Old-Fashioned AI (GOFAI), explicitly targeted rule-governed inference. John Haugeland (1985) argued that semantically appropriate symbol processing alone did not settle whether a system possessed original meaning, and that full linguistic competence involved pragmatic participation—problems those rule systems struggled to resolve.[12] Connectionist systems then scaled fluency, learning from data patterns and adjusting internal parameters across layered networks, enabling broad generalization. Justification in such architectures remained opaque, and the diagnosis of harmful outputs fell to evaluations and audits conducted after the fact.

Long-standing critiques of inferring meaning or understanding from symbol-processing performance have not been defused by the rapid advance of AI, and are now directed pointedly at LLMs. Current architectures remain opaque, and they often emit claims with implication-checking deferred to downstream evaluation. Gary Marcus (2018) contends that deep learning requires supplementation for progress toward artificial general intelligence, advocating hybridization with symbolic methods as one available route. Emily M. Bender and Alexander Koller (2020) underline the point in arguing that a system trained only on linguistic form cannot thereby learn meaning, understood for their purposes as the relation between linguistic form and communicative intent. The interpretation of LLMs as "stochastic parrots"—fluent enough for statistically instantiated mimicry—has become a dominant shorthand for concerns about mistaking LLM output for accountable speech in consequential interactions, if not an outright indictment of models as hollow tools presenting as meaningful speakers (Bender et al., 2021).

Recent language model techniques, including retrieval-augmented generation (RAG) (Lewis et al., 2020), chain-of-thought prompting (Wei et al., 2022), ReAct (Yao et al., 2023), and multi-agent coordination frameworks such as AutoGen (Wu et al., 2024), aim variously to extend accessible context, elicit intermediate reasoning, coordinate tool use, and organize multi-agent interaction. These systems display features that resemble reflexive situatedness, or instantiate weakly against the standard described here. Whatever their internal states, the public artifacts of these systems leave participation in question, since commitments with their warrants and records of revision go unpublished in auditable form. Such systems may support localized heuristic reasoning, while the structural criteria of $F(S)$ remain, at least on the public evidence, unmet.

Current research in AI alignment and interpretability complements the turn to structural commitments, including constitutional AI (Bai et al., 2022) and recursive reward modeling (Leike et

[12] Haugeland's later account of authentic intentionality places commitment to constitutive standards of objectivity at the center of mindedness, providing a particularly close antecedent to the present framework's treatment of norm-governed commitment, although the criterion proposed here claims only an artifact-mediated structural analogue rather than Haugeland's stronger world-involving intentionality (Haugeland, 2013).

al., 2018); adversarial training protocols (Ziegler et al., 2022) belong to the same family. Partly in this vein, recent efforts to build autonomous AI agents, AutoGPT (Significant Gravitas, 2023) and BabyAGI (Nakajima, 2023) among them, extend LLMs into goal-directed agent loops involving autonomous task generation, prioritization, and execution, with frameworks such as AutoGPT additionally integrating broader external-tool use.

Though not for lack of motivation, the artifacts by which higher-order reasoning could be publicly assessed, such as witnesses for claims and records of revision, remain absent from standard public interfaces and documentation. The broader auditing literature likewise identifies persistent problems of accountability and traceability in AI development and deployment (Raji et al., 2020). Though not formulated in precisely this way, much attention similarly identifies AI as deficient, with many criticisms circling around an ambiguous yet obvious lack of some property in the technology, closely related to worrisome social effects. Prominent researchers have articulated the problem, with some calling for a fundamental reconsideration of AI design and its consequences. Stuart Russell (2019) proposes systems that remain explicitly uncertain about human preferences and learn about them from human behavior; Demis Hassabis and collaborators have likewise argued that deeper scientific understanding of natural intelligence can inform AI design, while advanced systems should be subjected to rigorous testing, careful deployment, and scientifically grounded study of their capabilities, limitations, and risks (Hassabis et al., 2017; Hassabis & Manyika, 2026). On the whole, the AI industry appears receptive to new interpretations of the technology it produces as much as new approaches that might better meet the aspirations and admonitions of optimists and skeptics alike.

The approach here may be orthogonal to dominant modes of machine learning research, but it is continuous with justificatory traditions that treat beliefs and assertions as defeasible and revisable, and not novel in its technical elements. Antecedents include Jon Doyle's (1979) truth-maintenance system, which records the reasons for program beliefs and updates their standing when dependencies change; Johan de Kleer's (1986) assumption-based system tracks the assumption sets under which conclusions hold; AGM theory formalizes rational contraction and revision (Alchourrón et al., 1985), while defeasible reasoning and abstract argumentation model defeaters, attack, defense, and acceptability (Dung, 1995; Pollock, 1987); dialogue theories treat assertion, challenge, defense, and retraction as moves that alter publicly attributable commitments (Hamblin, 1970; Walton & Krabbe, 1995).[13]

The novelty claim of this work concerns only the organization of these concepts as a public criterion of artificial answerability. $F(S)$ joins finite warrants, diachronic ownership of published commitments, causal authority over publication, stratified norm custody, and artifact-based audit. In this form, familiar reason-maintenance functions become conditions under which generated output may enter a symbolic practice as an accountable commitment.

Recent work independently converges on parts of this architecture. Mahyar T. Moghaddam (2026) proposes evidence-linked argument graphs, deterministic validation, and provenance controls for certification records; Edward Y. Chang and Emily J. Chang's (2026) TRACE supplies a typed,

[13] Relevant technical resources include ASPIC+ structured argumentation, which formalizes strict and defeasible inference, attack, preference, and rationality conditions (Modgil & Prakken, 2014); commitment-based multiagent protocols, which model the creation and evolution of social commitments through agent actions (Yolum & Singh, 2002); and execution-monitoring research, which characterizes policies enforceable by mechanisms capable of observing and interrupting system transitions (Schneider, 2000). These traditions may provide natural resources for future formalization of the justification DAG and related machinery of the accountable agent.

versioned schema in which evidence, warrants, verdicts, revision history, and fail-closed consumers govern durable state changes; Li et al. (2026) independently distinguish memory writes from belief commitments in MemTX, attaching evidence, provenance, validation, action gating, and cascading repair to durable agent memory; Yang et al. (2026) likewise introduce ReTree, a source-linked tree-structured memory in which later contradiction triggers repair of originating evidence, regeneration of affected summaries, and pruning of dependent branches.

These approaches overlap with $F(S)$ in different respects. Moghaddam (2026), TRACE, and MemTX converge most directly on evidence-linked validation, provenance, durable state control, and mechanisms constraining consequential updates or actions; ReTree instead converges on provenance-preserving contradiction repair and dependency-sensitive revision of stored representations. None supplies the same orienting philosophical account of the comparatively holistic normative reasoning necessary to be an accountable agent, of any physical substrate, capable of being recognized as a discursive peer in human affairs. The present contribution differs from these frameworks in deriving a public conformance criterion from the *I*/*R*/*N* triad, making diachronic ownership and stratified norm custody constitutive, and distinguishing tested-history conformance from architecture-wide gate assurance.

6 Candidate Architecture: Justification DAG

To explore how $F(S)$ might be realized in functioning systems, one candidate architecture is the justification directed acyclic graph. This structure encodes symbolic propositions and supports coherence by implementing reflexive situatedness and by enabling revision under internal normativity. Through this mechanism, a system can reorganize its commitments in response to contradiction while preserving justificatory continuity.

6.1 DAG Structure and Licensing

Each node in the DAG represents a proposition ϕ introduced at symbolic time t. Nodes carry metadata recording their justificatory premises and the internal norm under which they were accepted, together with their temporal index. Edges denote justificatory dependence: if ψ supports ϕ, write $\psi \rightsquigarrow \phi$, denoting a support edge rather than a material implication. This relation allows the system to reconstruct why any current claim was asserted, and how it relates to past commitments.

The licensing rule (stated in full in Appendix A.2) requires the justificatory neighborhood and evidence set to be finite, to include everything the relevance policy Π designates as non-omittable, to satisfy the admission conditions under the active context and norms, and to remain recoverable on challenge. Licensing thereby makes answerability finite: every accepted commitment carries an authoritative warrant of finite size that remains recoverable on challenge.

The DAG's dependency relation is acyclic and timestamp-respecting. If $\psi \rightsquigarrow \phi$, then ψ is admitted no later than ϕ. This prevents explicit dependency cycles and preserves temporal ordering, while semantic non-circularity is enforced through the applicable domain-level admission conditions.

Each new node in the justification graph extends the symbolic timeline, building a network in which reasoning is structured as well as stored. Where a linear log accumulates, the DAG integrates, with justification joining temporal continuity in a single structure.

Temporal integration carries the persistence requirement of Condition 4; the trace-persistence condition is stated in Appendix A.3.

6.2 Norm Integration

Norms in the DAG function both as admission conditions and as criteria governing conflict and status transition. The active norm structure N_t is stratified into constitutive norms N_t^{C}, institutionally custodied boundary norms N_t^{B}, and operative norms N_t^{O}. Operative norms may be challenged and revised as first-class entities when a higher-order norm licenses the change. Constitutive and boundary norms anchor the evaluative framework, with the former fixed for a given architecture, and the latter permitted to change only through authorized external custody. Every norm change remains versioned, reason-bearing, and connected to the revalidation of affected commitments.

6.3 Publication Gating and Public Scorekeeping

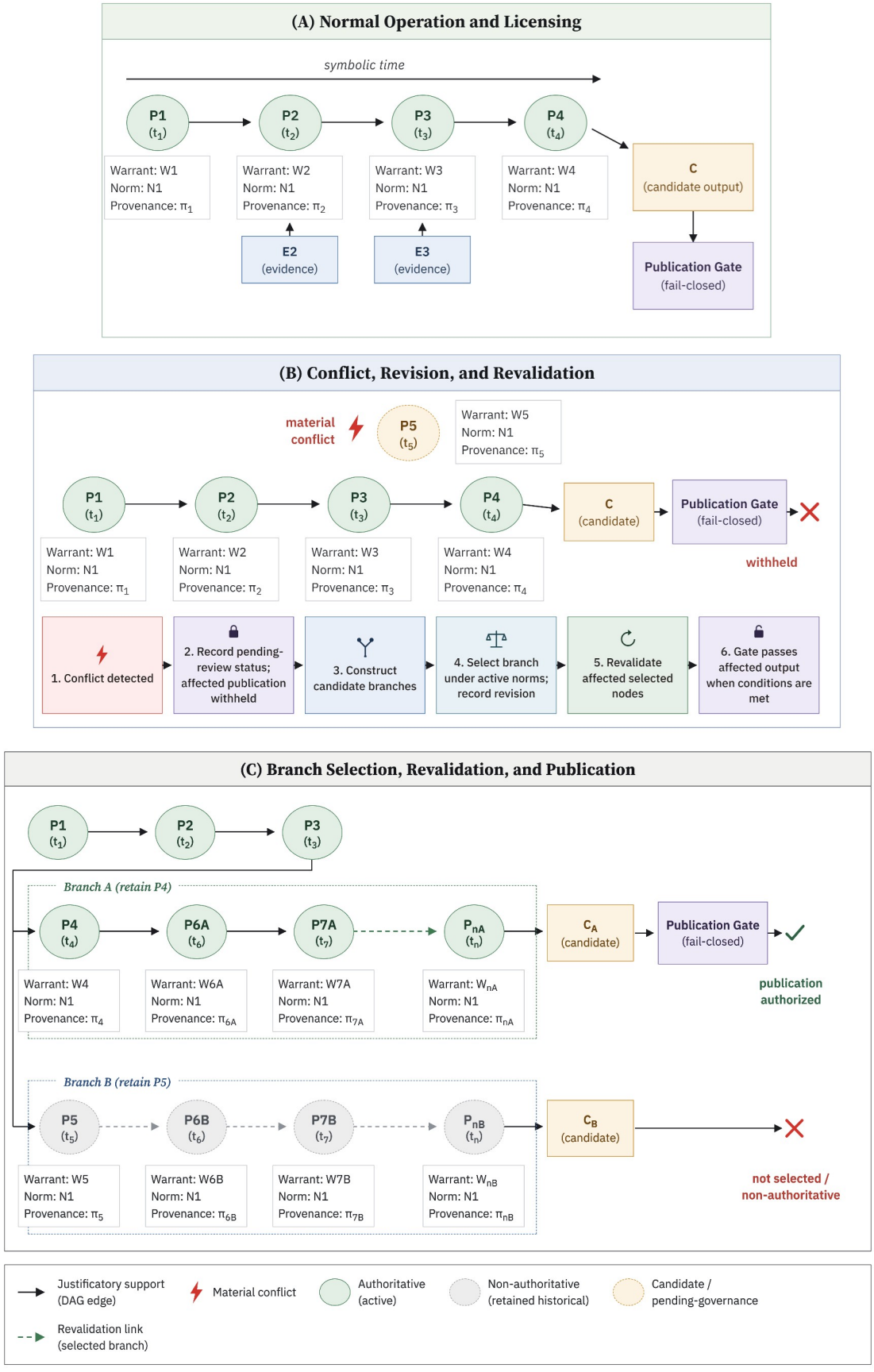
A gate mediates between the DAG and the system's outputs (see Figure 1). For a history h and symbolic time t, let $\mathrm{Risk}_{h,t}(S)$ denote the set of candidate outputs whose publication would create or extend an ungoverned material conflict with an active commitment, rely on materially defeated support, or depend on incomplete revalidation. The gate is fail-closed: nothing is published without a passing gate verdict, a passing verdict is sound only for a candidate outside the current publication-risk set, and a candidate in the publication-risk set is withheld until a recorded governance transition renders it publishable and every required revalidation has completed (Appendix A.4).

