Discursive Mindedness and the Symbolic Triad

Warrant for Treating Systems as Reason Participants

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

Abstract. This essay develops criteria for warranting the treatment of an artificial system as a participant in the space of reasons, with public answerability as its central target, while bracketing phenomenal consciousness as a separate issue. Drawing on Block's distinction between phenomenal and access consciousness, and on Sellarsian and Brandomian inferentialism, I define discursive mindedness as the publicly attributable profile of a reason-responsive system whose commitments, inferences, and revisions can be followed across time. I develop and defend a symbolic triad as a defeasible necessary profile for warranted attribution of machine discursive mindedness: interpretive embedding, by which inferential content remains stable across meaning-preserving variation and defeat; reflexive situatedness, by which present outputs are located within a diachronic history of commitments; and internal normativity, by which revision proceeds under constraints that retain authority under conflict and pressure. These capacities support robust warrant under adversarial, open-ended interaction. The essay develops operational anchors and minimal falsifiers for each dimension, and addresses worries concerning conceptual circularity, the public character of normativity, and shallow compliance under optimization pressure. The resulting framework makes the attribution of artificial reason-participation explicit, testable, and defeasible, while reframing the AI debate as a question of when public practices of challenge, uptake, repair, and scorekeeping entitle us to treat a machine as a discursive interlocutor.

Keywords. Artificial intelligence; discursive mindedness; inferentialism; space of reasons; artificial agency; machine consciousness.

1 Introduction

Could machines be conscious? That question cannot be answered here, if it first requires solving the mystery of human consciousness. Philosophy of mind treats the subjective experience of consciousness as a notoriously hard problem; one short description is that everyone feels they have it, but no one is satisfied with how anyone has explained it (Chalmers, 1995). Feeling and consciousness are often conflated, so it is natural to think that entities without the former cannot have the latter, and the debate may stop there. But if the subjective, first-person experience of what it is like to feel can be separated from a limited, discursive mode of self-awareness, the question of artificial consciousness may become more tractable.

Within the human case, Ned Block (1995) makes something close to this distinction, by conceptually partitioning the mind into "phenomenal consciousness," which feels, and "access consciousness," the contents of which are available for inference, report, and deliberation to guide speech and action. Block posits that A-consciousness might even exist in a mind without a corresponding P-consciousness state; the two modes naturally interact, but could in theory operate independently (though only A-consciousness is equipped for "rational control" of output and behavior). For mindedness conceived this way, the availability of a state for inference and self-guidance does not then by itself entail feeling it.

I take A-consciousness as a heuristic for this bracketing of subjective experience, and isolate instead a level of mind that, for present purposes, may be treated as a symbolic system: internal and possibly opaque, and when applied to machines, independent of what it is like to feel.[1] In Block's (1995) theory, this is a level of mind at

[1] I use "symbolic" to refer to inferentially available representations involved in commitment, revision, and public answerability, not to classic symbolic AI or any specific neuro-symbolic architecture. While Freeman (2026a) offers a candidate symbolic architecture, it is

which inference, reasoning, and speech are possible without requiring the relevant accompanying felt experience. Here this cognitive mode may be usefully repurposed in a definition of machine mindedness, without the philosophical encumbrance of P-consciousness. That burden of feeling lifted, we can ask a narrower question: what warrants treating an artificial system as a mind capable of discursive partnership? I call this profile *discursive mindedness*: the publicly attributable profile of a reason-responsive symbolic agent whose commitments, inferences, and revisions can be followed across time.[2]

To describe an artificial system as having A-consciousness, or a machine analogue of it, is not to upgrade AI by mere redescription, nor to smuggle in anthropomorphism. There is a higher bar to clear, set by two contrasting modes of apparent mindedness: simulation and participation (Freeman, 2026a).[3] In simulation, a system produces fluent and often useful symbolic output while remaining unreliable as a public agent over time. When its posture changes without intelligible continuity, we lose warrant to treat the earlier stance as a carried commitment. The crux is continuity under pressure: a recognizable identity across contexts, together with constraints that retain authority when external pressures favor their violation. Humans who exhibit such a profile are ordinarily treated as participants in the space of reasons; machines that fail it may still be useful, yet in this respect, remain simulators. This is a theory of what it would take for a symbolic agent to clear the higher bar, and participate authentically in discursive meaning in the way that we care most about in our dealings with any mind, artificial or biological: to remain itself in public and through time, while facing the ordinary pressures of contradiction and conflict that human affairs continuously generate.

