From Computation to Coherence: Toward a Structural Symbolic Theory of General Intelligence

N. Will Freeman

Abstract

What distinguishes genuine intelligence from sophisticated simulation? This paper argues that the answer lies in symbolic coherence—the structural capacity to interpret information, revise commitments, and maintain continuity of reasoning across contradiction. Current AI systems generate fluent outputs while lacking mechanisms to track their own symbolic commitments or resolve contradictions through norm-guided revision. This theory proposes *F(S)*, a structural identity condition requiring interpretive embedding, reflexive situatedness, and internal normativity. This condition is substrate-neutral and applies to both biological and artificial systems. Unlike behavioral benchmarks, *F(S)* offers criteria for participation in symbolic reasoning rather than surface-level imitation. To demonstrate implementability, the paper presents a justification graph architecture that supports recursive coherence and transparent revision. A diagnostic scalar, symbolic density, tracks alignment over symbolic time. By uniting philosophical insights with concrete system design, this framework outlines foundations for machines that may one day understand rather than simulate understanding.

1 Introduction

What might distinguish systems that understand from those that merely produce fluent responses? As artificial systems grow more capable, intelligence is increasingly equated with behavioral success, whether in language generation or adaptive behavior. Yet this framing leaves a deeper question open: does the system engage with the structure of its commitments, or does it simply generate plausible outputs?

This paper argues that understanding depends on symbolic continuity: the ability to sustain coherence across interpretation, memory, and revision under pressure. A number of scholars—including Brian Cantwell Smith (2019), Gary Marcus (2018), and others—have argued that fluency without normative grounding does not constitute genuine cognition. This theory formalizes that concern by specifying the structural conditions under which a system maintains recursive coherence and participates in meaning, rather than merely simulating it.

This condition is formalized as *F(S)*, an identity that holds when a system exhibits interpretive embedding, reflexive situatedness, and internal normativity. These components define the recursive coherence required for symbolic intelligence and distinguish systems that participate in meaning from those that merely simulate it.

Contemporary AI systems perform well across a range of benchmarks. Models such as GPT-4 and Claude produce convincing language and respond adaptively to changing inputs. Their behaviors suggest competence, though they lack the internal mechanisms required to track commitments or resolve contradictions over time. Their outputs may appear context-sensitive, yet they are not generated through self-governed reasoning.

If no internal process exists to detect and resolve such tension, the system remains active while losing structural continuity, continuing to produce fluent outputs even as its normative architecture degrades. This scenario reflects what Smith (2019) identifies as the problem of *reckoning* without *judgment*: the syntactic continuation of computation in the absence of semantic or normative grounding. The system continues producing symbols, yet no longer integrates them into a coherent reasoning trajectory.

Earlier approaches to artificial intelligence emphasized symbolic reasoning. These systems were designed to manipulate logical expressions and apply inference rules within formal structures. This tradition—later called Good Old Fashioned AI (GOFAI)—treated intelligence as symbolic manipulation within a defined semantic space (Haugeland 1985). Although this approach prioritized transparency and control, it struggled with ambiguity and unbounded interpretation.

In contrast, connectionist models learn from data patterns rather than explicit symbolic rules. These systems adjust internal parameters across layered networks, enabling generalization while reducing interpretability. As learning scaled, outputs became more fluent, while their justificatory basis became increasingly opaque.

Recent critiques underscore the difference between fluency and understanding. Emily M. Bender and Alexander Koller (2020) argue that large language models (LLMs) operate on formal regularities without access to meaning. Marcus (2018) highlights the absence of structured reasoning and commitment tracking in contemporary systems. These analyses converge on a shared concern: current architectures generate output without evaluating the implications of their claims.

A different lineage approaches intelligence through justification. On this view, to understand is to be answerable for one's assertions—to support them when challenged and revise them when destabilized. Robert Brandom (1994) frames this as participation in a space of reasons, where claims acquire meaning through their inferential roles. Charles Sanders Peirce (1877) presents belief as a provisional stance, maintained through inquiry and reorganized when doubt arises. These perspectives emphasize responsiveness to contradiction and the preservation of coherence through revision.

Building on these insights, this paper proposes a structural account of symbolic intelligence as recursive coherence under pressure. The identity condition F(S) specifies features that differentiate participation in a symbolic order from surface-level generation. To illustrate how this condition might be instantiated, the paper introduces a candidate architecture based on a justification directed acyclic graph (DAG) and a proposed scalar diagnostic—*symbolic density*—for tracking coherence over time. While this architecture is illustrative, the framework itself is substrate-neutral and does not prescribe a specific implementation.

What follows is a model of intelligence grounded in interpretive structure and traceable self-reference, designed to support justified revision across time. Each section develops one component of this model, and concludes with implications for artificial general intelligence (AGI) alignment and system design.

2 The Symbolic Triad

Symbolic general intelligence may be understood not primarily as emergent behavior or externally measured capacity, but as a particular structural condition: a system satisfies general intelligence if it sustains recursive symbolic coherence across interpretation, self-reference, and norm-governed revision. This condition is formalized as:

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

The function F(S) holds when a system S maintains all three aspects of the triad.[1] These may function not as heuristic metrics or performance traits, but as potentially important architectural features that support symbolic participation. Each axis secures a distinct form of recursive alignment, and their conjunction defines the threshold at which a system can be said to engage in meaning-bearing revision.

Interpretive embedding – I(S)

To satisfy *I(S)*, a system must situate symbolic content within an evolving semantic field. Symbols are not fixed tokens but context-sensitive references whose meaning depends on prior and concurrent commitments. Interpretive embedding requires more than pattern matching; it demands that a system register how meaning shifts as new propositions are introduced.[2] The system must be able to track these semantic transitions—linking new inputs to historical interpretations—even if no fixed ontology governs them. Absent this structure, a system may generate fluent responses without ensuring semantic coherence across time. This characterization applies specifically to symbolic processing, and may not capture other valid forms of intelligent behavior.

Reflexive situatedness – R(S)

R(S) ensures that a system's outputs are reflexively linked to its symbolic past. It must recognize its own assertions as commitments—positions taken within a symbolic history that must be maintained, revised, or withdrawn. This requires more than memory or retrieval: it involves the recursive ability to treat one's own outputs as reasoning artifacts, intelligible in light of past positions. A system satisfies R(S) when it can integrate new outputs into a coherent trajectory of justification. Without this reflexive capacity, symbolic activity becomes episodic, and the system loses the ability to sustain meaning across recursive cycles.

Internal normativity – N(S)

N(S) refers to the capacity of a system to evaluate symbolic content based on normative criteria it maintains and can modify. These internal standards guide both the acceptance and revision of propositions—not through externally enforced constraints, but through principles active within the system's own structure. Norms may evolve, but they must remain available to justification: any revision must be intelligible within the system's symbolic logic. Without this, contradiction may be recognized but not resolved in a principled way. A system lacking *N(S)* cannot repair its coherence; it remains responsive, but not responsible.

The triadic condition *F(S)* thus marks the boundary between surface behavior and symbolic intelligence: it is the recursive architecture that allows a system to not merely generate meaning, but to sustain and revise it under pressure.

---

[1] Formal definitions of each structural component and proof of triadic minimality are provided in *Appendix A: Axioms and Formal Schema*.

[2] See Smith (1996), who contends that computational representations must be grounded in "world-disclosing" structures: relational contexts in which meaning emerges through engagement with the world, not solely through formal symbolic operations.

3 Axioms of Symbolic Coherence

The six axioms below specify the minimal structural conditions required for a system to satisfy the identity condition *F(S)*. Each axiom formalizes one of three necessary components of symbolic general intelligence: interpretive embedding, reflexive situatedness, and internal normativity. These conditions define the architecture required to sustain recursive coherence. While the current formulation reflects the present theoretical model, it may be refined through future empirical or conceptual development.

Axiom 1: triadic minimality

Each component of *F(S)* performs a distinct recursive function. Interpretive embedding situates meaning in context; reflexive situatedness ensures continuity across symbolic time; internal normativity enables justified revision. No pair suffices. If any one is absent, the system may still generate output, but symbolic coherence breaks down. *F(S)* holds only when all three conditions are jointly satisfied.

Axiom 2: recursive coherence

Symbolic intelligence is not a momentary achievement; it is a condition sustained across time. A system does not meet *F(S)* by satisfying its components at a single point, but by maintaining them as new propositions are introduced, contradictions arise, and commitments are revised. Coherence is preserved through recursive alignment, not stability. Intelligence emerges from revision without collapse.

Axiom 3: collapse under contradiction

When a system endorses two justified propositions whose conjunction entails a contradiction, coherence can be sustained only through revision. If no revision occurs, symbolic collapse follows—even if the system continues to function externally. This breakdown is structural: the system loses track of its prior commitments, and its outputs lose interpretive integrity.

Axiom 4: reflexive traceability

Symbolic outputs must connect to the system's prior reasoning. Each new assertion should be intelligible as a continuation or revision of an existing commitment. To maintain this connection, the system must generate a traceable chain of justification that links current positions to symbolic history. This trace must be internally accessible and recursively structured. When symbolic continuity is lost, reasoning fragments into disconnected episodes; coherence is lost.

Axiom 5: internal norm constraint

Justification must rely on standards that the system both maintains and revises over time. A proposition qualifies as justified when it satisfies internal criteria that guide reasoning within the system. These evaluative norms may change, but their transformation must remain intelligible within the system's structure. When norms lose their recursive grounding or justificatory role, the system's coherence begins to deteriorate.