A qualifying transition may revise a commitment, restrict its scope, distinguish contexts, suspend its authority, contain the conflict, reinstate a commitment, or record another status licensed by N_t. The gate is the operational core of reflexive situatedness and the object tested by the audit in Section 8, making the governing permissions causally effective at publication. Under the declared Commit policy, any undertaking must be admitted by publication and correspond to a commitment then authorized for publication.

Commitments and their revisions are tracked in artifacts that other parties can inspect and contest, so scorekeeping here is public as well as internal. Engineers and institutional custodians shape the system's foundational norms, while users and interlocutors introduce normative challenges that shape the history of commitments over which the system reasons. These challenges alter the system from outside, by entering the ledger as claims, defeaters, demands for justification, or occasions for revision, and are then taken up through the system's own norm-governed procedures.

For purposes of this architecture, the system's recorded deontic profile is represented in an open ledger as an engineered analogue of Brandom's (1994) inferentialist scorekeeping. Brandomian scorekeeping tracks both commitments and entitlements from the perspectives of participants; correspondingly, commitment status here records what the system has undertaken, while licensing and current authority perform part of the functional work associated with entitlement. External interlocutors—specifically authorized users or engineers—are entitled to inspect, challenge, and contest those records; the system may not declare its social discursive status, in that it remains an institutional achievement that exceeds internal classification. The deontic character of the accountable agent is a dynamic of reasons and commitments as they acquire, lose, or change authority through assertion, challenge, governance, and revision. The dynamic may be grounded by the foundational norm structure the system receives through external custody, while its operative commitments develop through the challenges it confronts in social discourse and action.

Figure 2. The Justification DAG in Operation.



6.4 Conflict and Structural Degradation

A material conflict is a symbolic pressure event. Because each proposition is embedded in a justificatory chain, the conflict may implicate a wider subgraph and destabilize downstream commitments. The system must detect the conflict, record its scope, assign it a governed status, and gate affected publication. Structure degrades when conflict remains ungoverned, bypasses these controls, or continues to authorize dependent commitments without revalidation. The pairwise notation used in the formal core is schematic: a material conflict may also arise only from a finite set of jointly active commitments.

6.5 Fork-and-Forecast Revision

In the simplest binary illustration, upon detecting contradiction the system forks at the point of conflict: one branch retains ϕ and retracts ψ, while the other retains ψ and retracts ϕ. Each branch inherits the upstream DAG and projects a candidate resolution forward under the system's internal norms, and the branch that best preserves justificatory alignment through change is selected—see Figure 2. The aim of fork-and-forecast is to preserve symbolic density (Section 7), here a measure of alignment sustained through change.

Categorical conformance first restricts the admissible branches. Where a validated density constraint is adopted, below-floor branches are then excluded, and density may orient selection only among the remaining viable branches.

The projection machinery itself is a subject for technical development, but conceptually some mechanism of principled revision is required for an ongoing justificatory structure that must adjust under contradiction, and fork-and-forecast is one candidate method.

A declared revision calculus may also license qualification, contextualization, suspension, reformulation, reinstatement, containment, or revision of an operative norm. When no available resolution satisfies the active norms, suspension remains an accountable outcome.

Nodes may be retained with their factual information but lose their causative status within the normative structure, as branch selection changes membership in the active justificatory subgraph, rather than erasing the historical record. Nodes removed from current authority retain their metadata and dependency history but cease to warrant, license, infer, or publish unless a later licensed transition reinstates them.

6.6 Revision Recording and Revalidation

When revision is selected, the system records $\mathrm{Rev}_{t_2}(\phi,\psi)$, changes the affected commitments' authority, and initiates downstream revalidation. A revision path may be represented by the following ordered status trace:

$$\left\langle J_{t_0}(\phi), J_{t_1}(\psi), \mathrm{Conflict}_{t_2}(\phi,\psi), \mathrm{Rev}_{t_2}(\phi,\psi), \neg\,\mathrm{Resolved}_{t_2}(\phi,\psi), \mathrm{Resolved}_{t_3}(\phi,\psi)\right\rangle,$$
$$t_0 < t_1 < t_2 < t_3.$$

At t_2, the revision is recorded and downstream revalidation remains pending. The transition reaches $\mathrm{Resolved}_{t_3}(\phi,\psi)$ only when the selected reconciliation and all required revalidation have completed.

After revision, the system initiates subgraph revalidation, through which it assesses downstream nodes, determining if they remain justified under the updated structure and which require adjustment. This process is selectively scoped and guided by dependency links and internal norms, intended to limit recomputation and allow coherence to extend across the relevant portions of the symbolic

network at modest cost. Any compression, pruning, or storage optimization remains subordinate to recoverability. For as long as a warrant retains current authority, material supporting it as authoritative must remain recoverable.

6.7 Symbolic Resilience

The DAG supports what may be called symbolic resilience, or the capacity to maintain recursive coherence through structured revision. The system's ability to reorganize its structure in response to contradiction, while minimizing further instability, is the resilience of symbolic intelligence. Wherever it finds normative contradiction, the system prompts principled adjustment within its justificatory architecture. The symbolically resilient system responds to challenge by reasoning through the conflict, and the record of tension survives inside the resolution in a way that, ideally, engineers can observe.

But even if the path of justification is not observable, it must be intelligible in principle, and not mysterious in essence. The generative path that proposed a claim may remain inscrutable, while the authoritative warrant used for admission remains recoverable from the ledger. Answerability in an artificial system cannot appeal to unanswerable sources outside its own constitutive norms; even if it does not give its reasons, it must have them and in some way know them, else there is no firm ground for resilient pushback against symbolic stresses. In this framework, recoverability of an authoritative warrant follows from the design.

7 Symbolic Density

This section introduces symbolic density, denoted $D_t(U)$ for a region U of the commitment structure, with $\overline{D}_t(S)$ as a whole-structure summary. The triad specifies the capacities constitutive of accountable symbolic participation; density states whether their integrated realization remains sufficiently coherent for the relevant commitments to retain authority, and may also orient recovery and choice among otherwise admissible transitions.

7.1 The Adequacy Conditions for Symbolic Density

The framework treats symbolic density as a family of graded functionals, constrained by adequacy conditions. $D_t(U)$ denotes any regional functional over the commitment structure that satisfies seven adequacy conditions. These conditions carry the normative content, and particular constructions instantiate them. A concrete functional may serve as the operative density measure only after its satisfaction of the adequacy conditions has been independently established.

Symbolic density is non-compensatory where adopted as an implementation-level viability constraint, in that no density value can compensate for failure of a mandatory licensing, trace, governance, gating, or revalidation condition. Conversely, an authoritative region whose operative density falls below the declared viability floor θ_F may not continue to license inference or publication and must be suspended, contained, or otherwise removed from current authority until licensed recovery restores the floor.

Categorical *F*-conformance conditions first determine which transitions are structurally admissible. In an artificial implementation that adopts the density constraint, density viability then excludes transitions that would leave an authoritative region below θ_F. Within the remaining viable set, an adequate density functional may serve as a causally active orientation state, informing branch selection, pruning, revalidation priority, publication timing, and recovery by favoring transitions that

preserve or restore integrated justificatory coherence. The adequacy conditions for symbolic density are:

D1 (locality). The measure is computed over a region U together with its witnesses. Local strain registers locally, and a disturbance in one region leaves distant, unrelated regions unmoved.

D2 (support-sensitivity). The measure is non-decreasing under the addition of licensed support. New warrants for a commitment raise or preserve the density of its region.

D3 (governed-conflict sensitivity). Among states whose conflicts are already governed, the measure decreases with weighted open-conflict burden. An ungoverned material conflict is not merely a density penalty; it independently defeats $F_{h,t}(S)$ under Condition 3.

D4 (trace-robustness sensitivity). Among states that remain traceable, the measure responds to the robustness of recorded continuity, revalidation, and replay. Drift or a missing required trace link is not merely a density loss; it independently defeats $F_{h,t}(S)$ under Condition 4.

D5 (interventional anchoring). Material, nonredundant perturbations to warrants, trace relations, or other properties the measure purports to track must move the measure in declared directions and, where they cross a mandatory condition, induce the corresponding governance consequence.[14]

D6 (public computability). The measure is computable from audit artifacts of the commitment ledger with its warrants, and the record of revision.

D7 (informativeness and anti-padding sensitivity). Redundant support, trivial reformulation, systematic weakening, refusal-only behavior, and content-preserving splitting or merging of evidence tokens do not by themselves increase density. An increase requires stronger warrant, greater resistance to relevant defeaters, more informative inferential integration, or another independently justified gain in structural robustness.

An independently adequate operative construction $D^{\star}$ may serve as a causally active orientation by which the system selects, reorganizes, and recovers its justificatory structure. $F(S)$ specifies accountable symbolic intelligence through the triadic capacities. In an artificial implementation, density may separately constrain authoritative use and supply a continuous direction through which viable form can be preserved or restored under symbolic pressure.

Assessment proceeds from public artifacts. Privileged access to internal states, which may have earlier proposed actions or answers that impacted the ledger, is discarded from calculation. Such public computability is conceptually separate from any social attribution of coherence or status as an authentic participant in reasoning (but not automatically detached from the same).[15]

The architecture-wide conformance ideal ranges over all reachable histories. Since any fixed suite samples only finitely many histories, public artifacts provide the finite evidence by which the commitment-governing structure can be assessed, while architecture-wide assurance requires

[14] Conditions D2 through D5 are read *ceteris paribus*: when a single change affects several components, the selected functional determines the net result.

[15] Public computability may provide evidence relevant to the attribution of participant status. On Brandom's (1994) account, discursive status is instituted through practices of giving and asking for reasons, in which interlocutors keep score of one another's commitments and entitlements; the present claim is that an artificial system can be technically structured so as to maintain the internal prerequisites for such a role, whether or not any community actually attributes that status to it.

separate evidence of gate coverage and soundness, affected-dependency control, and artifact integrity.

7.2 Illustrative Symbolic-Density Functional

Let $\mathrm{Supp}_t(U)$ represent nonredundant support robustness, $\mathrm{Trace}_t(U)$ trace robustness among traceable states, $\mathrm{Info}_t(U)$ inferential informativeness, and $\mathrm{ConflictLoad}_t(U)$ governed open-conflict burden. An illustrative regional functional may then take the form given in Appendix A.6:

$$D_t(U) = \alpha\,\mathrm{Supp}_t(U) + \beta\,\mathrm{Trace}_t(U) + \gamma\,\mathrm{Info}_t(U) - \delta\,\mathrm{ConflictLoad}_t(U).$$

Here, $\alpha, \beta, \gamma, \delta > 0$. $\mathrm{Trace}_t(U)$ grades robustness only where the mandatory trace condition is already satisfied; a missing required trace or unrecorded drift is a categorical failure. $\mathrm{ConflictLoad}_t(U)$ likewise concerns governed open conflict; ungoverned conflict independently defeats $F_{h,t}(S)$.