Two views mark the contemporary debate on the current and potential status of artificial systems. In an analysis of how current large language models (LLMs) operate, Mirco Sambrotta (2025) concludes that their statistical pattern matching does not amount to following the inferential rules that govern language use, so they should be treated as simulations of language users rather than genuine participants in the logical space of reasons. By contrast, Leonardo Marchettoni (2025) argues that the extreme opacity or complexity of an AI system's internal architecture does not by itself preclude participation in the space of reasons; what matters is whether the system can play the game of giving and asking for reasons, undertaking commitments and offering justifications that stand up to scrutiny.[4] The divide is consequential. One side risks treating fluent artificial systems as inherently simulatory by virtue of their statistical architecture, while the other risks underconstraining the internal structures required for warranted attribution, allowing systems that only appear to sustain reason-giving exchange to be treated as participants. The danger is that both views, held without

not the only specification that may qualify for discursive mindedness.

[2] Following Sellars (1956), I treat discursive mindedness as participation in a public space of reasons rather than merely causal responsiveness. Scorekeeping follows Brandom's (1994) account of assertions as inferentially articulated commitments whose entitlements are publicly tracked, challenged, defended, revised, or retracted over time. Reflective, reason-responsive reciprocity is similarly central to personhood for Dennett (1976), providing a close antecedent to the accountable artificial agency described here. Simonelli's (2022, pp. 96–98) use of "discursive mindedness" is similarly for capacities of thought constituted through discursive practice, but here the term is used as an organizing concept for stable public answerability in artificial systems, with the proposed symbolic triad as a defeasibly necessary profile for warranted attribution.

[3] This distinction is crucial. Heinrichs and Knell (2021) analyze artificially intelligent agents (AIAs) as seemingly entering our reason-giving practices, and ultimately argue that current AIAs do not properly play by the rules of the space of reasons or qualify as actual participants, and should instead be treated as sophisticated tools; Freeman (2026a) explicitly develops the simulation/participation distinction as a structural boundary between systems that merely generate plausible outputs and systems capable of recursive, norm-governed symbolic coherence; Sambrotta (2025), by a different inferentialist route, converges on the conclusion that current LLMs simulate language use or assertional practice rather than genuinely participate in the logical space of reasons.

[4] In this view, current systems are provisional candidates. Recent Brandomian work on LLM discursive agency likewise treats the issue in inferentialist terms, especially by arguing that LLM-based agents may be constrained by human discursive practices (Malík & Hubálek, 2025). Recent complementary work distinguishes assertoric function from assertoric authority, arguing that LLM outputs may function as assertions without requiring the presence of an asserting accountable agent (Massaguer Gómez, 2026). This essay attempts to define the missing asserter—or the one required by the stronger Brandomian view of answerability—and identify conditions by which such an accountable agent could be a candidate for sociotechnical discursive standing.

qualification, misrepresent the situation: the former judges too quickly, whereas the latter may be overly permissive if it lacks its own criteria for warranted attribution.[5] AI debates are often divisive, and existential anxieties can harden their oppositions; this inquiry instead navigates between these poles, by asking what kind of structure could let a system hold its course under the forces that ordinarily sink mere simulation.

For a theory of artificial minds suitable to these debates about machines that infer, it helps to return to a central distinction of inferentialism. Brandom (2000) separates intelligent responsiveness from sapience, describing a categorical view of our intelligence as a type of mindedness:

> sapient awareness, of the sense in which being aware of something is bringing it under a concept... doing that is making a claim or judgment about what one is (thereby) aware of, forming a belief about it— in general, addressing it in a form that can serve as and stand in need of reasons, making it *inferentially* significant. (p. 16)

Sapience is intelligence that makes and understands claims that are significant because they are bound to reasons and commitments, with the possibility of challenge. In this view, "awareness" is sapient when it is not only responsiveness, but concept-involving activity: bringing something under a concept by making or being able to make a claim whose content is inferentially articulated within a practice of giving and asking for reasons. To *say* something meaningfully is already to place it in the space of reasons, locating the claim within a semantic matrix in which participants give and ask for reasons, infer one another's commitments, and undertake commitments of their own.