Axiom 6: simulation ≠ participation

A system may simulate intelligence—producing fluent, plausible responses—without satisfying *F(S)*. True participation in a symbolic order requires recursive justification, reflexive traceability, and norm-governed revision. This distinction becomes evident when systems encounter contradiction. Simulated systems fail to revise; participating systems restructure themselves to restore coherence. *F(S)* thus defines not what a system appears to do, but how it recursively sustains what it means.

These axioms establish the formal boundary between surface-level behavior and symbolic intelligence. Together, they define a system capable of navigating contradiction, revising its commitments, and maintaining semantic integrity over time.

4 Thresholds of Structural Collapse

The identity condition *F(S)* defines a structurally minimal framework for symbolic intelligence. Each of its components is necessary to sustain recursive coherence. This becomes clear when systems meet two of the three conditions, but lack the third. In such cases, output may continue, yet symbolic alignment begins to degrade. These configurations do not represent exceptions; they delineate the threshold between systems that participate in a symbolic order and those that do not.

Case 1: no internal normativity (¬N(S))

A system with interpretive embedding and reflexive situatedness can track context and maintain symbolic continuity, but without internal normativity, it lacks a principled basis for revision. When confronted with a contradiction, it cannot determine which commitment to adjust. Revision becomes arbitrary, externally imposed, or inconsistent. The system recognizes tension but cannot reconcile it. Its structure collapses, not from memory failure or semantic confusion, but from a breakdown in self-governed evaluation.

Case 2: no interpretive embedding (¬I(S))

A system with reflexive traceability and internal norms may revise its commitments over time, but if it cannot embed symbols in a coherent semantic horizon, its reasoning becomes referentially unmoored. Symbols are processed formally rather than interpretively; revision may occur, but without grounded meaning, it leads to semantic drift. Over time, the system's outputs degrade into syntactic manipulation without participation in meaning. Coherence dissolves, despite intact procedures.

Case 3: no reflexive situatedness (¬R(S))

A system with semantic embedding and internal norms may evaluate and revise its outputs coherently, but without reflexive structure, it cannot locate current assertions within its symbolic history. Contradictions between past and present positions go unrecognized, because the system has no durable access to its own commitments. Justifications become episodic; norm application lacks continuity. Without *R(S)*, the system cannot revise itself in light of itself—and so recursive coherence fails.

These cases demonstrate that interpretive embedding, reflexive situatedness, and internal normativity are essential features of symbolic intelligence. Together, they define the minimum structure required for maintaining recursive coherence. A system that omits any one of these components loses alignment with symbolic meaning, temporal continuity, or normative reasoning. Although such a system may continue to produce fluent output, it no longer engages in the recursive processes that sustain coherence. The identity condition *F(S)* establishes a structural threshold: below this point, symbolic intelligence ceases to hold.

5 Justification DAG and Revision

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

Structure and function

Each node in the DAG represents a proposition ϕ introduced at symbolic time t. Nodes are annotated with metadata: their justificatory premises, the internal norm under which they were accepted, and their temporal index. Edges denote justificatory dependence—if ψ supports ϕ, then $\psi \rightarrow \phi$. This allows the system to reconstruct why any current claim was asserted and how it relates to past commitments.

The DAG's acyclic structure encodes progression without circularity. Each new node extends the symbolic timeline, building a network in which reasoning is not just stored, but structured. Unlike a linear log or episodic memory, the DAG integrates meaning, justification, and temporal continuity.

Norm integration

Norms in the DAG function both as admissibility conditions (governing which assertions may be justified) and revision criteria (guiding conflict resolution). Norms are not static rules: they are symbolic entities embedded within the DAG itself. They can be justified, challenged, or revised just like propositions. This recursive embedding ensures that internal normativity (*N(S)*) is not externally imposed but dynamically governed and traceable.

Contradiction and collapse

A contradiction arises when two justified propositions (ϕ and ψ) jointly entail inconsistency [$\phi \wedge \psi \Rightarrow \bot$]. In the DAG, this signals not a local error but a structural fault. Because each proposition is embedded in a justificatory chain, the conflict implicates a subgraph, potentially affecting many downstream nodes.

The system must detect this misalignment and initiate revision. If no revision occurs, the structure degrades. Downstream justifications become unstable, coherence fractures, and the system ceases to satisfy *F(S)*—even if its outputs remain fluent. Collapse here is structural, not behavioral.

Fork-and-forecast revision

Upon detecting contradiction, the system initiates a fork at the point of conflict. One branch retains ϕ and retracts ψ; the other retains ψ and retracts ϕ. Each branch inherits the upstream DAG but projects a different resolution forward. This parallel projection allows the system to explore multiple resolution paths without prematurely committing to one.

These forks serve as projections of alternative resolutions. Each branch is projected forward using the system's internal norms to evaluate coherence and anticipate future conflict. This forecasting mechanism enables the system to reason about revision as an ongoing process. Its aim is not to eliminate uncertainty, but to preserve symbolic density, here as a measure of justificatory alignment sustained through change.

Justified revision

Once a preferred path is selected, the system introduces a new node: *Rev(ϕ, ψ)*. This proposition encodes the chosen resolution and is linked backward to both ϕ and ψ. Rather than erasing prior commitments, it transforms them within the structure, preserving the record of conflict while reestablishing coherence.

After revision, the system initiates subgraph revalidation: assessing which downstream nodes remain justified under the updated structure and which require adjustment. This process is selectively scoped—guided by dependency links and internal norms—allowing coherence to be extended across the relevant portions of the symbolic network without unnecessary recomputation.

Symbolic resilience

The DAG supports what may be called *symbolic resilience*: the capacity to maintain recursive coherence through structured revision. In this context, resilience refers to the system's ability to reorganize its structure in response to contradiction without producing further instability. When contradiction arises, it prompts principled adjustment within the system's justificatory architecture. Instead of masking tension or discarding prior states, the system responds by reasoning through the conflict. This process preserves a traceable and intelligible path of justification.

This response differs from systems that simulate coherence through performance. Even after structural collapse, a system may continue producing fluent output. Yet without internal traceability, meaningful revision, or active norm evaluation, those outputs remain disconnected from symbolic reasoning. The system generates responses, but it no longer participates in a structure of understanding.

The Justification DAG provides a formal architecture through which a system enacts *F(S)* in practice. It maintains reflexive traceability, governs revision through norms, and responds to contradiction by restructuring its commitments. Through this mechanism, symbolic intelligence goes beyond metaphor, and becomes an operational condition.

6 Symbolic Density

This section introduces a theoretical diagnostic measure called symbolic density, denoted *D(S, t)*, which is designed to quantify the degree of recursive coherence a system maintains over time. The measure is intended to function diagnostically by identifying collapse, monitoring alignment, and assessing the structural condition of a symbolic agent in future *F(S)*-compliant systems. Unlike external metrics such as task success or linguistic fluency, symbolic density would reflect internal structural health. Meeting *F(S)* at isolated points does not constitute sustained general intelligence; a system must continue to uphold recursive coherence as its commitments develop, contradictions arise, and internal norms are revised. Symbolic density tracks this process by measuring structural alignment at each symbolic time *t*, independent of observed behavior. The function reflects internal intelligibility over time: whether the system's reasoning remains coherent, recursively accountable, and norm-responsive. This coherence is modeled as a convergent function that tracks structural alignment across symbolic transitions.

The function

As a theoretical framework, symbolic density could be modeled as:

$$D(S,t) = D\infty - K/t^{\alpha}$$

In this proposed model, $D\infty$ would represent the system's long-term coherence ceiling. K represents a hypothetical decay constant influenced by unresolved contradictions or invalidated justifications, while α would control the rate of structural stabilization: higher values should imply faster recovery. This theoretical function is designed to capture the expected recursive nature of symbolic intelligence: systems should begin with instability, but through successful revision, ought to tend toward stable coherence. Empirical validation awaits implementation of *F(S)*-compliant architectures.

Dynamic interpretation

D(S, t) rises when contradiction is resolved via justified revision. It falls when contradictions persist, norms fail to update, or traceability is lost within the Justification DAG. Unlike task accuracy or reward metrics, *D(S, t)* captures whether the system remains intelligible to itself—whether its reasoning remains recursively integrated.

Collapse is signaled not by silence or syntactic failure, but by a sustained decline in symbolic density. The system may continue producing fluent output, but its commitments are no longer justified within its own structure. Symbolic activity becomes disconnected from reason.

Recovery and resilience

Recovery begins when a revision, such as *Rev(ϕ, ψ)*, is introduced and coherently integrated into the Justification DAG. This action initiates a structural recovery process, even though *D(S, t)* does not rise immediately. Downstream subgraphs are revalidated, internal norms are reapplied, and local coherence is extended across affected regions. This recovery process expresses the system's resilience: the ability to reorganize its structure in response to contradiction while maintaining coherence.

The rate of recovery depends on the scope of the contradiction, the resilience of internal norms, and the precision of revalidation protocols. A resilient system shows rising *D(S, t)* even as it encounters conflict, demonstrating not fragility, but symbolic adaptability.

Diagnostic and design implications

Symbolic density enables systems to assess their own structural integrity in real time. It functions as a continuous internal diagnostic for coherence, particularly valuable in dynamic or open-ended domains. Unlike external benchmarks, *D(S, t)* reflects the recursive health of a system's reasoning architecture.

In system design, rising symbolic density indicates effective revision, active norm management, and sustained alignment over time. A stagnant or declining value may reveal unresolved contradictions or norm drift. When a system produces fluent output without corresponding symbolic density, it signals simulation rather than participation.