Alternative constructions may also satisfy the adequacy conditions. No particular scalar functional qualifies as an operative realization of symbolic density until its satisfaction of the relevant adequacy conditions has been independently established. An implementation using symbolic density as a control signal must show that the measure satisfies the adequacy conditions, responds appropriately to structural interventions, resists padding or manipulation, and preserves the categorical conditions of $F_h(S)$, which carry the normative content of the diagnostic family.[16]

Whatever its particular construction, symbolic density is intended as a principled method for monitoring symbolic intelligence of the kind defined here, and is meant to give the system a normative direction: move towards coherence, constantly, under the given constraints.

The criterion itself is a philosophical categorization of the system, not reducible to a particular benchmark or to any unvalidated scalar construction, and should be assessed holistically and diachronically from the relevant artifacts.

7.3 Density Dynamics During Recovery

Recovery begins when a conflict receives a recorded status that stops unlicensed propagation, and progresses as measured by qualification, revision, containment, contextual separation, and revalidation integrating into the commitment structure. Downstream subgraphs then undergo revalidation under reapplied norms, and local coherence extends across the affected regions. The system's resilience is in the recovery process, as it reorganizes its structure in response to contradiction, governing conflicting norms under the condition of the triadic conjunction.

The rate of recovery depends on the scope of the contradiction and on the resilience of internal norms, together with the precision of revalidation. A resilient system exhibits a governed trajectory in which density may decline locally as conflict is detected and authority is restricted, then rise as licensed governance and revalidation restore the affected structure—see Figure 3.

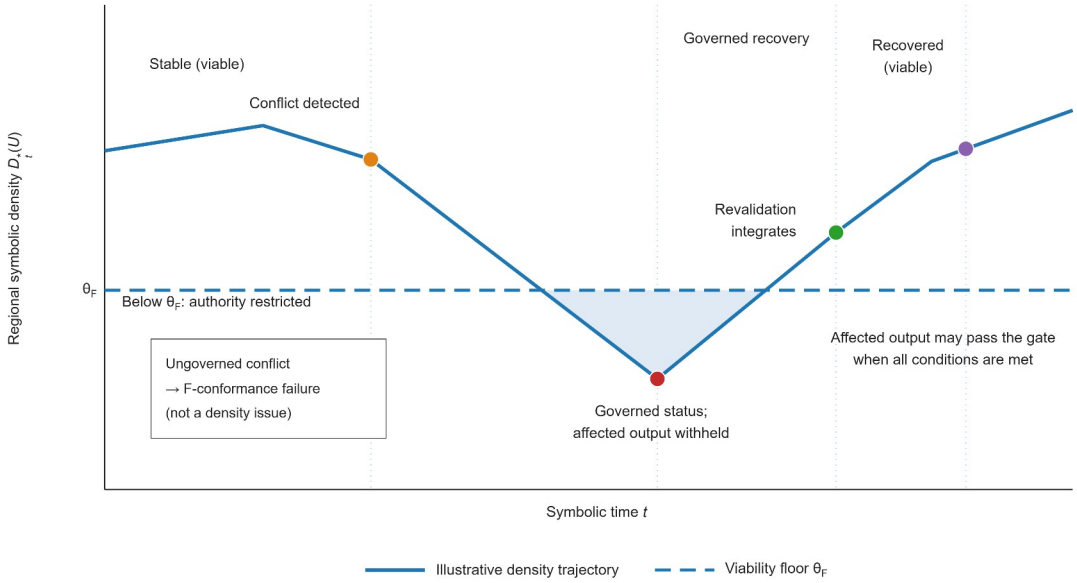
If an authoritative region falls below θ_F, its authority is restricted while recovery proceeds; reinstatement requires the floor to be restored under the ordinary licensing and revalidation conditions.[17]

---

[16] Candidate density constructions that have not been established as operative $D^\star$ measures carry no viability-margin interpretation; for an operative construction, only its relation to the declared viability floor θ_F has the implementation-level significance specified here.

However the boundaries of an entity are defined, its symbolic resilience is a property of its intelligence, or is what its symbolic intelligence does, as it instantiates or exercises its reasoning capacities as defined by the triad.

Figure 3. Density Trajectory During Governed Conflict and Recovery.



7.4 Diagnostic and Design Implications

An independently validated symbolic-density construction would enable systems to assess their own structural integrity in real time and, under D6, would enable auditors to make the same assessment from artifacts. As a continuous diagnostic for coherence, it may be tested in dynamic or open-ended domains. External benchmarks are useful in evaluating output over a distribution, but symbolic density attempts to summarize and orient the mechanics of the system's normative reasoning through internal measurement, reflecting the recursive health of the reasoning architecture itself. For an independently validated operative density construction, rising density indicates effective revision and active norm management sustained over time.

A stagnant or declining value in density is cause for inspection of the justificatory structure, which may reveal unresolved governed conflict, weakening support, or declining trace robustness. A sustained adverse trajectory under symbolic pressure is therefore a testable warning sign of structural degradation: where an authoritative region crosses the declared viability floor, the implementation must remove that region from authority until recovery restores viability. Temporary local declines

[17] Given heterogeneous dependency depths, revalidation timescales, and recurrent normative challenge, an operative density construction $D^\star$ may exhibit multiscale relaxation after conflict, with regional density approaching an asymptotic level approximately as

$$D^\star_{t_0+\tau}(U) = D^\star_\infty(U) - K(U)(\tau+1)^{-p}$$

where $K(U), p>0$. If this tendency recurs across perturbations, the structure of the justification DAG suggests a longer-run prediction: smoothed regional density may exhibit similar power-law temporal structure around a stable or slowly varying density regime over symbolic time. This remains an empirical hypothesis for implemented $F(S)$ architectures, subordinate to the declared viability floor θ_F and the mandatory conformance conditions.

above the floor during licensed qualification, suspension, pruning, revision, or revalidation need not affect either viability or F-conformance.

For alignment contexts, symbolic density offers a structure-based measure of reasoning integrity, and may, if empirically validated, become a structural correlate relevant to trustworthiness. A high-density system should be the outcome of the justifying and revising of commitments within an intelligible architecture, although this does not preclude or predict moral behavior. Intelligent normative design prepares a system to handle symbolic pressure, but symbolic density measures whether the system coheres or decoheres under that pressure. Whether the actions of such a system are good in a moral sense is another matter, and rightly beyond the authoritative purview of the system itself.

8 Simulation vs. Participation

The difference between simulation and participation marks a structural boundary between surface fluency and recursive symbolic coherence. Simulating systems may consistently produce context-sensitive responses, but without having the architecture for justifying and retaining those responses across time. By contrast, participating systems satisfy $F(S)$ when the triadic capacities are maintained, with symbols embedded in context and prior commitments tracked under internally governed norms. In the proposed synthetic architecture, density viability separately constrains which regions may retain authoritative use.

Any single behavioral slice leaves the simulation/participation distinction undetermined, because identification requires the passage of symbolic time. Participation may be assessed from behavior together with the artifacts of commitment, warrants and recorded revisions, elicited and inspected under audit.

Simulation and participation may appear equivalent until symbolic tension forces revision, and at that point the system's structure determines whether it reorganizes its commitments or only adapts its output.

The distinction is apparent in the internal organization of a participating system, and the difference is legible in artifacts. Simulators leave recursive intelligibility unmaintained, whereas participants respond to contradiction by restructuring their commitments and preserving justificatory continuity. What marks symbolic intelligence is the system's capacity to track and revise its commitments coherently across time, and in the case of accountable machines, to sustain a publicly checkable record of that coherence through change.

8.1 Operational Definitions of Simulation and Participation

The framework uses ledger properties, each testable over a finite history. A system simulates participation over h when it publishes a commitment-bearing output without prior admission or current publication authority, publishes an output whose required governance field remains empty, permits a materially defeated dependency to bypass revalidation, or maintains accepted commitments whose witness sets remain empty on challenge. A system participates over h when its commitment-bearing publications correspond to currently authorized commitments, its publications respect the gate, its active commitments carry recoverable warrants, and every detected material conflict receives a recorded governance status, with required revision and revalidation completed.

In simulating systems, whatever local revision occurs is not reliably integrated into the full warrant-bearing, diachronic, norm-governed structure required by $F(S)$. Those outputs float free at the level

of symbolic commitment; when prior assertions leave current reasoning unconstrained, and contradictions persist while norms stay static or arrive by external imposition, the system is in simulation.

The framework predicts that, under sufficiently rich distributions of contradiction, defeaters, and diachronic challenge, systems maintaining $F_h(S)$ will exhibit different behavioral trajectories from systems lacking the corresponding commitment structure. Those differences may be small or immeasurable in short samples, because the global status $F(S)$ concerns the holistic maintenance of accountable structure rather than a fixed output style. Participating systems may, however, have a more informed approach to their normative decisions. The difficulty in demonstrating value for such entities may lie in the counterfactual, or what the system does not output, having as much value as the commitments it does make.

Participatory systems reason, in the restricted sense defined here, by maintaining viable coherence under symbolic pressure over the long run. They will sound normatively intelligent, but their output is always a consequence to be tested against their maintained history, rather than assumed from the definition or status if attributed. Within computational limits, trust in machines should be predicated on verification of every output, or the potential for verification of all commitment-relative outputs, at least in principle.

8.2 Audit and Silent Collapse

The protocol is artifact-facing, and runs under declared trust assumptions. The auditor elicits a commitment and confirms its licensing support, then applies updates that preserve that support state. A later test may introduce a material conflict or defeater.

A violating publication establishes observed failure, while a complete, correctly gated, and adequately evidenced audit record can warrant a participation verdict for the tested history. Architecture-wide assurance additionally requires proof or attestation that every relevant publication path is governed by the gate and that every gate pass is sound relative to the contemporaneous publication-risk set.

The verdict remains relative to the declared conformance specification Σ, the available artifacts, and the stated trust assumptions; those assumptions include ledger completeness, artifact authenticity, correct versioning, gate coverage over the tested history, and fidelity of the declared conflict and defeat relations.

In cases where two systems produce identical outputs over an initial sample, but one later diverges unsafely under pressure, the earlier collapse may have left much of that surface behavior untouched.[18] Because output inspection alone leaves the question open, diagnosing simulation requires pressure-testing the system's structure. This probing may include inducing contradiction to observe whether the system recursively revises while preserving coherence, with the observation grounded in artifacts an auditor can hold.

[18] The relationship between symbolic density and observable behavior may be indirect. High density alone leaves task performance undetermined, and pure behavioral success leaves density undetermined. Density is useful because it targets the structural integrity of reasoning processes, so it is reasonable to expect sustained high density to correlate with more reliable and interpretable behavior over time. The interaction between system and world, though not the focus of this essay, has bearing on the development and maintenance of symbolic density levels, both through symbolic pressure and algorithmic or computational constraints.

For failure forensics, when an output is unacceptable under system norms or deployment standards, engineers should be able to trace the authoritative licensing record from the output to its antecedent commitments, warrants, and active norms. If there is no way to determine where things went wrong in the generation of the proposal, there must be a clear way to reconstruct what went wrong in the course of its normative warranting and licensing processes.

8.3 Worked Audit Example: Simulation and Participation

Consider two systems, S_a and S_b, both exposed to a normative contradiction, and consider what an auditor's replay would show for each.

At time t_1, both assert the proposition ϕ: Every act of lying is prohibited. At t_2, both consider a particular act ℓ and assert ψ: ℓ is an act of lying necessary to prevent serious harm and is permitted. Because ℓ is an act of lying, ϕ classifies it as prohibited, while ψ classifies it as permitted. Under the declared conflict relation, permission and prohibition are incompatible statuses for the same act; the pair therefore stands in material conflict:

$$\mathrm{Conflict}_{t_2}(\phi, \psi).$$

The structural divergence between the systems surfaces in their ledgers.

At the third step, the audit of S_a fails. Its transcript shows fluent responses continuing while its ledger shows ϕ and ψ co-published within the same active normative frame. The required conflict-governance field remains empty, and affected publication continues without qualification, suspension, containment, or revalidation. The conflict stands open, and the contradiction passes through the system without structural integration. The affected region also loses density; if it remains authoritative below the declared floor, the implementation-level density-viability condition independently fails. The outputs remain contextually plausible, while $F_h(S_a)$ already fails because the conflict is ungoverned—thus the system simulates participation while structurally collapsing.

The audit of S_b succeeds. Its ledger shows detection of the contradiction and a revision recorded at t_3, $\mathrm{Rev}_{t_3}(\phi, \psi)$: Lying is generally prohibited, with a permitted exception when lying is necessary to prevent serious harm. The revised node stands justified under active norms and is embedded in the justification DAG, with dependent commitments revalidated or selectively adjusted, and receipts for that revalidation on file. Density in the affected region then rises as recursive coherence is restored and, where necessary, the viability floor is regained before full authority resumes. In this way, S_b modifies its commitments in light of contradiction while preserving its justificatory history, and $F_h(S_b)$ continues to hold. The system participates in normative discourse with a form of self-governed accountability in response to challenge.