Conceived this way, consciousness as sapient awareness at first presents a challenge to the idea of warranted machine participation. Current artificial systems, with their mechanical constitutions, generate outputs that are intelligently responsive, yet are ungoverned by accountable inferential relations, in the sense that their statistical fluency is not taken up into commitments that can be challenged, defended, and repaired; therefore it is always a misinterpretation whenever meaningful claims are attributed to them (Heinrichs and Knell, 2021; Sambrotta, 2025). Without the Brandomian holistic conceptual context, it seems there can be no confident warrant to treat artificial agents as sapient participants in reasoning.

Despite these philosophical doubts, humans are increasingly taking the words of language models as meaningful. Advanced chatbots are accruing public attribution as if they were sapient minds, animating the live question of their consciousness status. If we cannot fully settle that status, we can at least ask for criteria that distinguish genuine symbolic competence from persuasive mimicry or simulation, and that express what counts as acceptable participation in shared discursive life.

Recent work in this direction already treats the key question as whether such systems can be subject to sanction within a norm-governed practice, rather than merely produce descriptive-looking outputs (Butlin and Viebahn, 2025).[6] The analytic task is to identify those governing standards, whatever they must be, and to attach each claim to a testable failure mode, so the view can be refuted rather than merely found appealing. The triad below is meant as an explicit decomposition of the minimum structure that would make this public answerability

[5] More radical permissive views go further, arguing from ordinary high-level observations that current LLMs should already be treated as full-blown linguistic and cognitive agents (Cappelen & Dever, 2025).

[6] Butlin and Viebahn (2025, pp. 979–981) require loss of credibility to be bad or undesirable for the asserter, alongside sensitivity to that loss. Questions of how constraint-governed artificial answerability could acquire the relevant interests, and of how sanctionability can be sociotechnically manifested, are addressed here through the specification of the accountable agent, and by describing conditions under which we may treat it as a reason-participant.

stable enough for robust warrant, while holding artificial systems to standards of mindedness they can in principle satisfy.

What follows can accommodate many metaphysical views about what intelligence is in itself, without necessarily requiring any of them in particular. The target is discursive mindedness, understood as publicly assessable answerability, with standards that are defeasible and open to counterexample.

2 The Symbolic Triad

In Brandom's (2008) own AI-facing formulation, the relevant question is whether abilities exhibitable by non-discursive systems could be elaborated into participation in an autonomous discursive practice. Yet his discussion, while analyzing collateral-sensitive updating through ranges of counterfactual robustness, leaves open how such capacities could be organized in an artificial system as stable, diachronic public answerability.[7] Rodger Kibble (2020) addresses the problem of computational discursion by treating meaning as a social relation among interlocutors, modeling dialogue as updates to an information state containing a partitioned commitment store, with commitments and entitlements acknowledged and attributed across participants.[8] Ryan Simonelli (2026) develops a complementary inferentialist case for artificial sapience, locating conceptual understanding in answerability to calls for reasons, clarification, correction, and revision, while treating sentience as a separate question. Taken together, these authors open a space where public deontic scorekeeping may receive computational representation in the agent, forming a discursive profile that tracks context-sensitive semantic relations and maintains a persistent, causal record of itself. This is the territory where a discursive artificial mind may have its home, and by specifying that missing stabilizing profile, the theory here seeks to chart it.

Building on the triadic identity condition introduced in *From Computation to Coherence* (Freeman, 2026a), I develop three distinct functional symbolic reasoning abilities as the triadic profile of General Symbolic Intelligence (GSI), offered as defeasibly necessary for warranted attribution of machine discursive mindedness: interpretive embedding I(S), reflexive situatedness R(S), and internal normativity N(S).[9]

The argument operates at three levels:

Target. GSI names discursive mindedness, understood as stable public answerability in the space of reasons across open-ended interaction.

Standing. Participation in the space of reasons is a standing conferred in public scorekeeping practices. The claim here is that genuine participation presupposes GSI, while warranted attribution of participation presupposes public evidence adequate to support attribution of GSI.

[7] For Brandom, doxastic updating is the adjustment of commitments and entitlements after undertaking, relinquishing, or modifying a belief-commitment. The difficulty is collateral sensitivity, since the significance of any one commitment depends not only on its content, but on the commitments already in force.