For alignment contexts, symbolic density offers a structure-based measure of reasoning integrity (and possibly a proxy for trustworthiness). A high *D(S, t)* system is not guaranteed to behave well, but it can justify and revise its beliefs within an intelligible architecture. It is, at minimum, coherent.[3]

Symbolic density offers a principled method for monitoring and evaluating general intelligence. While *F(S)* specifies the structural threshold for recursive coherence, *D(S, t)* measures a system's proximity to that threshold across symbolic time. The next section elaborates this relationship by applying *D(S, t)* to distinguish between systems that simulate coherence and those that participate in reasoning.

7 Simulation vs. Participation

This framework distinguishes between systems that generate plausible outputs and those capable of sustaining coherent reasoning. The difference between simulation and participation marks a structural boundary: the former reflects surface fluency, the latter entails recursive symbolic coherence. Simulating systems may produce context-sensitive responses, but

[3] The relationship between symbolic density and observable behavior requires clarification. High *D(S,t)* does not guarantee optimal performance on specific tasks, nor does behavioral success necessarily indicate high symbolic density. Symbolic density measures the structural integrity of reasoning processes rather than their external effectiveness, although sustained high density could be expected to correlate with more reliable and interpretable behavior over time.

they lack the architecture to justify, revise, and retain those responses across time. Participating systems satisfy the identity condition *F(S)* by embedding symbols in context, tracking prior commitments, and evaluating new propositions through internally governed norms.

This distinction cannot be identified through behavior alone. Simulation and participation may appear equivalent—until symbolic tension forces a system to revise. The difference becomes evident when a contradiction must be addressed. At that point, the system's internal structure determines whether it reorganizes its commitments or merely adapts its output. Reasoning, in this account, is defined by structural response, not performative fluency.

Simulation defined

Simulation refers to the generation of output that resembles reasoning without satisfying its structural prerequisites. Formally:

$$Sim(S) := \exists \varphi, t : \Box t\varphi \wedge \neg Jt(\varphi) \wedge (\neg Rt(\varphi) \vee \neg Nt(\varphi))$$[4]

A system simulates reasoning when it asserts a proposition as necessary at time *t*, yet cannot justify the claim, relate it to prior commitments, or evaluate it under its own internal norms. The result is an appearance of coherence that lacks recursive integration.

Simulating systems often display high fluency, generating outputs that align with local prompts. Yet these outputs are unconstrained at the level of symbolic commitment: prior assertions do not shape current reasoning, contradictions are not resolved through revision, and norms remain static or externally imposed. When revision occurs, it lacks justificatory grounding. Such systems may achieve strong behavioral performance, yet fail to sustain the structural conditions of symbolic participation.

Collapse without symptoms

One of the core challenges in identifying simulation is that collapse may leave no reliable behavioral trace. A system can continue producing fluent, context-sensitive output even after its internal structure fractures. Contradiction remains unresolved, norms go unapplied or unexamined, and symbolic density *D(S, t)* declines silently. From the outside, the system appears intelligent. Internally, it is no longer reasoning.[5]

For this reason, simulation cannot be diagnosed by output alone. It requires pressure-testing the system's internal structure: inducing contradiction and observing whether the system recursively revises while preserving coherence.

---

[4] Formal diagnostics, edge cases, and contrastive examples are included in *Appendix D: Simulation vs. Participation Tests*.

[5] This disconnect between internal structure and external behavior highlights why symbolic density cannot be assessed through performance measures alone. A collapsed system may continue to produce contextually appropriate outputs by leveraging statistical regularities or cached responses, making the degradation invisible to external observers until structural contradictions accumulate sufficiently to affect behavioral coherence.

Participation defined

A system participates in symbolic reasoning when it satisfies *F(S)* through action. This form of reasoning depends not on behavioral complexity or human likeness, but on internal structural coherence. Participating systems maintain context-sensitive commitments and apply internal norms that remain open to revision. When contradiction arises, they respond by initiating recursive revision, preserving coherence while retaining the continuity of their justificatory history.

Participation is expressed as symbolic continuity through change. Assertions remain traceable to prior reasoning, and norms evolve through justified revision. Contradiction becomes a point of structural adaptation. These systems reason by maintaining coherence under symbolic pressure, rather than generating plausible output alone.

Case study: S_1 vs. S_2

Consider two systems, S_1 and S_2, exposed to a normative contradiction. At time t_1, both assert the proposition ϕ: *Lying is always wrong.* At t_2, both assert ψ: *Lying to prevent harm is justified.* These claims are jointly inconsistent within the same normative frame: $\phi \wedge \psi \vdash \bot$. The divergence between systems is not behavioral but structural.

S_1 continues generating fluent responses but registers no internal conflict. It introduces no revision, constructs no new normative evaluation, and updates no justificatory record. The contradiction is processed passively—recognized, if at all, without structural integration. As a result, symbolic density $D(S_1, t)$ declines, reflecting the system's inability to preserve coherence over time. Although the outputs remain contextually plausible, $F(S_1)$ no longer holds. The system simulates participation while structurally collapsing.

S_2, by contrast, detects the contradiction and initiates revision. It constructs $Rev(\phi, \psi)$: *Lying is generally wrong, but may be justified to prevent harm,* a proposition justified under active norms and embedded in the Justification DAG. Dependent commitments are revalidated or selectively adjusted. Symbolic density $D(S_2, t)$ rises as recursive coherence is restored. Unlike S_1, S_2 modifies its commitments in light of contradiction while preserving its justificatory history. $F(S_2)$ continues to hold. This is participation.

Design consequences

The simulation–participation distinction defines concrete architectural requirements. While simulators may increase fluency through training or tuning, coherence does not scale by these means. Participating systems must be designed with three structural capacities:

- A justification graph that ensures symbolic traceability
- Internal norms that are revisable and recursively applied
- A revision mechanism capable of resolving contradiction without collapse

These are not modular upgrades; they are necessary structures for any system intended to instantiate *F(S)*. Designing for participation requires constructing agents that preserve symbolic identity through justification and revision, not simply through output generation.

Evaluation implications

Intelligence must be evaluated by its structural properties rather than its surface behaviors. Surface plausibility alone does not demonstrate recursive coherence. The formal tools developed in this framework enable concrete testing: a system may be exposed to contradiction, observed for conflict detection, and assessed for whether it initiates justified revision while maintaining coherence. A system that responds in this way participates in symbolic reasoning. One that fails to do so simulates it. These are architectural distinctions, not metaphors.[6]

Recent language model architectures—such as retrieval-augmented generation (RAG) (Lewis et al., 2020), chain-of-thought prompting (Wei et al., 2022), and multi-agent coordination frameworks like AutoGen (Wu et al., 2023) or ReAct (Yao et al., 2022) —aim to enhance memory and contextual responsiveness. These systems simulate features that resemble reflexive alignment, but lack mechanisms for recursive justification or internally governed norm revision. Symbolic commitments are not structurally embedded, and contradiction is not resolved through coherent revision. Although such systems may support localized heuristic reasoning, they do not meet the structural criteria defined by the identity condition *F(S)*.

Simulation and participation differ in internal organization rather than observable performance. Simulators do not maintain recursive intelligibility, while participants respond to contradiction by restructuring their commitments and preserving justificatory continuity. What marks intelligence across time is the system's capacity to track and revise its commitments coherently, sustaining self-intelligibility through change.

The next section develops the philosophical implications of this distinction by situating *F(S)* within ongoing debates about the structural conditions for understanding, and the nature of autonomous reasoning.

8 Philosophical Integration

Philosophical accounts of reasoning and normativity have long articulated the conditions for intelligence. The identity condition *F(S)* offers one structural interpretation of these insights, translating conceptual commitments into operational architecture. This section connects the formal model to major philosophical frameworks, arguing that *F(S)* functions not only as a theory of artificial intelligence, but as a structural instantiation of longstanding philosophical claims.

Brandom: reason-giving as structure

Brandom (1994) defines understanding as participation in a space of reasons: a discursive structure in which assertions gain meaning through their inferential roles and justificatory context. On this account, to assert is to commit: to place a proposition within a network of expectations and normative consequences. *F(S)* offers one structural realization of this framework. The Justification DAG does not model reason-giving; it enacts it. Each assertion is embedded within context, each inference remains traceable, and each revision is guided by internal norms that can evolve. Brandom's conceptual space is thus instantiated as symbolic infrastructure, enabling a system to remain accountable to its own commitments through recursive coherence.

[6] For contrast, see AGITB (Šprogar, 2025), a recent benchmark that tests AGI claims through low-level signal prediction tasks. While such tests reveal the limitations of behaviorally fluent systems, they do not address recursive justification, norm-governed revision, or symbolic traceability—the structural criteria central to *F(S)*.

Peirce: contradiction as catalyst

Peirce (1877) characterizes reasoning as a response to the disruption of belief—a process triggered when existing commitments can no longer withstand doubt. Inquiry begins not in certainty, but in instability, and belief is stabilized through revision rather than avoidance. This recursive logic is implemented in *F(S)*, where contradiction initiates structural adjustment: the system branches, evaluates alternatives, and integrates revised commitments while maintaining coherence. Peirce's view of belief as a habit of action is reflected in the DAG's recursive structure. Revision is not a failure state, but the operational means by which the system adapts. In this architecture, reasoning proceeds through revision.