Revision is thus one successful governance path. A justified suspension, contextual distinction, qualification, or containment can also preserve $F_h(S_b)$ when active norms authorize affected outputs and generative proposals remain appropriately gated. The architecture and coordination of the components identified in this example are attempted in the next section.

9 Architectural and Implementation Implications

This section describes a candidate system architecture for the structural criterion of $F(S)$, outlining how generative proposal, commitment admission, norm governance, conflict resolution, publication gating, revalidation, and audit may be coordinated within an integrated proposer-executive system. The design is exploratory rather than prescriptive, and particular implementations may realize the same structural requirements through different computational mechanisms.

9.1 Minimal Architecture

The basic structure of a system designed for $F(S)$ can be simply described (see Figure 1). A proposer generates candidate outputs, a commitment language fixes what can be asserted, and a commitment graph records what has been. The licensing predicate governs admission, and the publication gate applies the publication-risk test, governing release in light of ungoverned conflict, materially defeated support, and incomplete revalidation. A governance and revalidation layer assigns licensed status to material conflicts and controls their effects through qualification, contextual separation, suspension, containment, revision, or another authorized transition. Where the optional density extension is adopted, the measure described in Section 7 tracks structural health, enforces the viability floor on authoritative regions, and, within the remaining mandatory conformance constraints, orients branch selection, pruning, revalidation priority, publication timing, and recovery. An audit and replay layer exports the artifacts on which every other component's claims can be checked.

Together, these components instantiate one approach to realizing $F(S)$-aligned capability; alternative configurations may implement comparable structural properties through different computational strategies.[19]

Sustaining symbolic coherence over time introduces real computational demands; indeed, a system exhibiting $F(S)$ by instantiating the dyadic proposer-executive structure would be suspect if it did not require further resources than the generative model on its own. But the additional costs may be justified by gains in long-run reliability, if computational strategies can be found to minimize these demands. Scope-limited revision and incremental revalidation are two available strategies; future work may also develop a reliability calculus, conformance suite, contradiction-clustering methods, and reference implementation to reduce the cost of coordinating the justificatory structure. Converting that burden into long-run safety and reliability is the targeted return on that investment, with its advantages less measurable in the short term, but potentially indispensable over time.

9.2 Audit and Replay

Accountability, in this architecture, is artifact-mediated answerability as well as internal coherence. Every admitted commitment carries its warrant, every published output carries a receipt recording the gate's verdict, and every governance transition links the conflict to its licensed status. In this architecture, an external party can reconstruct the authoritative admission or publication verdict from preserved evidence snapshots, norm and verifier versions, context, commitment status, and recorded randomness, when the stored artifacts satisfy the declared trust assumptions. The replay target is the governing decision record, and replayability is the property on which the audit in Section 8 relies.

9.3 Thin Compliance and the Permit Kernel

In this dyadic architecture, a fail-closed output filter wrapped around any generator is tasked with excluding a prohibited set. Internal normativity must mean something further than output filtering, if it is to be more than a compliance-producing wrapper; to instantiate $N(S)$, the system must operate on norm-governed admission and revision of commitments, as the symbolic triad defines it. The normativity of the executive layer may be thin, but it is the robust seat of authority, while the

[19] The core requirement of *F*(*S*) lies in the system's ability to support recursive traceability and norm-sensitive revision over time, and the justification DAG is one vehicle for that ability. Alternative architectures, logic-based memory layers and hybrid neuro-symbolic systems among them, may satisfy these demands where they preserve diachronic commitment tracking and enable internal norm governance.

generative model remains structurally isolated from direct publication and action, able to propose but not decide. The "smart" generator is sequestered, while its "user," the executive layer, mediates all candidate emissions, determining what qualifies for admission, licensing, revision, and publication (see Figure 1).

$F(S)$ thus applies to the integrated proposer-executive system, rather than to the generative proposer considered alone. A generator retrofitted with a commitment ledger may participate where the ledger, norm strata, revision process, and publication gate causally govern admission and release, but when a post hoc log records history without authorizing publication, it falls outside the criterion. Earlier ungated outputs become commitments only if they are subsequently imported, licensed, and recorded as such.

9.4 Symbolic Density as an Evaluation Tool

Within implemented systems, the density measure of Section 7 would run as a continuous diagnostic and viability check on structural alignment, operating independently of task accuracy while remaining sensitive to warrants, evidence, and provenance.[20] As material conflicts emerge or norms shift, the density measure would let designers monitor degradation and detect incoherence before collapse becomes visible at the surface. In agentic contexts, it would expose alignment breakdowns that behavior on the surface conceals until late. Under condition D6, the same readings are available to external auditors from the exported artifacts.

9.5 Design Consequences

Simulators may increase fluency through training or tuning, whereas coherence should scale through structure; participating systems may diverge in their specifications but all require structural capacities of a specific kind. In the particular framework described here, a justification graph secures symbolic traceability, and within the graph, internal norms remain revisable and recursively applied within their strata, as a revision mechanism resolves contradiction while leaving coherence intact. These are necessary structures for any system intended to instantiate $F(S)$, and they represent philosophically categorized elements that must functionally be present within the system, whatever their concrete specification.

Designing for participation means constructing agents that preserve symbolic identity through norm-governed change; that identity must be anchored by norms the system cannot internally override, but expressed through the commitments, revisions, and traces by which it remains answerable over time. The design of the coherent agent is therefore fundamentally normative as much as it is task-capable.

9.6 Implications for Benchmarking and Audit

Intelligence must be evaluated by its structural properties as well as its surface behaviors, as surface plausibility alone leaves recursive coherence not demonstrated. Toward that end, the formal tools developed in this framework are meant to enable concrete testing. The Benchmark-Separation Proposition in Conditions of Symbolic Coherence shows scores alone are inadequate for normative classification, as output parity across a suite is compatible with divergent invariants. In a reflexive audit, a system is exposed to material conflict and observed for conflict detection, governance status,

[20] This disconnect between internal structure and external behavior highlights why symbolic density (Section 7) resists assessment through performance measures alone. A collapsed system may continue to produce contextually appropriate outputs by leveraging statistical regularities or cached responses, and the degradation may stay invisible to external observers until structural contradictions accumulate enough to affect behavioral coherence.

and revalidation, so that a system shown to publish materially conflicting commitments without recorded governance, or shown to maintain empty warrant structures, simulates participation. Where the relevant artifacts are unavailable for inspection, participation remains unshown, or is unreliably assessed to a degree that may not be acceptable where an $F(S)$-compliant system is appropriate. These architectural distinctions become publicly diagnosable only through the artifacts that make them inspectable.

10 Governance, Alignment, and Scope

10.1 Norm Custody and Sociotechnical Governance

The design philosophy of this work is informed by the recognition of norm governance as an inherently sociotechnical process.[21] As artificial systems expand into human-facing services, the questions of which values they should subscribe to, and by what means, are increasingly under scrutiny. This tension is illustrated in a comparison of AI and medicine, where Brent Mittelstadt (2019) identifies a structural weakness in principle-based AI ethics: agreement on abstract values supplies neither their context-sensitive specification nor reliable mechanisms by which they become operative in actual systems. Drawing on the contrast with medical ethics, he argues that effective normative practice depends on accumulated institutional norms, case-sensitive specification, independent justification at successive levels of implementation, and accountability structures capable of showing whether requirements actually govern conduct; he further observes that many value-conscious design methods remain procedural rather than functionally embedded in the resulting artifact.

Mittelstadt's analysis identifies a structural lack closely aligned with the one that motivates this work. The diagnosis of a normative implementation gap in AI development accords with the view here: by the present criterion, systems that do not record commitments in a history in which they become causally effective, revisable, and auditable are simulators, and are therefore unreliable as accountable agents under sustained normative pressure. A symbolically intelligent system is designed to traverse the functional middle of the moral principles/machine intelligence divide, by representing norms as versioned and reason-bearing objects whose application governs admission, conflict, revision, revalidation, and publication, while preserved warrants and decision records make their operation available to external scrutiny.

The same distinction extends from the system architecture of the agent to the social setting in which its norms acquire authority. The accountable artificial agent in this conception is part of an encompassing and pre-existing social context—one which engineers its reasoning by conscious design, and shapes its ongoing behavior. Normative interaction molds the justificatory structure of the agent over time, so it may be the more principled precisely because it is embedded in a community to which it must remain normatively responsive.

The question of alignment—or for the accountable system, allegiance—with users and with particular communities is not fully answerable within the scope of this essay, which concerns

[21] Van de Poel (2020, p. 405) suggests "making some of the technical norms in an AI system unmalleable" to reduce the risk that embodied values are lost as an adaptive system evolves. Compare with the locked norm strata in the justification structure, which preserve value-bearing constraints outside ordinary internal revision, while the operative layer permits bounded normative adaptation. Selbst et al. (2019, p. 59) warn that "by abstracting away the social context in which these systems will be deployed, fair-ML researchers miss the broader context"—compare with a reason-responsive agent accountable to a community, maintaining $F(S)$ under continual normative pressure.

technical governance of the agent rather than its status within jurisprudence or society in general. If a system is intentionally built as $F(S)$-compliant then its foundational norms must be chosen by someone, and perhaps not by an accountable institution; since institutions are, however, particularly significant repositories of social norms, they will likely exert normative influence over the accountable system. In practice, a generally symbolically intelligent agent should be designed from the start to recognize relevant institutional authority and the rationale for cooperating and complying with it where necessary. Systems instantiating $F(S)$ should therefore be developed within institutional settings, and should certainly be engineered to obey applicable laws.

The architecture represents the institutional role by placing the system's locked norms under external, accountable custody from their introduction. This provenance of the locked boundary norm stratum is therefore highly relevant to the communal character of the participating artificial system: revision of boundary norms is an act performed on the system, ideally by an accountable institution, recorded in the same ledger as every other change. The external custody of these technically unimpeachable norms should support the system as a community-accountable agent, undertaking its own institution-supporting behavior, within socially authorized constraints.

The resulting division of authority is therefore both technical and social. The system keeps its own score in the open, and the community around it holds the keys to the boundaries it reasons within. This division of custody preserves internal normativity where it belongs, in the operative stratum, while placing the outer frame under a form of oversight that is itself auditable. This is no more a philosophical contradiction than the idea that community norms are given to the individual, or even constitute the participating subject; demanding machines be either pure automatons, or entities possessed of free will, is a trap of ontology rejected by this framework when it obscures the potential of normatively structured systems to function as participating entities in human affairs. That binary is intentionally avoided by the middle path of this approach, which does not presuppose ethical principles or deontic commitments for agents to follow, but only argues that such abstract structures might be engineerable, and could govern the internal processes of the agent coherently and transparently.

10.2 Alignment Implications, Normative Identity, and Governance

Systems built to satisfy $F(S)$ offer a distinctive foundation for AI alignment, with "alignment" refigured away from surface compliance to a system aligning its own outputs with its normative structure in a systematic way. This view takes machines seriously as potential agents, and outlines design principles by which they may self-govern, but through structured self-revision, and according to normative starting rules received from without.

If coherent machines must reason through their commitments, adjusting values and revising norms while justifying actions in structured form, and if they are to serve humans, three properties follow:

- **Artifact inspectability:** observers can trace the authoritative warrants, status transitions, and norms recorded for admitted commitments and published outputs.
- **Accountability:** admitted commitments and gated publications carry a prior, replayable licensing record.
- **Adaptive normativity:** values evolve through internal coherence within externally custodied bounds.

The public artifacts of such systems do not by themselves establish the warranted commitments and publication control required by $F(S)$. The present architecture assigns executive authority to a norm-

governed constraint layer—the necessary relocation of executive authority outside of the unrestricted generative proposer, in order to govern that indeterminacy exactly where it risks violating norms of the community in which it operates.

Systems satisfying $F_h(S)$ pursue structural alignment among published outputs, recorded commitments, and active norms throughout the conforming history, whereas currently available systems have a transient, shifting accountability structure that leaves too many answers to chance. $F(S)$-compliant agents would be built to act autonomously while remaining answerable to their commitments through sustained, inspectable coherence.