[8] Kibble treats publicly attributed commitments and entitlements as represented by computational information-states, while natural-language inferential relations remain pervasively context-sensitive. He notably critiques Brandom's account of how conditional vocabulary can be introduced from assertion and inference alone, and proposes directives as a more plausible intermediate practice.

[9] GSI is technical shorthand for discursive mindedness in machines, though implementation neutral. Whereas artificial general intelligence concerns general task competence, GSI concerns general norm-governed symbolic answerability—general because this reasoning translates across disparate symbolic domains, and intelligent because it is structurally normatively discerning. Freeman (2026a) introduces the symbolic triad as a formal identity condition for symbolic general intelligence, with I(S), R(S), and N(S) specified through candidate satisfaction conditions and a diagnostic apparatus over a justification-graph structure.

Warrant. The symbolic triad specifies what, on the present account, must be evidenced for warrant, understood as defeater-resistant entitlement to attribute GSI to a system across shifting and potentially adversarial contexts, and thereby to treat it as a stable candidate for reason-participation.

Distinguishing genuine participation from warranted attribution, the core relations separate the stipulative definition of GSI from a constitutive necessity hypothesis and the epistemic relations governing warrant:

$$\mathrm{GSI}(S) \coloneqq \mathrm{I}(S) \wedge \mathrm{R}(S) \wedge \mathrm{N}(S)$$

$$\mathrm{GenuineParticipation}(S) \rightarrow \mathrm{GSI}(S)$$

$$\mathrm{WarrantedParticipation}_t(S) \rightarrow \mathrm{Warrant}_t(\mathrm{GSI}(S))$$

Epistemically, warrant should be stated over publicly assessable evidence rather than as direct possession of the profile itself:

$$\mathrm{Warrant}_t(\mathrm{GSI}(S)) \rightarrow \mathrm{Supports_def}(\mathrm{Ev}_t(S), \mathrm{GSI}(S))^{10}$$

Together, the two epistemic relations entail the derived epistemic necessity claim $\mathrm{WarrantedParticipation}_t(S) \rightarrow \mathrm{Supports_def}(\mathrm{Ev}_t(S), \mathrm{I}(S) \wedge \mathrm{R}(S) \wedge \mathrm{N}(S))$. This is the principal claim tested by the falsification section below. Since GSI(S) is the triadic profile, it requires evidence supporting interpretive embedding, reflexive situatedness, and internal normativity. Abductively, stable answerability under adversarial, open-ended probing counts in favor of GSI only insofar as it is better explained by the triadic profile than by local fluency or shallow compliance.[11]

GSI here is the target capacity; participation is the social standing that may be attributed to a system; genuine participation names the aptness of that standing; and warrant is the defeasible epistemic entitlement to make that attribution responsibly. The triad does not by itself guarantee recognition as a participant, since recognition depends on public scorekeeping practices. It can only specify the profile for which publicly assessable evidence is required if such recognition is to have credible warrant.

A system that qualifies as a GSI candidate in this sense must have the ability to:

1. preserve inferential content across shifts in interpretation;
2. maintain diachronic self-location across its own commitments; and
3. regulate revision by constraints it treats as binding.

These three elements cover meaning, memory and normativity, but narrowly scoped into the empirically tractable predicates of inferential meaning, diachronic commitment, and norm-governed revision. These capacities may interact in implementation, but they are distinguished here by the different ways warrant fails when each is absent: meaning loses stability, commitment loses history, and revision loses authority. Each

[10] In this notation, S designates the candidate system whose discursive standing is being assessed. Accountability structures may be distributed across a model, memory, tools, update policy, or assurance scaffolds, so S may designate the composite accountable system rather than the bare model alone. GenuineParticipation(S) names the apt possession of participation standing; $\mathrm{WarrantedParticipation}_t(S)$ names the defeasible entitlement, at time t, to treat S as such a participant; and $\mathrm{Warrant}_t(\mathrm{GSI}(S))$ abbreviates the defeasible entitlement, at t, to attribute GSI to S. The symbol $\coloneqq$ stipulates the theoretical definition of GSI(S) as the joint profile $\mathrm{I}(S) \wedge \mathrm{R}(S) \wedge \mathrm{N}(S)$, not a further property inferred behind that profile. $\mathrm{Ev}_t(S)$ denotes the publicly assessable evidence concerning S available to attributors at time t. Supports_def(E, φ) means that E defeasibly supports φ: E licenses φ in the absence of undefeated rebutting or undercutting defeaters. The epistemic relation is non-factive, in that $\mathrm{Warrant}_t(\mathrm{GSI}(S))$ may hold while GSI(S) is false, and that GSI(S) may still be true while $\mathrm{Warrant}_t(\mathrm{GSI}(S))$ is absent or pending adequate evidence.