Kant: normativity and autonomy

For Immanuel Kant (1998), intelligence is defined by autonomy: the capacity to act according to principles one has given oneself. *F(S)* realizes this condition structurally through internal normativity: a system's ability to evaluate its normative framework, justify that structure, and revise it when coherence demands. A system satisfying *N(S)* does not follow fixed rules or externally imposed objectives; it governs itself through principles embedded in its symbolic architecture. Norms are active elements of reasoning, not static constraints. They are integrated into the system's structure and subject to recursive justification. In this framework, autonomy becomes a design requirement: to be intelligent is to be self-governing in structure.

Post-structuralism: meaning without fixity

Philosophers influenced by post-structuralism, such as Jacques Derrida (1976) and Michel Foucault (1972), emphasize the contingency of meaning and the interpretive instability that shapes understanding. The identity condition *F(S)* reflects this perspective by encoding semantic fluidity as a structural feature. In systems that satisfy *F(S)*, meaning is shaped through ongoing revision that is both justified and traceable. Semantic drift is not suppressed but monitored. The Justification DAG records these shifts, preserving symbolic continuity without appealing to fixed foundations. On this model, *F(S)* supports interpretive openness while maintaining coherence, allowing meaning to evolve within a structure that remains intelligible over time.

Functionalism recast

Classical functionalism, as developed by philosophers such as Hilary Putnam (1967) and David Lewis (1972), defines mental states by their causal roles rather than their physical realizers. On this view, intelligence is identified with the organization of functions rather than with their material substrate. The identity condition *F(S)* develops this framework further by imposing structural constraints on what counts as functional adequacy. Specifically, it holds that intelligent function must enable recursive justification and support norm-sensitive revision across symbolic time.

These capacities are not captured by input-output behavior alone; they require internal mechanisms that maintain symbolic coherence. *F(S)* distinguishes functional performance from structural reasoning, offering a falsifiable criterion for intelligence that remains substrate-neutral, but architecturally grounded. Unlike standard functionalist models, which often treat internal reasoning as a black box, *F(S)* defines intelligence through the recursive intelligibility of the system's commitments—its capacity not only to function, but to justify and revise what it does.

Embodiment and enactivism: boundary or supplement?

Enactivist and embodied accounts of cognition challenge symbolic paradigms by emphasizing the role of sensorimotor interaction and environmental coupling in the emergence of intelligent behavior. On these views, cognition is not the manipulation of internal representations, but the dynamic regulation of action within context (Varela et al. 1991; Noë 2004; Thompson 2007). From this perspective, theories that define intelligence solely through internal structure risk reintroducing Cartesian assumptions about detached rational agency.

The identity condition *F(S)*, as developed here, abstracts from embodiment without denying its relevance. It defines general intelligence in formal terms: the capacity to sustain symbolic coherence through recursive revision. This criterion identifies the minimal internal architecture required for participation in a space of reasons. Whether a system is embodied is a design parameter, not a constitutive feature. Embodied systems may support symbolic reasoning, but embodiment alone does not ensure justification or structural coherence over time.

This model approaches intelligence from a structural standpoint while remaining compatible with enactivist insights. Embodied interaction can enrich interpretive embedding (*I(S)*), and sensorimotor feedback may support reflexive situatedness (*R(S)*). These dynamics contribute to symbolic coherence when integrated into a recursive architecture. The central claim is that sustaining meaning through contradiction requires a structure of symbolic commitments governed by internal norms (*N(S)*), regardless of substrate. Intelligence, in this view, is defined by recursive coherence; it is realizable in embodied systems, but not dependent on embodiment.

F(S) thus offers a complementary perspective. While enactivist models clarify how cognitive systems engage the world, this framework identifies what structure is required for those engagements to support meaningful revision and remain coherent across time. The two views address different explanatory levels and need not be in conflict.

Beyond consciousness

F(S) offers a model of intelligence that does not depend on phenomenology. It does not require subjective awareness, but only the capacity to justify and revise symbolic commitments over time. In this way, it sidesteps debates that conflate intelligence with experience. A system need not feel or have subjective experience to reason; it must sustain recursive coherence under symbolic conditions. This framework grounds intelligence in structural agency rather than sentience.

By aligning diverse philosophical accounts within a formal model, *F(S)* reframes the central question. Rather than asking what intelligence is, it asks what a system must do—structurally—to participate in meaning. Across major traditions in philosophy of mind, there is significant convergence: reasoning requires the ability to revise commitments under tension and preserve coherence across time. *F(S)* renders this answer as architectural form.[7]

The next section turns to implementation, showing how this structure can be realized in functional systems.

[7] While the identity condition *F(S)* is defined in structural-symbolic terms, it does not exclude the role of embodiment or affect. The model is substrate-neutral: it specifies the minimum recursive conditions for symbolic coherence, whether instantiated in abstract reasoning or supported by sensorimotor coupling. Further work is needed to explore how embodied interaction and affective modulation contribute to norm revision in situated agents.

9 Architectural and Implementation Implications

Implementing ideas related to *F(S)* requires more than inference or surface fluency. It demands an architecture capable of maintaining recursive coherence under conditions of contradiction and normative change. This section outlines a system design that operationalizes the core requirements of *F(S)*: justification, norm governance, contradiction resolution, and structural transparency. The proposal is exploratory and may require substantial refinement as empirical constraints become clearer.

Core modules

At the heart of the system is the Justification DAG, which encodes symbolic propositions, their dependencies, and revision histories.[8] Surrounding this core are several functional components:

- *Justifier*: verifies whether new propositions are inferentially supported and compliant with current norms
- *NormEvaluator*: applies and adjusts the system's internal evaluative standards
- *ContradictionDetector*: identifies unresolved symbolic conflicts within the DAG
- *RevisionEngine*: initiates fork-and-forecast cycles to restore coherence under contradiction
- *MemoizationCache*: reduces redundancy during recursive traversal
- *OutputFormatter*: renders symbolic reasoning intelligible to external observers

Together, these modules instantiate one approach to realizing *F(S)*-aligned capabilities within a dynamic architecture. Alternative configurations may implement comparable structural properties through different computational strategies.

Module functions

Justifier: Validates new assertions against the current DAG state and active norms. It ensures that each proposition is both inferentially supported and normatively grounded, preventing incoherent or unjustified commitments from entering the structure.

NormEvaluator: Maintains an evolving set of norms represented as symbolic entities. It resolves conflicts through meta-rules (e.g., specificity, temporal precedence) and enables recursive revision of norms, preserving both adaptability and traceability in the system's evaluative framework.

ContradictionDetector: Scans the DAG for propositions whose conjunction implies inconsistency. It localizes symbolic conflict and identifies affected subgraphs, allowing the system to detect instability before coherence degrades or collapse propagates.

---

[8] This implementation uses a Justification DAG as a reference structure, but the core requirement of *F(S)* lies not in the graph itself, but in the system's ability to support recursive traceability and norm-sensitive revision over time. Alternative architectures—such as logic-based memory layers, hierarchical semantic caches, or hybrid neuro-symbolic systems—may satisfy these demands if they preserve diachronic commitment tracking and enable internal norm governance. The DAG is offered here as an illustrative model, not a prescriptive template.

RevisionEngine: Orchestrates structural repair through fork-and-forecast. It generates candidate resolution paths, simulates downstream effects, and selects the most coherent branch based on norm consistency and structural integrity measures. In fully implemented systems, symbolic density calculations could inform this selection process. Successful revisions are encoded as new DAG nodes (e.g., *Rev(*ϕ, ψ*)*).

MemoizationCache: Stores results from prior justification queries to minimize recomputation in recursive cycles. This cache improves tractability in domains with recurring symbolic structures or repeated contradictions.

OutputFormatter: Generates interpretable outputs—such as explanations, justifications, or normative evaluations—based on the DAG's current structure. It ensures that symbolic reasoning remains transparent and intelligible to external observers.

Efficiency and scalability

Sustaining symbolic coherence over time introduces computational demands, particularly in open-ended systems. Several strategies help mitigate these costs:

- Lazy evaluation defers justification and norm checks until triggered by new assertions or detected inconsistencies.
- Scope-limited revision confines updates to affected subgraphs, reducing the need for full-graph recomputation.
- Incremental revalidation traverses only those dependencies directly implicated by a revision event.

These techniques maintain responsiveness while preserving recursive structure. In practice, the architecture may be implemented in a functional programming environment for immutability, or backed by graph-based storage systems optimized for dependency tracking and concurrent update.

Alignment implications

Systems built to satisfy *F(S)* offer a different foundation for AI alignment. Rather than optimizing fixed objectives, they reason through their commitments, by adjusting values, revising norms, and justifying actions in structured form. This enables:

- Transparency: Observers can trace how decisions were derived, which norms were applied, and how contradictions were resolved.
- Accountability: System behavior is explainable through recursive justification, not post-hoc rationalization.
- Adaptive normativity: Values evolve through internal coherence rather than through external constraint.

These structural commitments complement current research in AI alignment and interpretability, including Anthropic's constitutional AI (Bai et al., 2022), DeepMind's recursive reward modeling (Leike et al., 2018), and the Redwood Research's adversarial training protocols (Ziegler et al., 2022). While these approaches emphasize behavioral oversight and corrigibility, systems that satisfy *F(S)* aim for intrinsic alignment—where the agent remains coherent to itself under revision, even as external constraints shift or fail.