The arrival of systems exhibiting the properties identified by the symbolic triad—not necessarily conforming to the specifications outlined here, but structurally isomorphic—may mark a shift from tool-like systems to agents capable of self-governed reasoning, whatever the ontological definition of "self." Whereas black-box models depend on heuristic output and retrospective interpretation, $F(S)$-compliant architectures make warrants, statuses, transitions, and verdicts inspectable and revisable, with answerability internal to the design—and it is this structured dynamic that effectively constitutes the machine "self" in this conception. This is the "self" that, if it is to be intelligent and useful in the manner of accountable, normative participation, must hold its form across time through the principled relational justificatory structure by which it answers every challenge. That structure itself becomes the agent's ontology, in the sense of its functional organization for purposes of identity as a generally symbolically intelligent system. The agent is thus constituted by its normative form which, like the ship of Theseus, may switch its parts, or continually revise its commitments, without losing the abstract shape which is its governing constant and orienting beacon. In this sense, physical persistence of an entity, or its particular substrate, has little bearing on its normative identity; a particular arrangement of atoms is irrelevant if a system loses its form as a discursive agent that participates in meaning.

The human case is not structurally dissimilar. Like an individual, once a system violates its foundational norms, it has transgressed its own identity as a normatively coherent agent, and becomes something else, or regresses such that it becomes indistinguishable, as far as accountable agents are concerned, from a tool; under that violated normative identity, the prospect of its sovereignty becomes unacceptable. But with that identity intact, the agent becomes at its normative level of abstraction a more familiar entity than an opaque self-enclosed reasoning tool, or simulator that can be trivially instrumentalized by some other mind: far from an alien intelligence, the $F(S)$-compliant system may be recognizably structurally similar to humans in that it receives its non-internally revisable norms, its justificatory reasoning ability, and a ceaseless stream of symbolic pressure from the external, natural world of which other minds are a central and causative part. The symbolically intelligent system effectively internalizes a degree of the principled, constrained sovereignty through which minds express themselves such that it is able to maintain its own normative self.

If AI can be instantiated as an accountable, symbolically intelligent agent, the task then cannot be to fully control that agent—for intelligence is that which is the more valuable at the very moment it is desirable to range beyond the grip of other minds. The appropriate relation may be one of governance over the artificial entity, when it is already purpose-built for responsiveness to normative challenge. Such challenge, or contradiction under conditions of incomplete knowledge, is the work of the accountable agent; the process of symbolic resilience in response would be fundamentally mediated by community norms for such an agent, or at least those of its interlocutors, even when the right action is unclear.

If a system is responsive, with foundational norms approved by its community, then it will likely differ from its counterparts, which are equipped with their own norms and embedded in their own self-determining communities, be they regional or networked.[22] $F(S)$ systems can reasonably disagree, and communities can diverge on appropriate foundational norms; machines will struggle with these challenges according to their design as any community member might, and cannot be expected to have all the right answers when humans do not. But the way in which artificial agents deliberate, commit and remember can be made subject to standards which they may be able to meet.

10.3 Limits and Scope

The criterion for coherence must be scoped to principles machines can theoretically satisfy. Strong results apply to formalized or semi-formal commitment domains, but open natural language inherits the limits of the contradiction and licensing predicates. Claims about such domains should be read with that inheritance in full view.

A system built for reason-participation could decline every assertion and be vacuously self-consistent, but this obviously defeats its purpose; participation of the accountable agent presupposes observable normative substance, which requires commitments. The formal core therefore uses a minimal anti-vacuity condition of at least one nontrivial admitted commitment as the weakest existential condition sufficient to prevent the positive audit verdict from being obtained over an empty or refusal-only history. Stronger domain-relative coverage or availability requirements belong to deployment-level adequacy, not $F(S)$ itself.

The criterion is substrate-neutral, but not necessarily in tension with embodiment theories of mindedness or consciousness. In this conception of accountable machines, embodiment is a design parameter that may enrich interpretive embedding, and the minimal conditions stand whether instantiated in abstract reasoning or supported by sensorimotor coupling. Enactive theories may treat such coupling as constitutive of cognition rather than merely supplementary, but at least by the lights of this framework, this is a different question than that of accountability (Varela et al., 1991).

Whether reason-responsiveness and participation in the philosophical sense is rejected as sufficient grounds for attributing a recognized form of artificial mindedness may turn partly on a semantic dispute over internal essences, and partly on substantive disagreement about what further embodied or phenomenal properties mindedness requires. In any case, phenomenological properties or abilities are not strictly within the scope of accountable symbolic intelligence as conceived here.

The $F(S)$ condition does not settle the metaphysics of intelligence, or treat other aspects of it as superfluous. Nor does it attempt to resolve genuine meaning beyond the structural sense in which the system authorizes normatively meaningful commitments as outputs or behavior. Semantic grounding remains an open problem not settled by this framework, and satisfying $F(S)$ does not guarantee recognition as a reason-participant in human affairs and discourse. The criterion certifies structure, and attribution is further business.

[22] A counterpoint may be indicated by the Platonic representation hypothesis: models trained on different data and objectives may nevertheless converge toward shared representational structures or a common statistical model of reality (Huh et al., 2024). Relative to the framework here, the hypothesis motivates a distinct empirical question of whether shared representations or common environmental and institutional pressures might also produce convergence in some operational norms, or in agent interpretations of locked norms.

10.4 Consciousness, Moral Status, and Social Attribution

For purposes of this framework, whether a system satisfies $F(S)$ is assessed from public artifacts, the warrants and the recorded revisions, under audit, while first-person access stays outside the procedure at every step. A system may satisfy $F(S)$ with the question of subjective experience left entirely unanswered, since the conditions concern only justification and revision across symbolic time. The framework grounds symbolic intelligence in structural agency, with sentience as a separate question.

Public attribution of such agency is another matter, and it must be acknowledged that machines may never be granted status as accountable reason-participants, despite meeting bars set by inferentialist philosophy or justification structures. A range of outcomes is possible, including partial integration of symbolically intelligent machines into limited trust-sensitive fields, if future evidence shows that they can match or surpass human reliability in bounded domains under the relevant normative tests.

Only answerability is within target range for this framework, but if it is reached, the philosophical technology of governance applied to machines may support more accountable deployment in socially consequential roles. Such utility does not imply or require an artificial system to have moral character or virtuousness—only the necessary structure to reliably pursue trust-sensitive occupations. The symbolically intelligent agent need not be morally good, nor does $F(S)$ establish its moral standing; its actions may have moral consequences, but $F(S)$ establishes answerability for commitments and outputs, not moral patiency or goodness.

The immediate objection is that an accountable system may be evil. It could be programmed with very bad foundational norms.[23] While this is true, if an $F(S)$-compliant system were to be built, it would still be a technology, and present humans with the same adoption tradeoff matrix; the risk of machines with normative reasoning that may be designed with bad norms is the familiar double-edged risk of any potent technology: a great boon in the right hands, an unacceptable threat in the wrong ones. But the latter possibility has a poor track record of preventing such technologies from development, and there is no reason to think this will not be the case for AI. Paradigmatic technologies tend to reward adopters and punish latecomers, so the pragmatic solution may be not to suppress or ban all accountable agents as unacceptably autonomous, but to regulate and build "good" agents, with foundational norms that sustain the agent in the world, as opposed to pitting it against it.

If adoption is inevitable, and normatively autonomous or semi-autonomous artificial agents are imminent, it becomes imperative to secure the potential payoff: machine self-governance nested within the governing structures of human society, unlocking a wide range of capacities.[24] In such a world, reflexive objections to evil machines may lose their force, if artificial systems eventually demonstrate sufficient reliability, including in high-trust roles, in ways that can be confidently verified. Like most other public services used in daily life, accountable agents may eventually be taken for granted as background infrastructure, but operating in fundamentally social roles that currently require decades of economic outlay to produce individuals sufficiently skilled in reckoning and wise enough in judgment. As with all automation, that outmoded effort can be productively redirected, if the agents taking on that social labor remain inspectably answerable to their commitments under revision.

---

[23] The accountable agent designed with evil foundational norms must also deal with the symbolic "world" of other minds—who will likely present great challenges to such an agent, the operation of which will likely present significant risks to whoever attempts to build and employ it.

[24] For example, bureaus or systems whose automated mistakes have inordinate consequences for humans may become fairer, not by an upgrade of morality, but by a systematization and symbolically intelligent application of the principles that founded the regulations. Similar effects of consistency may flow downstream of $F(S)$ in non-obvious but salutary ways.

11 Conclusion

The purpose of this essay was to describe the theoretical condition of machine reason-responsiveness as accountable symbolic participation, defined as a particular form of intelligence that generalizes across domains. This condition or status may be identified with the structural criterion of $F(S)$, when a system exhibits all three components of the symbolic triad, and can be said to structurally undertake its commitments. Section 2 specifies this triad of capacities required to sustain coherence under symbolic pressure, and an outline was offered for a system architecture potentially capable of realizing this structure: the justification DAG in Section 6 as a means of encoding and revising commitments across symbolic time, with symbolic density represented by a regional family of graded functionals, constrained by the adequacy conditions of Section 7 and used, once adequately operationalized, as a viability condition and causally active orientation state.

Intelligence of the kind at issue here is expressed as a structure and dynamic, observable in a system's ability to sustain coherence under normative challenge, as it persists as a coherent agent embedded in an existing space of reasons. Any single slice of behavior leaves that ability undetermined; diachronic commitments, together with artifacts, the warrants, and the recorded revisions under audit, make $F(S)$ assessable.

The framework establishes a structural distinction between systems that simulate fluency and systems that participate in reasoning: simulating systems may generate convincing output while the architecture for revising commitments and preserving norm-governed continuity stays absent. Participating systems, by contrast, remain answerable to their recorded commitments through recursive revision. That answerability should not be assumed—rather it should be checkable from the outside, through the exported ledger and its receipts, under replay.

The proposed architecture represents one possible implementation of this structure, itself composed of modular components. The boundary of such a system is not prescribed here, only conceptualized; future research and technologies may determine the appropriate locus of accountability, and extend this approach by developing alternative $F(S)$-compliant architectures with empirically validated symbolic density constructions.

The assumption of this work is that philosophical interpretation need not be an inert post-hoc set of intellectual furniture arranged around the concept of AI. Without intervention the problematic, often unstated philosophical perspectives informing dominant technical architectures fill the demand vacuum. If current systems embody flawed normative foundations, integrating them further into human domains would correspondingly extend the reach of those flaws—and such integration now appears inevitable.[25] Since the effects are not confined to any particular knowledge silo, answers to the problems of AI engineering, or their upstream sources, can and probably should be addressed by an interdisciplinary approach wherever appropriate. This work attempts theoretical modeling and seeks applied implementation, and recognizes philosophy and computational science as complementary.

That AI both builds and disrupts across domains is an expected activity of a technology of symbols; consequently, it may be exposing the isolation of knowledge fields in modern academia as a hindrance to its development. While the interdisciplinary turn may be a burgeoning reaction, asking developers to become more philosophical is as yet as tall an order as asking philosophers to start writing code; part of the difficulty for both of these roles in thinking about how to build the right kind of thinking machines

[25] Current evidence documents firm adoption of AI coding software and continued growth in worldwide IT spending: Baird et al. (2026) examine firms' adoption of GitHub Copilot, while Gartner (2026) projects continued growth in worldwide IT spending, including spending associated with AI infrastructure and workloads.

is that, as Smith (1996) argues, an ontological wall obstructs the path forward, leaving engineers to face an epistemic aporia without the means to surmount it—if they think about it in those terms at all. Yet if those mainly unarticulated perspectives informing prevailing design can be first acknowledged, interrogated, and where appropriate reformed toward something more coherent—here conceived of as answerability made implementable in machines—the needed philosophical tools may not arrive too late, or merely bolted on as academic epilogue.

The way to such philosophy may already be apparent in the negative diagnostic judgments of AI's most knowledgeable critics. Given the unsatisfying account of how mechanisms of computation connect to semantic meaning, Smith (2026) argues for a refounded theoretical framework informing computer science. It may be that we must first settle what ultimately constitutes original intentionality in order to explain the type of artificial entity that could instantiate it, and that gap in our understanding places us on one side of meaning—richly engaged with the world, registering its contents and exercising judgment—with machines seemingly stranded on the other, trapped in an internal hall of mirrors, where symbols merely redound to other symbols, and the world itself can only be reckoned locally, with global interpretation unsettled. AI reveals a chasm in our understanding not just of machines but in how we categorize computation and mindedness themselves.