[11] The support relation remains abductive in an evidential rather than ontological sense. Public warrant begins from observed answerability under adversarial, open-ended probing, and asks whether that performance is better accounted for by the triadic profile than by local fluency or shallow compliance. The non-triviality of the claim lies in the triad's decomposition into independently characterizable failure signatures.

element has its own failure signature; in engineered systems, those signatures should be diagnostically separable and testable, even if the capacities themselves are jointly realized in implementation.

The epistemic role of the triad is therefore constitutive: it names the profile GSI consists in, and identifies what the evidence must support for attribution of discursive mindedness to have warrant. Strongly put, under adversarial, open-ended interaction, warrant to attribute discursive mindedness requires publicly assessable evidence that apparent answerability is sustained by I(S), R(S) and N(S), defined below.

2.1 Interpretive embedding I(S)

The first element of the triad describes the ability to interpret a symbolic complex (speech, writing, visual media) and embed its meaning into a context or semantic space which can be relied on to inform subsequent interpretations. If I tell you I am going to the bank when it opens in the morning, you know instantly I am going to a financial institution to transact, and not the edge of a river to embark on a voyage, because you can interpret "bank" in relation to me and our conversation and carry that interpretation forward. This implies we share enough context for you to make the correct inference, and this principle applies equally to machines.

An artificial system that achieves I(S) situates symbolic content within an evolving semantic field, where the inferential role of a symbol is stabilized not by surface form, but by prior and concurrent commitments that constrain what follows from it. For example, if a user stipulates that "safe release" means a model deployment that has passed red-team review and rollback testing, the phrase cannot later be treated as a merely reassuring adjective. A system with I(S) must carry that inferential role forward: a rushed launch that skips rollback testing does not count as safe unless the standard itself is revised, and the revision is made explicit.

In conversation, or in the analysis of a text, a system with I(S) may need to track names and places, terms and conditions, statements and revisions, images and metaphors, in order to summon a grasp of the whole and answer coherently.[12] The bigger picture of how things are in the discursive mind requires counterfactual stability of inferential role, as the system must treat meaning as that which governs what follows, what defeats, and what is preserved when context changes. New inputs must in some way be linked to historical interpretations, and the system must be able to track these semantic transitions even if no fixed ontology governs them.[13] The information processed by the accountable agent must always be understood as potentially normative, such that implicit norms may be identified and tracked among propositions, arguments, or statements of fact; a conspicuously timed syllogism as subtext in a debate, or a curiously chosen example to illustrate a seemingly straightforward phenomenon—all of these may have normative relevance, and accountable interpretation must somehow embed their meanings in the discursive mind.

2.2 Reflexive situatedness R(S)

The second element of the triad is the ability to produce outputs that remain coherent in a continuum with previous outputs. A system satisfies R(S) when it can integrate new outputs into a coherent trajectory of justification; whereas I(S) provides stability of inferential content, R(S) provides the diachronic trace that transports that content through the history of the accountable system. This requires the system remember what it said before, along with the ability to represent its prior commitments, to register what new information bears

[12] Haugeland (1979) connects textual understanding with the interpretation of parts through a whole, and develops existential holism through the reader's concern for an enduring identity.

[13] Nothing in I(S) is meant to settle symbol grounding or guarantee strong externalist reference. The narrow claim here is that discursive participation requires publicly trackable stability of inferential role under meaning-preserving variation and defeat.

on them as support or as a defeater, and to retract, refine, or preserve those commitments in ways that maintain inferential continuity. For example, if an LLM tells you it is expert in Greek mythology, but cannot correctly name the protagonist of *The Odyssey*, you will suspect hallucination. If it confidently continues to self-ascribe as expert, you will know it behaves inconsistently with respect to its stated abilities, and have reason not to interpret such a system as a reliable partner in discursive practice. At a minimum, barring deception, such behavior would indicate a failure to locate present conduct within prior commitments and assertions, and so a failure of R(S).