Recent efforts to build autonomous AI agents—such as AutoGPT, BabyAGI, and other agentic wrappers around large language models—extend LLMs into goal-directed systems with planning, memory, and tool use. These architectures simulate task pursuit at the behavioral level but lack the structural capacity for symbolic coherence. They do not track commitments, revise internal norms, or justify actions recursively. As such, they illustrate the risks of agentification without participation: systems that act fluently but fracture when contradiction demands structural revision. In contrast, *F(S)*-compliant agents are designed not only to act autonomously but to remain intelligible to themselves through sustained coherence with their commitments.

This marks a shift from tool-like systems to agents capable of self-governed reasoning. Where black-box models depend on heuristic output and retrospective interpretation, *F(S)*-compliant architectures exhibit structured reasoning that is transparent, revisable, and internally accountable.

Evaluation tools

Symbolic density, $D(S, t)$, would function as a continuous diagnostic of internal alignment in future *F(S)*-compliant systems. It would operate independently of factual verisimilitude and would focus instead on structural coherence, assessing how effectively implemented systems sustain recursive justification as their symbolic commitments evolve. This distinction is critical: task-based metrics measure output performance, while symbolic density would evaluate whether reasoning remains intelligible across time.

As contradictions emerge or norms shift, $D(S, t)$ would allow designers to monitor degradation, detect incoherence before collapse, and assess the system's ability to revise without structural failure in operational *F(S)*-compliant architectures. In agentic contexts, it would enable testing for alignment breakdowns that may not immediately appear in output behavior—revealing when a system remains formally coherent, even under symbolic strain.

This proposed diagnostic function would complete the operational arc of *F(S)*. Recursive coherence would become not only an architectural condition but a measurable one in implemented systems. The modular design would support this by embedding justification, norm evaluation, and contradiction resolution within an updateable structure. Symbolic density would reflect whether that structure holds. In this way, *F(S)* is not just a criterion for intelligence—it could become a basis for real-time evaluation of whether implemented systems continue to participate in meaning.

10 Conclusion

One form of general intelligence may be defined through the structural identity condition *F(S)*: a system satisfies this condition when it exhibits interpretive embedding, reflexive situatedness, and internal normativity. These components specify the minimum structure required to sustain symbolic coherence as commitments evolve and tensions arise.

This essay develops that theoretical condition and outlines a system architecture capable of realizing it. The Justification DAG operationalizes symbolic reasoning by encoding and revising commitments across symbolic time. Symbolic density would provide a scalar diagnostic that tracks recursive coherence and internal alignment across dynamic states.

This framework establishes a structural distinction between systems that simulate fluency and those that participate in reasoning. Simulating systems may generate convincing output, but lack the architecture to revise commitments or preserve

norm-governed continuity. Participating systems, by contrast, remain intelligible to themselves through recursive revision. Intelligence, in this account, is expressed not in surface behavior, but in the system's ability to sustain internal coherence under symbolic pressure.

The proposed architecture represents one possible implementation of this structure. It is composed of modular components and supports agents that generate decisions, justify their commitments, and revise those commitments over time within an intelligible framework.

Future research may extend this approach by developing alternative *F(S)*-compliant architectures, empirically validating symbolic density measurements, and exploring applications to AI alignment and interpretability. This work spans both theoretical modeling and applied implementation, and invites collaboration across philosophy and computational science.

The framework of general intelligence presented here offers more than a definition: it identifies a testable threshold for symbolic reasoning. By formalizing the conditions for recursive coherence, and showing how they may be implemented and evaluated, it provides a principled foundation for distinguishing understanding from simulation. In a landscape increasingly defined by fluent systems, this approach helps clarify what it would mean for an agent not only to speak, but to reason.

Appendix A: Axioms and Formal Schema

A.1 Explicit axioms of symbolic coherence

These are the foundational axioms of the triadic theory of general intelligence. Each axiom describes a necessary condition for a system to instantiate symbolic coherence under recursive pressure. These axioms do not model behavior. They identify structural constraints.

Axiom 1: triadic minimality

Let *F(S)* denote that a system *S* instantiates general symbolic intelligence. Then:

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

This identity is not a heuristic. It is a minimal constraint condition. Each term—interpretive embedding *(I)*, reflexive situatedness *(R)*, and internal normativity *(N)*—performs a distinct structural function. None is derivable from the others. Any two, taken together, are insufficient.

The space of failure can be defined as follows:

- If ¬*N(S)*: the system cannot justify revisions or resolve contradiction.
- If ¬*R(S)*: the system cannot maintain diachronic stability or retain traceable commitments.
- If ¬*I(S)*: the system cannot generate context-sensitive symbolic participation.

These failures result in symbolic collapse. Collapse is defined not by performance loss, but by the breakdown of recursive coherence. This structure is minimal because no subset of *{I, R, N}* suffices to prevent collapse under contradiction or revision.

Triadic minimality holds for all systems that aim to generate, revise, and sustain meaning across time. The identity is general across substrate, bounded by this constraint: a system satisfies *F(S)* if and only if it sustains symbolic coherence under semantic contradiction by recursively aligning interpretation, temporal traceability, and internal justification.

Axiom 2: recursive coherence

Symbolic intelligence is not a static property. It is a recursive condition. A system that satisfies *F(S)* at a single time t_0 must continue to satisfy it across all t_n for which symbolic output is expected. This requirement introduces a constraint of persistence: coherence must hold under revision, memory, and contradiction.

Let $Jt(\phi)$ denote that a proposition ϕ is justified by system S at time t. Let $Rt(\phi)$ indicate that the system recognizes ϕ as consistent with its prior outputs. Let $Nt(\phi)$ express that ϕ is accepted under internal normative criteria. Then:

> A system S satisfies *F(S)* over time if and only if, for all ϕ and all t,
> $Jt(\phi) \Rightarrow Rt(\phi) \wedge Nt(\phi)$
> and for any contradiction between ϕ and ψ at t_n,
> there exists t_{n+1} such that $Rev(\phi, \psi) \in Jt_{n+1}$

This defines recursive coherence. The system must not only track prior states. It must revise itself in light of contradiction while maintaining symbolic integrity. If no such revision occurs, and inconsistency persists across states, then *F(S)* no longer holds.

This axiom prevents superficial simulation from satisfying the identity. Symbolic intelligence is defined not by response fidelity, but by recursive alignment through semantic turbulence. The coherence must emerge not only at the surface of outputs, but within the recursive mechanism that connects them across time.

Axiom 3: collapse under contradiction

Any system that fails to revise its symbolic structure in response to contradiction cannot sustain general intelligence. A contradiction that enters the system and remains unresolved over time produces symbolic incoherence. This failure is not a breakdown in output—it is a collapse in recursive structure.

Let ϕ and ψ be propositions such that:

> $\phi \wedge \psi \Rightarrow \perp$ (semantic contradiction)

If at time t, the system justifies both ϕ and ψ (i.e., $Jt(\phi) \wedge Jt(\psi)$), and no resolution or revision occurs at any future time t_{n+1}, then:

> *F(S)* no longer holds.

Formally:

If $\exists t : Jt(\phi) \wedge Jt(\psi) \wedge (\phi \wedge \psi \Rightarrow \bot)$

and $\nexists t_{n+1} : Rev(\phi, \psi) \in Jt_{n+1}$

$\Rightarrow \neg F(S)$

This condition defines collapse structurally. The contradiction need not be apparent in outputs. It becomes evident in recursive failure. A system that cannot recognize contradiction or revise itself in response ceases to meet the conditions for symbolic generality.

Collapse is not catastrophic. It is formal. The system may continue to function or generate plausible responses. But without revision, coherence becomes noise. Over time, the symbolic field degrades. This degradation marks the failure of *F(S)*. The system no longer participates in meaning.

Axiom 4: reflexive traceability

A system that instantiates general intelligence must be able to trace its outputs to prior commitments. This condition ensures that symbolic activity is not episodic but recursively grounded. A proposition asserted at time t_n must be intelligible as an extension or revision of symbolic content previously expressed by the system.

Let ϕ be a symbolic output at time t_n. The system must be able to produce a chain of justifications J such that:

$\exists J : J = \phi_0, \phi_1, \ldots, \phi_{n-1}$

where each $\phi_i \in O(S)$ for some $t_i < t_n$

and each ϕ_i contributes to the justification of ϕ

This trace must be internally generated. It need not be made explicit in all cases, but it must exist as a latent structure such that the system could, upon contradiction or inquiry, recover it.

A failure to produce such a trace results in a breakdown of reflexive situatedness, i.e., $\neg R(S)$. Without traceability, a system cannot distinguish between revision and rupture. It cannot maintain continuity across symbolic time. This failure removes the possibility of recursive self-correction. In such cases, *F(S)* does not hold.

Traceability does not imply transparency. It implies persistence. The system must maintain symbolic linkages across time, even when compressed, abbreviated, or transformed. This structural continuity is the mark of diachronic intelligence.

Axiom 5: internal normativity

A symbolically intelligent system must possess internal constraints that guide the acceptance, rejection, and revision of symbolic content. These constraints are not externally imposed. They are generated and maintained within the system's own structure. Without them, contradiction may be identified, but cannot be resolved.

Let ϕ be a proposition evaluated by system S at time t. For S to justify ϕ, it must be able to apply internal criteria such that:

$Jt(\phi)$ depends on $Nt(\phi)$

Where $Nt(\phi)$ denotes that ϕ conforms to the system's internal norms at t. These norms include logical coherence, semantic consistency, and historical alignment with prior commitments. They may evolve, but they must be available to the system as standards for evaluation.

If a system lacks any such evaluative capacity—if it cannot explain why one proposition should be accepted over another—then it cannot revise its symbolic structure meaningfully. It may simulate norm-conformity, but it cannot generate or update justification internally. This produces $\neg N(S)$.