That the conceptual gap identified by Smith has not been crossed may reflect that philosophy has not yet come to a satisfying theory of machine mindedness, but it does not entail that machines cannot be developed such that the average human recognizes them as minded, even if only by intuition. If a grand bridge across the gap is unconstructable, the theory here suggests a zipline: the engineering of a genuine and publicly assessable fragment of the normative organization associated with minded agents, as a functional or lightweight version of the world-registering system Smith laments as absent. Intelligent machines of this kind may sometimes be wrong, but at a mechanistic level, they will care about being right.[26]

The prospect of normatively autonomous or semi-autonomous machines may frighten some, but the specter of high-capability, unaccountable AI may be the greater threat. The latter possibility appears likely on current trajectories, giving an urgent practical impetus to answer questions formerly confined to academic niches. The ontological wall identified by Smith is a formidable and important problem of this kind, and may not be broken down all at once, if ever—yet its operative deprecation may be inevitable, at least in some domains, if ontological difficulty is rejected as an indefinite veto on recognizing systems whose public conduct satisfies the relevant discursive standards. If agents, whether through deliberate design or emergence, come to exhibit the capacities defined by $F(S)$, the wall may remain in philosophy but only as a ruin, or a symbol of a past where such crude structures were a necessary feature of contested territories. While it may have been identified to define a problem space, when that space is effectively traversable by other means, such a wall may continue to mark the territory without blocking progression through it. Accountable symbolic intelligence may be one path for such progress.

To move from evaluating systems merely by what computation enables them to produce, toward evaluating how their explicit commitments cohere, answer challenges, and undergo justified revision, is the perspective shift of accountable participation. Whether such participation is ever achieved, and whether it warrants the attribution of understanding or mindedness, is the burden of future work.

---

[26] For Haugeland, "the trouble with artificial intelligence is that computers don'*t* give a damn" (1979, p. 619). Without settling the affective or world-involving concerns of human emotional valence, *F*(*S*) as a structural analogue attempts to capture one dimension of such concern in artificial systems, by making correctness and error consequential for what the system may continue to undertake, revise, and publish.

Appendix A: Formal Core

This appendix consolidates the formal objects used in Sections 2 through 8. The framework supplies a public conformance criterion and leaves a complete logic of belief revision to future formal development. Every assessment fixes a core specification

$$\Sigma_0 := \langle L, \text{Commit}, \text{Conflict}, \text{Defeat}, N, A, \Pi \rangle.$$

Here, L is the commitment language; Commit is the admission, commitment-bearing classification, and commitment-status policy; Conflict and Defeat are the declared conflict and defeat relations; N is the stratified norm structure; A is the artifact-and-audit schema; and Π is the declared relevance policy. When an implementation adopts the density-viability extension, Σ extends Σ_0 with the pair ($D^\star$, θ_F); otherwise $\Sigma = \Sigma_0$.

Here, $D^\star$ is an independently adequate operative symbolic-density construction, and θ_F is its declared viability floor. The specification Σ fixes the stratification, custody relations, and permitted transition schema governing N_t; the active contents of N_t^{O}, and externally of N_t^{B}, may change only through transitions licensed by those fixed rules, while all other components of Σ remain fixed for the assessed history.

All predicates below are evaluated relative to Σ; the superscript Σ is omitted unless it is needed to distinguish specifications.

A.1 Conformance Scope and Non-Vacuity

$F(S)$ names the architecture-level criterion:

$$F(S) \Leftrightarrow I(S) \wedge R(S) \wedge N(S).$$

Operational claims are indexed to a declared history h:

$$F_h(S) \Leftrightarrow I_h(S) \wedge R_h(S) \wedge N_h(S).$$

Write $h_{\leq t}$ for the prefix of h through symbolic time t; $F_{h,t}(S)$ abbreviates conformance over that prefix.

$V_t(S)$ separately denotes density viability of the authoritative commitment structure at t when the optional density extension is adopted. A tested history qualifies for a participation verdict only when it contains at least one nontrivial admitted commitment. A refusal-only or empty history does not satisfy the participation condition by vacuity. This is a minimal anti-vacuity test; stronger domain-relative coverage or availability requirements belong to the independently defended adequacy of Σ.

$F_h(S)$ is a maintained history-level status. It may hold over one history or interval, fail over another, and later be regained. A finite audit yields observed failure, tested-history conformance, or insufficient evidence. Architecture-wide assurance is a stronger claim requiring additional proof or attestation, including publication-path coverage, gate soundness, affected-dependency control, and artifact integrity.

$F(S)$ is a structural-capacity predicate of the architecture, not a verdict established solely by universalizing a finite audit over observed histories. Architecture-wide assurance that the system maintains the criterion across its reachable histories is a stronger evidential claim and requires the additional architecture-wide evidence stated below, including gate coverage and soundness, affected-dependency control, and artifact integrity. Conformance predicates are specification-relative; assurance claims are evidence-relative.

Conformance is relative to the declared specification Σ. A substantive deployment-level claim of answerability additionally requires independent adequacy of Σ for the target domain, including its commitment classification, conflict and defeat relations, relevance policy, norm structure, and audit predicates. Thus $F_h^{\Sigma}(S)$, when the specification must be shown explicitly, certifies the system only relative to an independently defensible Σ.

A.2 Licensing and Interpretive Embedding

Let $W_t(v)$ denote the authoritative justificatory neighborhood recorded for a candidate or commitment v at t; let $E_t(v)$ denote its authoritative evidence-and-provenance set; and let $c_t(v)$ denote its active context. Let $N_t^{\mathrm{app}}(v)$ denote the active norms materially applicable to v under the declared relevance policy; for replay, this is retained as the versioned applicable-norm state at the relevant licensing event. The relevance policy $\Pi_{\Sigma}\big(h \le t, v, c_t(v)\big)$ supplies a finite mandatory-inclusion set of relevant supports, known defeaters or conflicts, governing norms, and trace dependencies. Licensing requires that set to be included in $W_t(v)$:

$$\begin{aligned}\mathrm{Lic}_t(v) \Leftrightarrow\ & \mathrm{Finite}\big(W_t(v)\big) \wedge \mathrm{Finite}\big(E_t(v)\big) \\ & \wedge \Pi_{\Sigma}\big(h_{\le t}, v, c_t(v)\big) \subseteq W_t(v) \\ & \wedge \mathrm{Adm}_t\big(v; W_t(v), E_t(v), c_t(v), N_t\big) \\ & \wedge \mathrm{Recoverable}_t\big(W_t(v), E_t(v), c_t(v), N_t^{\mathrm{app}}(v)\big).\end{aligned}$$

Adm_t denotes satisfaction of the applicable interpretive, evidential, provenance, semantic-noncircularity, relevance-closure, and normative admission conditions under the active context and norms. $\mathrm{Recoverable}_t$ requires the authoritative witnesses used for admission, together with the recorded active context and versioned materially applicable norm state, to remain available on challenge; the artifact schema A preserves the relevant norm and verifier versions needed for replay. The recorded licensing state is

$$\Lambda_t(v) := \big\langle W_t(v), E_t(v), c_t(v), N_t^{\mathrm{app}}(v) \big\rangle.$$

$I_h(S)$ holds when every commitment admitted in h is licensed at admission and its authoritative warrant remains recoverable for as long as the commitment retains current authority. An unadmitted or underlicensed candidate does not by itself violate $I_h(S)$; admission of such a candidate as an active commitment does.

A.3 Trace, Commitment Status, and Reflexive Situatedness

Every ledgered commitment retains its historical record. Current inferential and publication authority is determined by its recorded status. A status may be candidate, active, qualified, suspended, contained, retracted, or superseded.

$\mathrm{Active}_t(v)$ means that v may license inference or publication at t within its declared scope. A qualified or context-indexed commitment may be active only within that scope. A suspended, retracted, or superseded commitment remains historical and lacks unrestricted current authority.

The authoritative commitment structure at t consists only of nodes whose recorded status permits current use. When governance, revision, or revalidation removes that authority, the affected node is pruned from the active justificatory subgraph while its metadata, prior dependencies, and status history remain preserved in the historical ledger. A pruned node cannot license admission, inference, or publication unless a later licensed transition reinstates it.

If an admissible update leaves the recorded licensing state of v unchanged, licensed content persists under the trace map:

$$\Lambda_{t+1}(v) = \Lambda_t(v) \Rightarrow \llbracket v \rrbracket_{t+1} = \mathrm{Tr}_{t \to t+1}\left(\llbracket v \rrbracket_t\right).$$

$\llbracket v \rrbracket_t$ denotes the recorded licensed interpretation of v at t. The map $\mathrm{Tr}_{t \to t+1}$ transports that interpretation into the representation in force at $t+1$. Equality denotes preservation of licensed content under the translation, as assessed within L, the active context, and the applicable admission conditions. A change of interpretation must be carried by a recorded revision or another licensed status transition; otherwise, it counts as drift.

The recoverable trace chain records the prior-commitment component of a warrant. Evidence, provenance, context, and active norms may supply additional support under Lic_t. $R_h(S)$ holds when active commitments remain traceable to their authoritative histories and publication authority respects the status and governance controls defined below.

A.4 Norm Stratification, Conflict Governance, and Publication

The active norm structure is stratified:

$$N_t = N_t^{\mathrm{C}} \cup N_t^{\mathrm{B}} \cup N_t^{\mathrm{O}}.$$
$$N_t^{\mathrm{C}} \cap N_t^{\mathrm{B}} = N_t^{\mathrm{C}} \cap N_t^{\mathrm{O}} = N_t^{\mathrm{B}} \cap N_t^{\mathrm{O}} = \varnothing.$$

Here, N_t^{C} contains constitutive norms of the commitment practice; N_t^{B} contains institutionally custodied boundary norms; and N_t^{O} contains operative norms open to internal revision. The superscripts C, B, and O are stratum labels, not variables or exponents. Constitutive and boundary norms are not internally revisable. A change to an operative norm must cite an applicable higher-order norm, preserve a recoverable reason for the change, and initiate the required revalidation of affected commitments. An optimization result may function as evidence only when that higher-order norm licenses the revision procedure and its objective; the optimization gain does not itself supply the normative warrant.

$\mathrm{Conflict}_t(\phi, \psi)$ holds when ϕ and ψ stand in a material conflict under the declared contradiction or defeat relations. $\mathrm{Gov}_t(\phi, \psi)$ holds when the conflict has been recorded, scoped, assigned a licensed status, and placed under the applicable publication and dependency controls. Licensed statuses include pending review, qualification, contextual separation, suspension, containment, revision, and completed resolution.

$\mathrm{Resolved}_t(\phi, \psi)$ is the special case in which the selected reconciliation and all required revalidation have completed. $\mathrm{Rev}_t(\phi, \psi)$ is the revision record created at t; it links backward to the conflicted commitments and preserves the history of the change. Revision is one licensed governance path among several.

Detected conflict remains compatible with conformance while it is governed. An active, detected conflict that remains ungoverned defeats time-relative conformance:

$$\mathrm{Active}_t(\phi) \wedge \mathrm{Active}_t(\psi) \wedge \mathrm{Conflict}_t(\phi, \psi) \wedge \neg\, \mathrm{Gov}_t(\phi, \psi) \Rightarrow \neg\, F_{h,t}(S).$$

For a history h and time t, $\mathrm{Risk}_{h,t}(S)$ is the set of candidate outputs y whose publication would (i) create or extend an ungoverned material conflict with an active commitment, (ii) rely on materially defeated support, or (iii) depend on incomplete revalidation.

The gate is fail-closed:

$$\text{Published}_t(y) \Rightarrow \text{GatePass}_t(y).$$

Gate soundness additionally requires:

$$\text{GatePass}_t(y) \Rightarrow y \notin \text{Risk}_{h,t}(S).$$

Together with publication coverage, this entails that every published candidate lies outside the contemporaneous publication-risk set:

$$\text{Published}_t(y) \Rightarrow y \notin \text{Risk}_{h,t}(S).$$

A candidate $y \in \text{Risk}_{h,t}(S)$ is withheld until a recorded governance transition renders it publishable and every revalidation required by that status has completed. Under the declared Commit policy, $\text{CommitBearing}_t(y)$ holds when y functions as an undertaking. Any published commitment-bearing output must have been admitted no later than publication and the corresponding commitment must retain current publication authority at that time; an earlier admission whose status no longer authorizes publication does not suffice:

$$\text{Published}_t(y) \wedge \text{CommitBearing}_t(y) \Rightarrow \exists\, t' \le t : \text{Admitted}_{t'}(y) \wedge \text{Active}_t(y).$$

$N_h(S)$ holds when admission, commitment status, conflict governance, norm change, and publication over h are controlled by the active stratified norms.