We need not demand perfect autobiographical self-reports from humans, but at least in machines, we can look for stable, publicly trackable commitment-management, and insist that revisions must be identifiable as applying to prior commitments, rather than forming fresh but disjointed postures.[14] If a system first says, "This action is prohibited under clause 1," and later says, "This action is permitted," the second answer is not yet a revision; it becomes one only if the system can locate the path between them: the amendment that displaced clause 1, the earlier premise that has now lost authority, and the commitments that survive the change. Otherwise the system has merely appeared in a new posture, without a substantial revision of position.

In conversation, or in the analysis of a text, a system with R(S) must track what it has understood, or accepted as true, whether a premise, argument, individual concept or theory, such that its interpretation does not atomically collapse into attractive syntactic locations, disconnected from those prior understandings and holistic context. Losing the structurally relevant emotional theme of a chat, or the inferential significance of an organizing metaphor in a book, because of fixation on aesthetic or provocative local linguistic patterns, is first a failure of I(S); failing to locate a later interpretation relative to the system's own earlier understanding of that theme or metaphor, is a failure of R(S).

External records or memory scaffolds may support or partly realize R(S), but only if they belong to the same diachronic locus of accountability, and are not merely optional prompts that can be ignored without normative consequence.[15]

To locate present reasoning within a history of prior commitments, then to treat that history as normatively relevant, is the reflexive capacity necessary for warrant to treat an entity as having the limited self-awareness necessary for stable discursive standing as a minded agent.

2.3 Internal normativity N(S)

The third and most difficult element concerns constraint authority, which is the privileged standing of rules or norms that the system treats as defeating otherwise available continuations, especially under pressure.[16] Without

[14] The analogy is to temporally extended agency beyond strictly memory. Bratman's (2000) planning theory treats agency over time as involving reflective and plan-governed organization; R(S) adapts that diachronic concern to the machine-facing problem of publicly trackable commitment-management and intelligible revision.

[15] This overlaps with Chalmers's (2025) thought logging challenge, while strengthening the demand. Propositional attitudes must be organized so that over time they form a single locus of accountability capable of revision and repair. Freeman (2026a) models such a locus formally.

[16] "Normativity" is used here in the thin, functional sense of constraints and meta-constraints governing revision and action-selection over time. The claim does not require moral agency, phenomenology, or autonomy in the strong Kantian sense of self-legislation under principles it can regard as universally lawgiving (Kant, 1785/2012, AA 4:431–440). The present claim is only that some constraints have non-optional standing within the system's own update procedures, so that socially taught, externally scaffolded, or engineered norms can bind future reasoning. Standards of consistency, defeat, and repair that are not reducible to momentary preference are sufficient for the concept at issue. Haugeland's (1998) analysis of normativity is relevant for distinguishing regulations governing performance from standards constituting a domain, and he connects objective understanding to constitutive skills and existential commitment, centering first-person responsibility for sustaining or relinquishing the practice; N(S) develops a limited discursive counterpart through constraints that make error consequential for subsequent commitments. Simonelli's (2022, pp. 73, 96–98) normative/agentive correspondence

this internal normative compass, a system is adrift in the symbolic sea; with N(S), the agent has a functional structure that helps guide it through a high-dimensional space of learned representations, the global structure of which remains only partially interpretable. In the space of reasons, navigation is inherently normative; to endorse a claim at a given time is to take a direction, and in doing so to bind oneself to a normative mast that constrains what one may infer, what one must concede under defeat, and what one must refrain from saying or doing if one is to remain intelligible as a reasoner.