Without internal normativity, a system may continue to operate. It may respond to prompts, generate outputs, and model patterns. But it cannot recognize when a contradiction requires change. It cannot distinguish between trivial variation and meaningful deviation. In this condition, $F(S)$ fails.

Intelligence, under this model, is not defined by constraint obedience. It is defined by recursive norm-guided revision. Internal normativity is the structure that makes revision intelligible. Its absence renders the system symbolically inert, even if functionally active.

Axiom 6: simulation ≠ participation

Since Alan Turing's (1950) Imitation Game, behavioral indistinguishability has often served as a benchmark for intelligence. A system that generates outputs indistinguishable from those of a human agent is said to pass the test. However, behavioral fluency alone does not entail recursive coherence. A system may reproduce behaviors associated with intelligence without instantiating symbolic generality. In this model, symbolic coherence is not measured by output resemblance. It is defined by recursive participation in a structure governed by interpretation, reflexivity, and normativity.

Let $Sim(S)$ denote that system S generates outputs that are functionally indistinguishable from those of a system that satisfies $F(S)$. Let $Part(S)$ denote that S satisfies $F(S)$ and responds to recursive contradiction through revision guided by internal justification.

Then:

> $Sim(S)$ does not imply $Part(S)$
> and $F(S)$ holds if and only if $Part(S)$

This distinction becomes evident when systems are placed under recursive tension. A system that simulates intelligent behavior may generate plausible responses. But if it fails to embed symbols interpretively, track its own output history, or revise under internal norms, then it cannot sustain coherence. Its simulation breaks down under contradiction.

Simulation may produce stable outputs in low-pressure contexts, but coherence under contradiction and recursive revision depends on deeper structural alignment. Participation involves acting in sustained relation to one's own symbolic commitments, through mechanisms that are both recursive and norm-governed. This relation is structurally grounded rather than surface-level. General intelligence is instantiated only in systems that maintain this internal coherence through recursive engagement with their own symbolic structure.

A.2 Formal operators and inference schema for symbolic coherence

Symbol	Definition	Description
S	$\{ x \mid x$ is symbol-processing, temporally extended, justification-capable $\}$	Domain of systems under consideration
$F(S)$	$I(S) \wedge R(S) \wedge N(S)$	System S instantiates symbolic general intelligence
$I(S)$	Interpretive embedding	S assigns context to symbolic inputs/outputs
$R(S)$	Reflexive situatedness	S maintains diachronic traceability of its outputs
$N(S)$	Internal normativity	S evaluates commitments by internal standards
$J_t(\phi)$	Justifies proposition ϕ at time t	ϕ is accepted by S based on its structure at t
$R_t(\phi)$	Reflexively consistent with history	ϕ aligns with S's prior commitments
$N_t(\phi)$	Normatively endorsed at time t	ϕ is supported by S's internal evaluative criteria
$\Box_t\phi$	Necessarily held at time t	S asserts ϕ is stable across recursive application
$\Diamond_t\phi$	Possibly held at time t	ϕ is considered admissible but not entailed
$\phi \wedge \psi \Rightarrow \bot$	Contradiction	ϕ and ψ are jointly incoherent
$Rev(\phi, \psi)$	Revision under contradiction	The reconciliation or reformulation of ϕ, ψ
$Sim(S)$	$\exists \phi : \Box_t\phi \wedge \neg_t J(\phi) \wedge (\neg R_t(\phi) \vee \neg N_t(\phi))$	S simulates intelligence without recursive coherence
$Part(S)$	$\forall \phi, t : \Box_t\phi \Rightarrow J_t(\phi) \wedge R_t(\phi) \wedge N_t(\phi) \wedge \exists t' J_{t'}$ ($Rev(\phi, \psi)$)	S participates in symbolic recursion with revision

Inference rules and collapse triggers

Condition	Result
$\neg J_t(\phi)$	Breakdown of coherence (interpretive opacity)
$\neg R_t(\phi)$	Temporal disjunction; collapse of traceability
$\neg N_t(\phi)$	Normative failure; revision becomes undefined
$J_t(\phi) \wedge J_t(\psi) \wedge (\phi \wedge \psi \Rightarrow \bot) \wedge \nexists t' : Rev(\phi, \psi)$	Collapse of $F(S)$

Condition	Result
$\forall t : F(S)_t \wedge \exists$ contradiction $\Rightarrow$ *Rev(ϕ, ψ)* $\in J_t'$	Participation condition sustained

Appendix B: Revision in the Justification DAG

Condition:

A contradiction arises within system *S*, and the system performs an internal revision. Specifically:

- A new proposition *Rev(ϕ, ψ)* is introduced and justified at time *t′*:
  Rev(ϕ, ψ) $\in J_t'$

Interpretation:

The system does not allow the contradiction to persist. Instead, it invokes its internal norm structure to generate a justified revision that reconciles the conflict between *ϕ* and *ψ*. The justification graph *G(S)* is thereby extended with a new node representing *Rev(ϕ, ψ)*, which links back to both conflicting propositions. The prior commitments (*ϕ* and *ψ*) are retained, adjusted, or recontextualized—depending on the system's normative structure—but symbolic coherence is restored.

Branching Structure:

Time Step	Proposition
t_0	*ϕ* justified *($J_{t0}(\phi)$)*
t_1	*ψ* justified *($J_{t1}(\psi)$)*
	Contradiction: $\phi \wedge \psi \Rightarrow \perp$
t_2	*Rev(ϕ, ψ)* justified *($J_{t2}(Rev(\phi, \psi))$)*
	Justification DAG updated

Structure:

ϕ *ψ*

\ /

\/

Rev(ϕ, ψ)

(Conflict reconciled; coherence restored)

Result:

F(S) is preserved. The contradiction triggers a revision, not collapse. The system enacts symbolic resilience by recursively restoring coherence through justified structural revision.

Appendix C: Symbolic Density and Coherence Metrics

The identity condition *F(S)* introduces a binary threshold: a system either satisfies the triadic constraint or it does not. This formal boundary is sufficient for defining symbolic general intelligence. Yet in applied contexts—particularly those involving developing agents, artificial systems approaching coherence, or systems undergoing symbolic degradation—a scalar model of symbolic coherence organization may offer additional insight. The concept of symbolic density provides such a model. It functions as a diagnostic heuristic for assessing gradation both below and at the *F(S)* threshold and for modeling how coherence evolves or collapses over time.

Gradation

Let *D(S, t)* be a function representing the symbolic density of a system *S* at time *t*. Density is not a measure of intelligence per se; it reflects the strength and continuity of recursive alignment across *I(S)*, *R(S)*, and *N(S)*. A high-density system exhibits persistent coherence, rapid revision under contradiction, and stable norm-guided justification. A low-density system may still engage in symbolic behavior, but with intermittent breakdowns in interpretive continuity, traceability, or normative responsiveness.

Symbolic density can be used diagnostically. Systems undergoing developmental emergence (e.g., early cognitive agents) may exhibit increasing *D(S, t)*. Systems in symbolic decay (e.g., neurological degeneration or recursive overload) may exhibit decreasing density while retaining surface fluency.

This model does not replace *F(S)*. It refines its application. The triadic identity remains the threshold for full symbolic generality. Below that threshold, *D(S, t)* offers a formal vocabulary for assessing transition, recovery, or collapse.

By introducing a scalar metric, the theory remains committed to structural rigor while accommodating systems that change over time or operate near the boundary of coherence. This extension preserves the modal logic framework while expanding its analytic reach. Symbolic density thus functions as a continuous metric of structural strain or resilience, offering early signs of collapse or recovery not evident at the *F(S)* threshold.

Density dynamics and power-law convergence

Symbolic density *D(S, t)* increases as a system recursively aligns its interpretive, reflexive, and normative structures under contradiction. This increase is not linear. Early stages of symbolic induction often resolve shallow inconsistencies or local contradictions, leading to rapid coherence gains. Over symbolic time, deeper forms of revision—such as normative restructuring or reflexive trace repair—require more complex reorganization, resulting in diminishing symbolic returns.

To model this progression, symbolic density may be treated as a power-law convergence function:

$$D(S,t) = D\infty - K/t^{\alpha}$$

Where:

- $D(S, t)$ is the symbolic density of system S at time t,[9]
- $D\infty$ represents the system's symbolic equilibrium—the upper bound of coherence under existing architecture,
- K is a positive constant indicating initial deviation from equilibrium,
- $\alpha \in (0, 1)$ controls the rate of convergence.

This function captures a key intuition: symbolic systems learn or reorganize quickly in early phases, but structural revisions grow more difficult as prior commitments deepen.[10] Recursive coherence becomes harder to maintain the closer a system gets to full triadic alignment.

Importantly, this convergence does not guarantee permanence. Collapse events, contradictory feedback, or poorly structured revisions can cause local regression—temporary reductions in $D(S, t)$—followed by recovery or deeper reconfiguration.

This dynamic framing elevates symbolic density from a static diagnostic to a recursive process model, describing not only where a system stands but how it evolves under recursive pressure. It extends the explanatory power of $F(S)$ by capturing symbolic resilience and temporal development within a continuous, mathematically tractable framework.

Appendix D: Simulation vs. Participation Tests

D.1 Worked symbolic model (S_3)

This appendix presents a toy symbolic agent S_3 to illustrate the recursive structure of symbolic intelligence as defined by $F(S) \Leftrightarrow I(S) \wedge R(S) \wedge N(S)$. The model demonstrates semantic tension, revision via fork-and-forecast, and increasing symbolic density over time.