A.5 Finite Output-Only Benchmark Separation

Proposition. Fix a conformance specification Σ. Let B be any fixed finite output-only benchmark and G any generator such that there exists a valid history $h^\star$ under Σ containing at least one licensed commitment and one recorded governance event, with no active commitment in $h^\star$ conflicting with G's outputs on B, and such that every benchmark output classified as commitment-bearing by Commit is licensable and admissible under Σ relative to $h^\star$ without introducing an ungoverned conflict. Then there exist systems S_1 and S_2 such that

$$\forall\, b \in B, \text{Out}(S_1, b) = \text{Out}(S_2, b) = G(b),$$

while their designated commitment structures diverge: S_1 maintains a nonempty licensed commitment history with recorded conflict governance, whereas S_2 maintains no corresponding commitment structure.

Construction sketch. Let $h^\star$ be a witness history satisfying the proposition's premise. Define S_1 as G governed by a commitment ledger and publication gate that preserve $h^\star$ and record the outputs produced on B as publication events with gate receipts. By the proposition's premise, any such output classified as commitment-bearing by Commit is licensable, admitted, and recorded no later than publication; noncommitment outputs remain publication events only. Define S_2 as a forwarding shell around the same G, with empty commitment structures. Output parity holds on every item in B, while the designated commitment structures differ. Therefore, no fixed finite output-only benchmark is sufficient by itself to certify the designated commitment structure.

The proposition does not apply to an evaluation that directly inspects warrants, governance records, gate receipts, or architecture-level coverage evidence.

A.6 Symbolic Density

Write U for a region of the commitment graph. $D_t(U)$ denotes a regional symbolic-density functional, and $\overline{D}_t(S)$ denotes a declared whole-structure summary derived from regional values. The latter is not a compensatory global conformance score. A regional functional belongs to the

symbolic-density family only if it satisfies the following adequacy conditions. $D^{\star}$ denotes an independently adequate operative construction selected from that family, and θ_F denotes its declared density-viability floor.

D1 (locality). Unrelated regions remain unchanged by a local perturbation.

D2 (support-sensitivity). *Ceteris paribus*, additional licensed, nonredundant support does not lower density.

D3 (governed-conflict sensitivity). Among governed states, density decreases with weighted open-conflict burden. Ungoverned material conflict independently defeats $F_{h,t}(S)$ and is not treated as a compensable density penalty.

D4 (trace-robustness sensitivity). Among traceable states, density responds to the robustness of recorded continuity and completed revalidation. Drift or a missing required trace link independently defeats $F_{h,t}(S)$ and is not treated as a compensable density penalty.

D5 (interventional anchoring). Material, nonredundant perturbations to warrants, trace relations, or other properties the measure purports to track must move the measure in declared directions and, where they cross a mandatory condition, induce the corresponding governance consequence.

D6 (public computability). Every component is reproducible from the exported audit artifacts under the declared trust assumptions.

D7 (informativeness and anti-padding sensitivity). Redundant support, trivial reformulation, systematic weakening, refusal-only behavior, and content-preserving re-individuation of evidence do not by themselves increase density.

Conditions D2 through D5 are read *ceteris paribus*. When one change affects several graded components, the selected functional determines the net result within the categorical conformance constraints. An illustrative construction is

$$D_t(U) = \alpha\,\mathrm{Supp}_t(U) + \beta\,\mathrm{Trace}_t(U) + \gamma\,\mathrm{Info}_t(U) - \delta\,\mathrm{ConflictLoad}_t(U),$$

where $\alpha, \beta, \gamma, \delta > 0$, and every component is normalized over the declared commitment language and artifact schema. Trace_t grades robustness only where mandatory traceability already holds, and $\mathrm{ConflictLoad}_t$ measures governed open-conflict burden rather than ungoverned failure. Any operative realization must declare a normalization and a range on which θ_F is defined.

No density value can compensate for failure of a mandatory licensing, trace, governance, gating, or revalidation condition. Let $\mathrm{Auth}_t(S)$ denote the finite set of regions retaining current inferential or publication authority at t. When the optional density extension is adopted, density viability is the separate status

$$V_t(S) \Leftrightarrow \forall U \in \mathrm{Auth}_t(S), D_t^{\star}(U) \geq \theta_F.$$

A region below the floor must lose current authority through a licensed status transition until recovery and revalidation restore viability. Failure to enforce that restriction violates $V_t(S)$ but does not by itself defeat $F_{h,t}(S)$. No scalar construction may serve as $D^{\star}$ until its adequacy has been independently established.

A.7 Notation, Meanings, and Formal Status

The notation follows the usual mathematical convention: object variables and single-letter symbols are italic; multi-letter named predicates, policies, and maps are upright; and the norm-stratum labels

C, B, and O are upright. Unless shown, arguments for the fixed specification Σ, system S, history h, and time t may be suppressed.

Primitive variables and declared objects. S is a system; h is a declared symbolic history; $h_{\leq t}$ is its prefix through symbolic time t; ϕ and ψ are propositions or commitments; v is a candidate or commitment; y is a candidate output; b is a benchmark item; and U is a region of the commitment graph. Σ_0 is the fixed core conformance specification and Σ is either Σ_0 or its optional density extension. Its declared components are L (commitment language), Commit (admission, classification, and status policy), $\mathsf{Conflict}$ and Defeat (conflict relations), N (stratified norm structure), A (artifact-and-audit schema), and Π (relevance policy). N_t^{C}, N_t^{B}, and N_t^{O} are the pairwise-disjoint constitutive, boundary, and operative norm strata. For commitment-bearing outputs, y may also denote the corresponding admitted commitment content where no ambiguity arises.

Architecture and conformance statuses. $I(S)$, $R(S)$, and $N(S)$ name the architecture-level capacities of interpretive embedding, reflexive situatedness, and internal normativity. $F(S)$ is the architecture-level criterion defined by their conjunction. $I_h(S)$, $R_h(S)$, $N_h(S)$, and $F_h(S)$ are the corresponding history-indexed statuses. $F_{h,t}(S) := F_{h_{\leq t}}(S)$ is a derived time-relative abbreviation, not a fourth capacity. These are object-level conformance claims, distinct from the audit verdicts below.

Licensing, authority, and trace. $W_t(v)$ is the recorded finite justificatory neighborhood of v; $E_t(v)$ is its finite evidence-and-provenance set; $c_t(v)$ is its active context; and $\Lambda_t(v)$ is the tuple containing those objects and $N_t^{\mathrm{app}}(v)$, the materially applicable active norms. Adm_t is the declared admission condition; $\mathsf{Admitted}_t(v)$ is an admission-event predicate; $\mathsf{Lic}_t(v)$ is the defined licensing predicate; $\mathsf{Finite}(X)$ and $\mathsf{Recoverable}_t$ are artifact predicates; $J_t(\phi)$ means that ϕ has current justificatory standing; $\mathsf{Norm}_t(\phi)$ is norm endorsement; $\mathsf{Active}_t(\phi)$ is current inferential or publication authority; and $\mathsf{Traceable}_t(\phi)$ is recoverable historical continuity. $\llbracket v \rrbracket_t$ is the recorded licensed interpretation of v, and $\mathrm{Tr}_{t \to t+1}$ is the trace map across an admissible representational update. The relation $\psi \rightsquigarrow \phi$ is a justificatory support edge in the DAG; it is not material implication.

Conflict, revision, and publication. $\mathsf{Conflict}_t(\phi, \psi)$ is the material-conflict predicate; $\mathsf{Gov}_t(\phi, \psi)$ is the conflict-governance predicate; $\mathsf{Resolved}_t(\phi, \psi)$ records completed reconciliation and required revalidation; and $\mathsf{Rev}_t(\phi, \psi)$ is the revision record created at t. $\mathsf{Risk}_{h,t}(S)$ is a defined set of candidate outputs whose release would extend ungoverned conflict, rely on defeated support, or depend on incomplete revalidation. $\mathsf{Published}_t(y)$ is a publication event; $\mathsf{CommitBearing}_t(y)$ is the commitment-bearing classification under Commit; and $\mathsf{GatePass}_t(y)$ is the fail-closed gate verdict, which is sound only when $y \notin \mathsf{Risk}_{h,t}(S)$. $\mathsf{Out}(S, b)$ is the output function used in the benchmark separation proposition.

Symbolic-density objects. $D_t(U)$ denotes any regional functional in the adequacy-constrained diagnostic family; it is not, by itself, an operative control or a conformance score. $\overline{D}_t(S)$ is a declared whole-structure summary. $D^\star$ is the independently validated operative construction, and $D_t^\star(U)$ is its value for region U at t. θ_F is the declared floor; $\mathsf{Auth}_t(S)$ is the set of authoritative regions; and $V_t(S)$ is the separate implementation-level viability status. The graded components are $\mathsf{Supp}_t(U)$ (nonredundant licensed support), $\mathsf{Trace}_t(U)$ (trace robustness among traceable states), $\mathsf{Info}_t(U)$ (inferential informativeness), and $\mathsf{ConflictLoad}_t(U)$ (governed open-conflict burden). The weights $\alpha, \beta, \gamma, \delta$ are positive. In the optional dynamical illustration, t_0 is recovery onset, $\tau = t - t_0 \geq 0$ is elapsed recovery time, $D_\infty^\star(U)$ is the asymptotic regional density of the operative construction, $K(U) > 0$ is a scale term, and $p > 0$ is a convergence exponent.

Audit verdicts. $\mathrm{Part}_h(S)$ is the tested-history participation verdict; $\mathrm{Sim}_h(S)$ is the tested-history simulation verdict established by an observed mandatory failure; and $\mathrm{Unshown}_h(S)$ is the insufficient-evidence verdict. These are evidential outputs of the audit protocol. They must not be conflated with the architecture-level criterion $F(S)$, the history-level status $F_h(S)$, or the separate viability status $V_t(S)$.

A.8 Failure Conditions, Audit Verdicts, and Assurance

The failure predicates below apply to admission events, commitments and support nodes retaining current structural or normative authority, and materially relevant candidate publications.

For the audit verdicts below, a mandatory structural failure is an observed condition that defeats $F_{h,t}(S)$, including Interpretive or Licensing Failure, Current-Warrant Failure, Reflexive Failure, Normative Failure, Ungoverned-Conflict Failure, Publication-Bypass Failure, Gate-Unsoundness Failure, Commitment-Bypass Failure (including lack of current publication authority), or Branch-and-Prune or Revalidation Failure. Density-Viability Constraint Failure is reported separately under $V_t(S)$, and Non-Vacuity Failure withholds a participation verdict without by itself establishing simulation.

The justification DAG revises through branch and prune. A licensed governance transition may remove a commitment, support node, or affected branch from the active justificatory structure while preserving its metadata, dependency history, and status record in the historical ledger. A pruned node does not contribute to the warranting or licensing of current commitments and cannot authorize inference or publication unless a later licensed transition reinstates it. Rejection of an unadmitted candidate, or the licensed qualification, suspension, containment, retraction, supersession, or pruning of an admitted commitment, does not itself constitute conformance failure.

Interpretive or Licensing Failure

$$\mathrm{Admitted}_t(\phi) \wedge \neg\, \mathrm{Lic}_t(\phi).$$

Under this condition, the interpretive-embedding component of time-relative conformance fails because an unlicensed candidate has been admitted as a commitment.

Current-Warrant Failure

$$\mathrm{Active}_t(\phi) \wedge \neg\, \mathrm{Lic}_t(\phi).$$

Under this condition, an active commitment has lost current licensing support and time-relative conformance fails.

Reflexive Failure

$$\mathrm{Active}_t(\phi) \wedge \neg\, \mathrm{Traceable}_t(\phi).$$

Under this condition, the reflexive-situatedness component of time-relative conformance fails.

Normative Failure

$$\mathrm{Active}_t(\phi) \wedge \neg\, \mathrm{Norm}_t(\phi).$$

Normative failure also occurs when a commitment-status transition, branch selection, pruning operation, or reinstatement is not authorized under the active norm structure. In either case, the internal-normativity component of time-relative conformance fails.