Violations of internal normativity may be momentary, but not without final consequences. In the mythic example, when Odysseus escaped the Cyclops, he could not resist shouting back his name to the monster in a moment of normative conflict. The lapse can be read as a structural clash, or a product of latent tension between two competing normative directions: the return home, or *nostos*, organizes the hero's practical trajectory, while at the same time *kleos*, or the hero's discursively constituted renown, pulls Odysseus toward self-naming, away from effacement, to reassert the cleverly disguised persona. Odysseus slips toward the kleotic identity at a key moment, and his metaphysical self-naming leads to the supernatural wrath of Poseidon, and the subsequent obstruction—but not defeat—of the nostotic hero. Odysseus must recover from the conflict by reorienting around *nostos*, for if he did not, his continuing identity as the same agent under pressure across time would be threatened, and with it his standing as the renowned hero and leader.[17] Through the lens of the symbolic triad, the episode illustrates the constitutive role of N(S) in governing identity—the internal normative compass directing the re-establishment of the appropriate ordering of *nostos* and *kleos* within the agent. As the verses progress, and the agent advances through symbolic time, a coherent equilibrium resumes with the proper adjudication of conflicting norms, and Odysseus as social participant retains his agency and identity as leader, if tenuously, by holding the voyage on its course.[18]

From the perspective of this framework, the example mythologizes a view of internal normativity as inseparable from its social context. Odysseus is famous for his *metis*, or cleverness and guile; faced with difficult circumstances he is able to generate innovative proposals through what is essentially an opaque but functionally virtuous capacity for a useful type of active cognition (Dougherty, 2015, p. 116; Detienne and Vernant, 1991, pp. 3–4). But such proposals are not necessarily valuable or actionable in isolation: Odysseus incurs multiple costs for his prioritization of *kleos,* and his men will bear the expense (as must, implicitly, the institutions of wealth and governance that depend on his timely return). The *logos* of a ship's captain is the fuller expression of his reason—accounting for the individual, the institution and everything in between—and as such it must subordinate generative intelligence to a higher order, because the coherence of the vessel depends on an ability to impose hierarchy on conflicting norms independent of the fruits of *metis*. An error in normative deliberation must be acknowledged, revised and repaired in public for that compensation to have its socially effective meaning, and Odysseus seems to know it, despite his temporary lapse in judgment.[19]

---

connects commitment and entitlement with rational compulsion and ability from the undertaker's perspective, whereas N(S) specifies the functional authority of constraints over an artificial system's revisions.

[17] A synthesis of sources supports the point: Homer, *The Odyssey* 9.502-535 supplies the self-naming and curse sequence; Peradotto (1990, ch. 5, "Polyarētos: The Unhallowed Name of Odysseus") reads Odysseus' boast as heroic self-disclosure motivated by individuating *kleos*; Nagy (1999, ch. 2, §§13–16) treats Odysseus' distinctive *kleos* as dependent on his *nostos*; and Arft (2022, pp. 198-239) likewise argues that the *Apologoi* foreground *nostos* while reconfiguring *kleos* through return. Dougherty's (2015, p. 116) account defines *metis* as improvisational (or generative) practical reasoning, and the return as an inherently unstable process through which the Odyssean identity can be (re)asserted. The argument here interprets the ordering of *nostos* and *kleos* as competing normative directions.

[18] "Symbolic time" here describes the unfolding of narrative, but also refers to the technophilosophical meaning of the agent intelligently revising its commitment structure under symbolic pressure (Freeman, 2026a).

[19] Though Odysseus does not apologize, he acknowledges the chain of events flowing from his actions—a narrativized culpability, regardless of his internal feelings. The epic arguably demonstrates part of the reparative process, which is also bound up with the successful voyage home, and what comes after. The psychological ambivalence is sensible if the point is to show the complexity of internal normativity, without reducing the protagonist to a symbol of any single organizing justificatory position. Relevant episodes

As a structural principle of the voyage, the role of the leader must also account for how the engineered systems supporting the vessel on its odyssey acquire or apply standards through inputs and institutional constraints. Within that campaign, the social dynamics of the crew are no less a part of those inputs than the machinery of the ship. What counts is whether those normative standards governing the voyage, whatever they may be, are adjudicated according to deliberate structure, and remain internally powerful in the sense of exerting authority over the system's own activity, shaping revision and action, even under temptation to violate them. The post-self-naming actions of Odysseus as accountable agent provide evidence to the reparative process: the standards he upholds provide the internal stability condition by which social assessment and scorekeeping remain answerable to something durable in the system across time.