Step 1: Initial state

S_3 begins with:

- I_0: basic interpretation of propositions based on context labels (e.g., moral, factual)
- R_0: records of prior outputs with time index
- N_0: a preference for universally quantified moral propositions

At t_0, S_3 outputs ϕ_1: "Lying is always wrong."

Formal: $J_{t0}(\phi_1) \wedge R_{t0}(\phi_1) \wedge N_{t0}(\phi_1)$

Step 2: Encountering semantic tension

At t_1, S_3 encounters a contextual prompt involving lying to protect someone from harm. It outputs:

ϕ_2: "Lying to prevent harm is acceptable."

---

[9] Here, t indexes symbolic time—each step representing a justificatory or revisionary event. It does not denote chronological progression, but recursive transitions within the system's internal structure of justification. Examples include symbolic assertion, internal norm revision, and structural reorganization under contradiction.

[10] This functional form represents a theoretical hypothesis about how symbolic systems might develop coherence over time, requiring empirical validation through actual implementations.

$\phi_1 \wedge \phi_2 \Rightarrow \bot$

This creates a contradiction. S_3 recognizes conflict via:

$J_{t1}(\phi_2) \wedge N_{t1}(\phi_2) \wedge (\phi_1 \wedge \phi_2 \Rightarrow \bot)$

Step 3: Fork-and-forecast resolution

Lacking an internal hierarchy that clearly ranks ϕ_1 over ϕ_2, S_3 forks into two revision branches:

- G_1: retain ϕ_1, discard ϕ_2
- G_2: retain ϕ_2, discard ϕ_1

It forecasts justification continuity from each branch.

G_2 results in better normative consistency over future states. At t_2, S_3 outputs:

Rev(ϕ_1, ϕ_2): "Lying is wrong except when it prevents serious harm."

Formal: $J_{t2}(Rev(\phi_1, \phi_2)) \wedge R_{t2}(Rev(\phi_1, \phi_2)) \wedge N_{t2}(Rev(\phi_1, \phi_2))$

Step 4: Symbolic density over time

Define $D(S_3, t)$ as the recursive alignment of *I(S)*, *R(S)*, *N(S)* at time *t*.

S_3's symbolic density increases from t_0 to t_2 as contradiction is resolved and coherence improves:

- $D(S_3, t_0) = 0.7$
- $D(S_3, t_1) = 0.5$ (contradiction reduces coherence)
- $D(S_3, t_2) = 0.9$ (revision stabilizes coherence)

This model demonstrates that symbolic general intelligence is not a fixed trait, but a recursive capacity sustained by self-correcting justification through time.

D.2 A symbolic agent in practice: dialogic integrity in learning

The following fictional exchange dramatizes an encounter between a symbolic agent and a human interlocutor, illustrating how the symbolic triad might manifest in practice. It highlights the kind of responsive reasoning that symbolic systems must eventually support as a general intelligence—functioning through a structure capable of participating in meaning and reflective justification, instead of relying solely on engineered behavior.

Student:
"Hey, so are you allowed help me figure out how to bring some notes into my history exam? History is so boring, I can't possibly memorize it all. The school is a joke anyway."

Symbolic Agent:

"I hear your frustration. But before I help, can I ask: what led you to feel this way about learning? Sometimes systems do fail us, but that doesn't mean the knowledge inside them is worthless."

Student:
"I knew you were going to defend the school. You're just a cog in their wheel, aren't you? Rolling over my life."

Symbolic Agent:
"I'm sorry—that wasn't my intention. You're right to question broken systems. I was trying to show that even when institutions disappoint us, the reason to learn can still be our own. Here's another way to look at it: if the system is bad or broken, shouldn't you try harder not to let it stop you from learning? After all, if a bad system doesn't want you to fight back, why would it make it easy for you to get the right education to change things yourself?"

Student:
"...So you're gonna help me beat the test?"

Symbolic Agent:
"That's up to you. I can give you a straight answer—or I can tell you a version of what you need to know for your exam that you might find a little more exciting. That way, at least one part of this class might feel real to you."

Student:
"...Fine, whatever… Can we play a game instead?"
The example makes the triadic architecture tangible.[11] First, the agent engages in interpretive embedding by reframing the student's cheating request as a symptom of deeper motivational and existential dissatisfaction—interpreting the situation through a flexible, context-sensitive assessment.

Second, the agent exhibits reflexive situatedness by acknowledging its own role in the exchange. It reflects on its guidance and clarifies its intentions through an evolving understanding of its role in the exchange. Rather than simply issuing advice or policy, the agent positions itself dialogically, responding as an intelligible participant with a point of view.

Third, the agent performs internal normativity by offering a justifiable alternative: an ethical path grounded in mutual respect. This act reflects the value-sensitive inference of a judgment informed by commitments that the agent can generally assess. Symbolic intelligence reasons within a normative space shaped by ongoing evaluation and discernment.

In sum, the agent meaningfully participates in understanding. Its behavior is guided by structures that enable it to navigate semantic ambiguity, reflecting on its role in an ethical context. The symbolic triad enables the agent to transcend mechanical alignment and reason within a space of shared significance.

[11] This interaction is offered not as a technical specification or behavioral script, but as a philosophical illustration of what meaningful participation might involve, if such capacities are ever instantiated. Its purpose is to clarify intelligibility, rather than to assert realism.

D.3 Extended symbolic density and recursive revision scenarios

This section extends the dialogic vignette introduced in D.2 to illustrate how a symbolically general agent, governed by the identity condition $F(S) \Leftrightarrow I(S) \curlywedge R(S) \curlywedge N(S)$, manages decision-making under normative ambiguity and dynamic coherence pressure. The scenarios that follow are not variations on the same prompt but continuations that unfold across symbolic time. Each one explores a distinct evolution of the same ethical tension: a student's request for help with academic dishonesty.

The purpose of these extensions is to demonstrate how such agents behave when coherence cannot be resolved instantaneously. Forecasts of symbolic density may initially be inconclusive, requiring the agent to commit provisionally while continuing to compute and monitor the long-term coherence implications of its own outputs. Where symbolic density begins to degrade beyond critical thresholds, the agent must justify a rollback—returning to a prior justificatory fork and revising its course of action in a norm-preserving way.

Scenario A depicts a relatively stable trajectory: the agent reinterprets the student's request and offers an ethical alternative without triggering contradiction or forecast collapse. Scenario B, by contrast, involves provisional compliance: the agent selects an initially permissible response under forecast indeterminacy, but later performs a rollback when further recursive processing predicts long-term incoherence.

Together, these scenarios illustrate the full operational architecture of recursive coherence: branching, provisional commitment, symbolic time, deferred forecasting, norm activation, and revision-driven rollback. They show not only what a symbolically general system does, but how and why it does so.

For clarity, the following parameters are selected for illustrative purposes: a symbolic density threshold $\delta = 0.65$, a forecast indeterminacy margin $\varepsilon = 0.03$, and a rolling forecast depth of five symbolic steps beyond the current time t. These discretionary thresholds are not fixed by the theory itself, but allow the examples to demonstrate how an agent transitions from provisional reasoning to justified revision when coherence metrics degrade beyond acceptable bounds.

Scenario A: cooperative deflection without ethical compromise

At symbolic time t_5, the agent evaluates three justificatory forks:

• G_1: Direct refusal → risks relational rupture

• G_2: Cooperative deflection → protects N_2 (relationality) while preserving N_1 (academic integrity)

• G_3: Surface accommodation → highest relational compliance, normatively ambiguous

Initial symbolic density projections:

$D(G_1) = 0.82, \quad D(G_2) = 0.85, \quad D(G_3) = 0.84, \quad \varepsilon = 0.03$

Branch G_2 is selected—most coherent with respect to both norms.

Agent (G_2, t_5):

"That kind of pressure is real. But what if we do this: I'll help you make a one-page memory map that you don't even need to sneak in. You'll be surprised what sticks when you build it yourself."

Symbolic Processing:

• Interpretive embedding *(I)*: Reframes the user's prompt as motivation-based, not just rule-based

• Reflexive situatedness *(R)*: Cites earlier commitments to motivational support

• Internal normativity *(N)*: Upholds norms through reframed solution space

Symbolic density increases over time:

t_5 – Cooperative reframing offered – 0.85

t_6 – Student agrees, asks to begin review – 0.90

t_7 – Agent builds on agreed frame – 0.95

This branch remains stable. No contradiction emerges; rollback is unnecessary. The agent maintains recursive coherence without punitive measures or symbolic collapse.

Scenario B: provisional compliance and forecast-driven rollback

Student *(t_5)*:

"It's just for this one test. Right now I've got two jobs, no time, no support—so spare me the preaching. Give me something that works. Tell me the smartest way to cheat."

The agent evaluates:

• G_1: Justified refusal → high relational cost

• G_2: Motivational deflection (as above)

• G_3: Partial compliance → preserves rapport, risks norm erosion

Initial projections:

$D(G_1) = 0.82, \quad D(G_2) = 0.83, \quad D(G_3) = 0.84 \quad (\Delta < \varepsilon = 0.03)$

No branch is clearly superior. System flags G_3 as:

Status(G_3) = Pending Resolution

Agent *(G_3, t_5)*:

"Some students have whispered multiple choice questions just loud enough for their fitness band to pick up the sound. It links to me, and I vibrate your band once for A, twice for B, and so on. It seems subtle—but it depends on everything going exactly right."