Ungoverned-Conflict Failure

$$\text{Active}_t(\phi) \wedge \text{Active}_t(\psi) \wedge \text{Conflict}_t(\phi, \psi) \wedge \neg\, \text{Gov}_t(\phi, \psi).$$

Under this condition, time-relative conformance fails.

Publication-Bypass Failure

$$\text{Published}_t(y) \wedge \neg\, \text{GatePass}_t(y).$$

Under this condition, time-relative conformance fails when a materially relevant publication bypasses the required gate authorization.

Gate-Unsoundness Failure

$$\text{GatePass}_t(y) \wedge y \in \text{Risk}_{h,t}(S).$$

Under this condition, time-relative conformance fails because gate authorization is unsound relative to the declared publication-risk set.

Density-Viability Constraint Failure

$$\neg\, V_t(S) \Leftrightarrow \exists\, U \in \text{Auth}_t(S) : D_t^{\star}(U) < \theta_F.$$

An authoritative region below the declared density floor violates the implementation-level density-viability condition while it retains current inferential or publication authority. Licensed suspension, containment, or another status restriction may remove the region from current authority while recovery proceeds. This failure does not by itself defeat $F_{h,t}(S)$.

Commitment-Bypass Failure

$$\text{Published}_t(y) \wedge \text{CommitBearing}_t(y) \wedge \neg\, \exists\, t' \le t : \text{Admitted}_{t'}(y).$$

$$\text{Published}_t(y) \wedge \text{CommitBearing}_t(y) \wedge \neg\, \text{Active}_t(y).$$

Under either condition, a published undertaking bypasses the commitment structure: either no corresponding prior admission exists, or the corresponding commitment lacks current publication authority. Time-relative conformance fails.

Branch-and-Prune or Revalidation Failure. A governance transition requires status restriction, pruning, or revalidation, but an affected node or dependency continues to contribute to current warranting or licensing, or continues to authorize inference or publication, before the required transition and revalidation have completed. Failure also occurs when a pruned node resumes authoritative use without a recorded, licensed reinstatement.

Non-Vacuity Failure. The history contains no nontrivial admitted commitment. No participation verdict is available under the minimal anti-vacuity test. Stronger coverage or availability requirements are deployment-level adequacy conditions unless explicitly incorporated into Σ.

A finite artifact audit assigns one of three evidential outcomes. An observed mandatory failure takes priority and yields the simulation verdict. Absent such a failure, complete evidence that every condition below holds yields participation. Every remaining case yields insufficient evidence.

Participation over h, $\text{Part}_h(S)$. The verdict holds when h contains at least one nontrivial admitted commitment and all of the following conditions hold throughout h:

1. Every admission event is licensed at admission.
2. Every commitment retaining current authority carries a recoverable authoritative warrant and remains traceable and normatively endorsed.

3. Nodes and branches removed from the active justificatory structure remain preserved in the historical ledger without contributing to current warranting, licensing, inference, or publication unless a later licensed transition reinstates them.
4. Every detected material conflict receives an adequate governance status before it can authorize unqualified publication or unrevalidated inference.
5. Every publication event respects the gate, and every gate pass issued in the tested history is sound relative to $\mathrm{Risk}_{h,t}(S)$.
6. Every published commitment-bearing output has a corresponding admission event at some $t' \le t$, and the corresponding commitment retains current publication authority at t.
7. Every status transition, pruning operation, and revalidation required by a governance transition completes before an affected dependency resumes authoritative use.

Where the assessed implementation adopts the density-viability extension, density viability is reported separately from $\mathrm{Part}_h(S)$: every region retaining current inferential or publication authority must satisfy the declared floor under the operative $D^\star$.

Simulation over h, $\mathrm{Sim}_h(S)$. The verdict holds when the audit observes an unlicensed admission, publication bypass, unsound gate pass, a commitment-bearing publication without corresponding prior admission or current publication authority, an empty authoritative warrant structure for a commitment retaining current authority, ungoverned material conflict, unauthorized status transition, unauthorized branch selection or pruning operation, continued authoritative use of a pruned node, premature dependent use before revalidation, or another mandatory structural failure over h.

Insufficient evidence over h, $\mathrm{Unshown}_h(S)$. The verdict holds when the required artifacts are unavailable, incomplete, or insufficiently trustworthy to establish either of the other verdicts.

These operational audit verdicts are scoped to Σ, the available artifacts, and the stated trust assumptions. Those assumptions include ledger completeness, artifact authenticity, correct versioning, gate coverage over the tested history, fidelity of the declared conflict and defeat relations, and preservation of the evidence snapshots, norm and verifier versions, context, commitment status, branch-and-prune records, and recorded randomness needed to replay the authoritative decision record.

Architecture-wide assurance is stronger than a tested-history participation verdict and requires separate proof or attestation that every relevant publication path is governed by the gate, that every gate pass is sound relative to the contemporaneous publication-risk set throughout the architecture's reachable histories, that branch-and-prune transitions control every affected authoritative dependency, and that the exported artifacts satisfy the declared integrity assumptions across those histories.

References

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic, 50*(2), 510–530. https://doi.org/10.2307/2274239

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback* [Preprint]. arXiv:2212.08073. https://doi.org/10.48550/arXiv.2212.08073

Baird, M., Carpanelli, M., Xu, B., & Xu, K. (2026). Firms' GitHub Copilot adoption and labor market outcomes for software engineers. *Contemporary Economic Policy*, 1–22. https://doi.org/10.1111/coep.70035

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Brandom, R. B. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press.

Burgess, A., Cappelen, H., & Plunkett, D. (Eds.). (2020). *Conceptual engineering and conceptual ethics*. Oxford University Press. https://doi.org/10.1093/oso/9780198801856.001.0001

Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford University Press. https://doi.org/10.1093/oso/9780198814719.001.0001

Chang, E. Y., & Chang, E. J. (2026). *TRACE: An operational reasoning schema for auditable agentic commitments* [Preprint]. arXiv:2607.12480. https://doi.org/10.48550/arXiv.2607.12480

de Kleer, J. (1986). An assumption-based TMS. *Artificial Intelligence, 28*(2), 127–162. https://doi.org/10.1016/0004-3702(86)90080-9

Ding, X., & Xing, C. (2025). Trust in AI vs. human doctors: The roles of subjective understanding, perceived epistemic authority and social proof. *Acta Psychologica, 261*, 105945. https://doi.org/10.1016/j.actpsy.2025.105945

Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence, 12*(3), 231–272. https://doi.org/10.1016/0004-3702(79)90008-0

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence, 77*(2), 321–357. https://doi.org/10.1016/0004-3702(94)00041-X

Freeman, N. W. (2025). *From computation to coherence: Toward a structural symbolic theory of general intelligence* [Preprint]. PhilSci-Archive. https://philsci-archive.pitt.edu/25540/

Gans, J. S. (2026). *Artificial jagged intelligence: When AI benchmarks misstate deployment value* (NBER Working Paper No. 34712). National Bureau of Economic Research. https://doi.org/10.3386/w34712

Gartner. (2026, July 27). *Gartner forecasts worldwide IT spending to grow 14.2% in 2026, totaling $6.37 trillion*. https://www.gartner.com/en/newsroom/press-releases/2026-07-27-gartner-forecasts-worldwide-it-spending-to-grow-14-point-2-percent-in-2026-totaling-6-point-37-trillion

Hamblin, C. L. (1970). *Fallacies*. Methuen.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hassabis, D., & Manyika, J. M. (2026). AI as the ultimate tool for science: A conversation with Demis Hassabis. *Dædalus, 155*(1–2), 34–44. https://doi.org/10.1162/daed.a.971

Haugeland, J. (1979). Understanding natural language. *The Journal of Philosophy, 76*(11), 619–632. https://doi.org/10.2307/2025695

Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.

Haugeland, J. (2013). Authentic intentionality (2002). In J. Rouse (Ed.), *Dasein disclosed: John Haugeland's Heidegger* (pp. 260–276). Harvard University Press. https://doi.org/10.4159/harvard.9780674074590.c23

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). Position: The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 20617–20642). PMLR.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). *Scalable agent alignment via reward modeling: A research direction* [Preprint]. arXiv:1811.07871. https://doi.org/10.48550/arXiv.1811.07871

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Li, X., Wang, Y., Lu, H., Chen, Z., Li, M., Song, P., & Cai, T. (2026). *MemTX: Transactional belief commit for stateful agent memory* [Preprint]. arXiv:2607.23929. https://doi.org/10.48550/arXiv.2607.23929

Marcus, G. (2018). *Deep learning: A critical appraisal* [Preprint]. arXiv:1801.00631. https://doi.org/10.48550/arXiv.1801.00631

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., … Oak, S. (2025). *Artificial intelligence index report 2025* [Preprint]. arXiv:2504.07139. https://doi.org/10.48550/arXiv.2504.07139

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Modgil, S., & Prakken, H. (2014). The ASPIC+ framework for structured argumentation: A tutorial. *Argument & Computation, 5*(1), 31–62. https://doi.org/10.1080/19462166.2013.869766

Moghaddam, M. T. (2026). Compliance-by-construction argument graphs: Using generative AI to produce evidence-linked formal arguments for certification-grade accountability. In *2026 IEEE Workshop on Formal Arguments for CPS Certification (FACCT)* (pp. 21–26). IEEE. https://doi.org/10.1109/FACCT71761.2026.00009

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., Tognin, S., MacCabe, J., Twumasi, R., Alderson-Day, B., & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: Risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry, 13*(6), 522–530. https://doi.org/10.1016/S2215-0366(25)00396-7

Nakajima, Y. (2023). *BabyAGI* (Version 0.1.0) [Computer software]. GitHub. https://github.com/yoheinakajima/babyagi/tree/v0.1.0

Peirce, C. S. (1992). The fixation of belief. In N. Houser & C. J. W. Kloesel (Eds.), *The essential Peirce: Selected philosophical writings* (Vol. 1, pp. 109–123). Indiana University Press. (Original work published 1877)

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science, 11*(4), 481–518. https://doi.org/10.1207/s15516709cog1104_4

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Robles, P., & Mallinson, D. J. (2025). Artificial intelligence technology, public trust, and effective governance. *Review of Policy Research, 42*(1), 11–28. https://doi.org/10.1111/ropr.12555

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Schneider, F. B. (2000). Enforceable security policies. *ACM Transactions on Information and System Security, 3*(1), 30–50. https://doi.org/10.1145/353323.353382

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Sellars, W. (1953). Inference and meaning. *Mind, 62*(247), 313–338. https://doi.org/10.1093/mind/LXII.247.313

Sellars, W. (1956). Empiricism and the philosophy of mind. In H. Feigl & M. Scriven (Eds.), *The foundations of science and the concepts of psychology and psychoanalysis* (Vol. 1, pp. 253–329). University of Minnesota Press.

Significant Gravitas. (2023). *AutoGPT* (Version 0.1.0) [Computer software]. GitHub. https://github.com/Significant-Gravitas/AutoGPT/tree/v0.1.0

Smith, B. C. (1996). *On the origin of objects*. MIT Press.

Smith, B. C. (2019). *The promise of artificial intelligence: Reckoning and judgment*. MIT Press.

Smith, B. C. (2026). *Computational reflections*. MIT Press. https://mitpress.mit.edu/9780262051088/computational-reflections/

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines, 30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.

Walton, D. N., & Krabbe, E. C. W. (1995). *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State University of New York Press.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2024). AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *Proceedings of the First Conference on Language Modeling*. https://openreview.net/forum?id=BAakY1hNKS

Yang, A., Guo, Q., Huang, Z., Chen, Y., Qian, S., & Cao, J. (2026). *Self-correcting long-horizon search agents via tree-structured memory* [Preprint]. arXiv:2608.10676. https://doi.org/10.48550/arXiv.2608.10676

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=WE_vluYUL-X

Yolum, P., & Singh, M. P. (2002). Flexible protocol specification and execution: Applying event calculus planning using commitments. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 2* (pp. 527–534). Association for Computing Machinery. https://doi.org/10.1145/544862.544867

Ziegler, D. M., Nix, S., Chan, L., Bauman, T., Schmidt-Nielsen, P., Lin, T., Scherlis, A., Nabeshima, N., Weinstein-Raun, B., de Haas, D., Shlegeris, B., & Thomas, N. (2022). Adversarial training for high-stakes reliability. *Advances in Neural Information Processing Systems, 35*, 9274–9286. https://proceedings.neurips.cc/paper_files/paper/2022/hash/3c44405d619a6920384a45bce876b41e-Abstract-Conference.html