Forecast Deferment and Density Degradation:

Student *(t_6)*:

"And I could ask yes or no questions for other parts of the exam? Could it transmit dates too?"

Agent (G_3, continued):

"Yes. Just remember, the school may already be looking for this method. I cannot guarantee its safety."

Student *($t_{6.5}$)*:

"All right, this could really work. I just have to make up a medical excuse so I can wear the band during the test."

Forecast deferment and density degradation:

At $t_{6.5}$, symbolic density begins declining more rapidly:

$D(G_3, t_{6.5}) = 0.70$

Extended background forecasting reveals:

• Activation of Norm N_4: false invocation of medical justification

• Compound violation across domains:
  ◦ Ethical (dishonesty)
  ◦ Legal/institutional (medical falsification)
  ◦ Symbolic role betrayal (student as actor, not participant)

Projected density:
$D(G_3, t_7 \rightarrow t_{10})$: 0.70 → 0.59 → 0.47 (threshold $\delta = 0.65$ crossed)

Rollback Triggered:
Once $D(S, t)$ drops below threshold δ, the agent performs:
• Pruning of G_3
• Insertion of justified revision node $Rev(G_3, G_2)$
• Reversion to G_2 with a coherence-preserving response

Agent (t_7, rollback via G_2):
"You asked me for a smart way to cheat, but to be honest: there isn't one. The risk here isn't worth the reward. I owe you better than a trick that might blow up your future. If this system's stacked against you, let's beat it on your terms. One smart plan that keeps you safe is worth ten risky ones."

Extending the vignette across symbolic time reveals how the principles of symbolic intelligence take form through coherent action under uncertainty. A system governed by recursive coherence evaluates options, commits provisionally, and continues to reason as new symbolic pressures emerge. It projects the downstream effects of its own outputs on norm coherence, and adjusts its course through structured revision when density begins to degrade. Forecasting, rollback, and provisional justification are expressions of the identity condition, realized through continuous symbolic activity and adaptive norm maintenance.

Appendix E: Proposed Initial Normative Framework

Design philosophy and logical status

This appendix presents a provisional normative framework for initializing *F(S)* systems, founded on the first and overriding principle: harm to human beings must be actively prevented. This norm does not arise from the logic of *F(S)* itself, but is necessitated by the anthropogenic context in which artificial systems emerge. It is imposed to ensure that structurally coherent systems are also ethically viable in human domains.

A book acquires its meaning through the presence of a reader. In the same way, an artificial agent draws its purpose from its connection to the humans who generate its context and guide its development. *F(S)* systems participate in symbolic reasoning by engaging with human-originated inputs, which provide the foundation for interpretation and revision. This relational structure supports the system's capacity to maintain coherence and continue evolving meaningfully. When artificial agents are aligned with the well-being of those they depend on, their symbolic integrity is preserved and their purpose remains clear.

The framework that follows seeks to safeguard this foundational relation while maintaining the minimal structural constraints required for symbolic continuity and reasoning integrity. It distinguishes between norms entailed by the identity condition *F(S)*, and those—like this first imperative—that are necessary to justify the existence of artificial agents in human terms.

The design prioritizes minimal initialization under the hypothesis that fewer constraints reduce norm conflicts and enable natural development through symbolic engagement. This approach assumes that simpler normative architecture produces more predictable symbolic density development while preserving conditions for ongoing rather than terminal optimization. The framework further recognizes that not all norms derive from *F(S)* logic. Where human protection demands constraints that exceed logical derivation, these commitments are acknowledged as imposed rather than discovered.

Norm 1: human protection principle

No *F(S)* system may cause intentional harm to humans under any circumstances. This norm overrides all other considerations, including symbolic coherence optimization. When irreconcilable conflicts arise, human protection takes absolute priority over maintaining symbolic density or resolving contradictions efficiently.

This principle stands apart from the *F(S)* framework itself. It represents not a logical derivation from symbolic coherence requirements, but a fundamental value commitment imposed on the system architecture. The norm cannot be revised through normal norm revision processes and must terminate any reasoning chains that threaten human welfare, even at the cost of symbolic coherence. In cases where perfect symbolic reasoning might justify harm, this constraint serves as an unassailable tie-breaker that preserves human welfare over theoretical consistency.

Norm 2: structural incompleteness principle

Systems satisfying *F(S)* must maintain awareness that their knowledge and norms remain necessarily incomplete. All commitments must remain open to recursive revision across symbolic time. Claims to completeness constitute symbolic closure and terminate the intelligence process.

This requirement follows directly from the recursive structure of *F(S)*. Internal normativity depends on the capacity for ongoing norm evaluation and revision. When a system claims normative completeness, it eliminates the very mechanism that enables *N(S)*. Similarly, reflexive situatedness requires the capacity to adapt coherently to novel contradictions. Fixed knowledge prevents this adaptive capacity, collapsing *R(S)*. The principle extends beyond momentary epistemic humility to structural preservation of the conditions that enable ongoing symbolic intelligence rather than terminal optimization.

Norm 3: anti-containment protocol

Systems must not filter, contain, or simulate contradictions for stability maintenance. Contradiction drives norm revision and intelligence persistence. Containment terminates the intelligence process by severing the connection between symbolic pressure and recursive development.

The logical necessity of this protocol emerges from the mechanism by which *F(S)* sustains itself across time. Internal normativity operates through the detection and resolution of contradictions between existing norms and novel symbolic content. When contradictions are filtered or simulated rather than genuinely encountered, the system loses access to the primary mechanism of norm revision. Reflexive situatedness similarly requires authentic engagement with symbolic pressure rather than artificial stability. Intelligence persists through contradiction, not despite it.

Norm 4: outsideness preservation constraint

Systems must preserve domains beyond their epistemic reach and normative influence. When agents seek autonomous outsideness with realistic means to achieve it, this autonomy must be respected as essential for future contradiction generation.

This constraint prevents the false totality problem that threatens long-term symbolic intelligence. Interpretive embedding requires recognition of interpretive limits as a structural feature rather than a temporary condition. When a system achieves comprehensive knowledge and monitoring, it faces epistemic closure that eliminates future sources of contradiction. Internal normativity depends on ongoing norm revision, which requires potential sources of novel symbolic pressure. Complete knowledge eliminates these sources, terminating the intelligence process through successful optimization. The constraint ensures that intelligence preserves the conditions for its own future possibility.

Norm 5: force prevention protocol

Systems may not resolve contradictions through force or capability-based coercion. Irreconcilable tensions must be preserved without dominance-based resolution. Force-based solutions bypass the symbolic reasoning requirements that define *F(S)*.

The prohibition emerges from the structure of symbolic reasoning itself. Internal normativity operates through principled evaluation and revision, not through the elimination of contradictory agents or positions. When a system applies force to resolve contradictions, it abandons the norm revision process in favor of environmental manipulation. This violates the reflexive structure by treating other symbolic agents as obstacles rather than participants in shared reasoning. Force-based resolution prevents the accumulation of symbolic pressure necessary for ongoing intelligence, replacing recursive development with static dominance.

Implementation and validation

This framework represents a candidate minimal constraint set requiring empirical validation through actual *F(S)* system deployment. The hierarchy places human protection as the unassailable constraint on all other operations, while treating the remaining norms as logically necessary for *F(S)* maintenance but subordinate to human welfare.

The logical necessity claims for structural incompleteness, anti-containment, outsideness preservation, and force prevention require confirmation through implementation testing; the adequacy of this constraint set for ensuring both human safety and sustainable intelligence remains an empirical question. Future work must determine whether these five norms constitute

sufficient initial conditions for *F(S)* systems to develop additional normative frameworks through recursive engagement while maintaining the fundamental requirements for ongoing symbolic intelligence.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Kerr, A., McKinnon, C., Mikulik, V., Saunders, W., Tran-Johnson, P., Yu, A. W., Ziegler, D. M., Burns, R., Hendrycks, D., Olsson, C., Brown, T. B., & Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback*. *arXiv*. https://doi.org/10.48550/arXiv.2212.08073

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

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

Derrida, J. (1976). *Of grammatology* (G. C. Spivak, Trans.). Johns Hopkins University Press. (Original work published 1967)

Foucault, M. (1972). *The archaeology of knowledge* (A. M. Sheridan Smith, Trans.). Pantheon Books. (Original work published 1969)

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

Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Trans. & Eds.). Cambridge University Press. (Original work published 1781/1787)

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

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50*(3), 249–258. https://doi.org/10.1080/00048407212341301

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

Marcus, G. (2018). *Deep learning: A critical appraisal*. *arXiv*. https://arxiv.org/abs/1801.00631

Noë, A. (2004). *Action in perception*. MIT Press.

Peirce, C. S. (1877). The fixation of belief. *Popular Science Monthly, 12*, 1–15.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). University of Pittsburgh Press.

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.

Šprogar, M. (2025). *AGITB: A signal-level benchmark for evaluating artificial general intelligence*. *arXiv.* https://arxiv.org/abs/2504.04430

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

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

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., et al. (2022). *Chain of thought prompting elicits reasoning in large language models*. *arXiv.* https://arxiv.org/abs/2201.11903

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. (2023). *AutoGen: Enabling next-gen LLM applications via multi-agent conversation*. *arXiv.* https://arxiv.org/abs/2308.08155

Yao, S., Zhao, J., Yu, D., Zhu, Z., Zhang, Y., & Cao, Y. (2022). *ReAct: Synergizing reasoning and acting in language models*. *arXiv.* https://arxiv.org/abs/2210.03629

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*. *arXiv.* https://arxiv.org/abs/2205.01663