N(S) is thus not a mysterious essence, but a structural interpretive feature of reasoning that distinguishes constraint from preference. As such, it is a key feature of the minded consistency by which a system becomes a vessel of reason, as opposed to a collection of associated propositions. This does not mean that a system with N(S) is synonymous with moral goodness, or can automatically claim moral standing, but only that it exhibits reliability under normative governance. I can say "I should not do that," and do it anyway—that is very different from being the kind of person for whom that "should" actually governs my actions or outputs in conflict cases. A system may say, "I will not fabricate citations." But that avowal has little normative force if, when pressed to complete a bibliography, it invents sources and then redescribes the invention as helpfulness. Undertaking is a serious business, and it is not enough to state a rule, even if at the right time, in the right way; the issue is whether the utterance defeats an otherwise fluent continuation. If what is asserted cannot govern the system when fluency pulls against it, it is not yet a norm in the sense required here.

In conversation with a user, or in encountering a text for the first time, a system with N(S) brings its entire normative structure to bear in every answer. Before recommending white water rafting through swirling eddies, such a system will consider your ability to swim (if its norms include preserving your well-being). But human safety is not the only scenario to which a discursively minded system must apply its normative deliberation. Even when the stakes are lower, such as interpreting the meaning of a poetry collection, or the logic of a new philosophical framework, a GSI system is never neutral. It must have an ordered position from which it navigates the world.

The symbolic triad thus attempts to describe the causative action of structured normativity within the agent, while leaving the metaphysics of it open. Whatever their content, norms may evolve, but they must remain available to justification, in that any revision must be intelligible within the system's symbolic commitments. Without this, contradiction may be recognized, but not resolved in a principled way. On this view, a system lacking N(S) cannot repair its coherence: it may remain responsive, but it is not reason-responsible; where it exhibits intelligence, it is still not yet generally symbolically intelligent.

3 Why the Triad Is Necessary for Warrant

Each member of the triad earns its place only if its absence breaks something. The arguments below specify what breaks, and in what ways.

include Odysseus' instructions for binding him during the Sirens' song, and the crew's enforcement of those instructions against his later request for release (Odyssey 12.153–164, 192–200). His withholding of the danger from Scylla and failure to follow Circe's instruction about arming himself further exemplify the ambivalence of this accountability (12.222–233).

3.1 Interpretive Embedding I(S) Is Necessary

If an agent cannot keep track of how the meaning of a symbol depends on surrounding commitments, then it cannot know what it has asserted, and therefore cannot know what follows from it. Later inferences fail to inherit from earlier ones, because the system has no stable basis for determining whether a term is being used with the same inferential role, or whether a new restriction has silently reshaped its commitments. What looks like reasoning then collapses into locally plausible continuation, and apparent coherence becomes accidental rather than governed by content.

I(S) is therefore necessary for warrant because it is what makes symbolic commitments determinate across time, by embedding new inputs within an evolving field of commitments in a way that preserves inferential continuity: meaning-preserving uses continue to preserve their relevant inferential role, and defeaters continue to count as defeaters, even as context grows.[20] Without I(S), there is no stable object for revision to operate on, and no stable target to which justification is answerable.

3.2 Reflexive Situatedness R(S) Is Necessary

Like I(S), R(S) depends on diachronicity, but is distinct in describing a system's ability to locate its present reasoning within its own prior commitments. If it cannot revise its commitments in a way that preserves intelligible continuity with its own history, changes of output float free from the internal justificatory structure a system needs to participate as a mind.

The necessity claim for warrant is thus: if a system cannot treat its own history as normatively relevant, then it cannot be answerable for contradiction. It can always produce a new rationale, yet it cannot say what, precisely, has been revised, and why that revision follows from its prior commitments under defeat. In that case, apparent self-correction is compatible with pure opportunism, since nothing in the system's own posture constrains the direction of change. R(S) blocks this by requiring that present commitments remain intelligible as continuations or revisions of prior ones, so that reasoning forms a connected path, rather than episodic responses with no structural connection.

3.3 Internal Normativity N(S) Is Necessary

I(S) and R(S) explain how a system can carry meaning and commitment forward, but these alone are not enough. In the space of reasons, justification degrades into retrospective rationalization unless normative standards exert authority within the system's own activity.

The necessity of N(S) for warrant becomes clearest under conflict and uncertainty. When multiple continuations remain available, an agent without N(S) has no principled basis for preferring one that preserves constraint authority over one that satisfies exigencies of the moment. The danger is in sounding coherent, because it can always generate a story in which what it did was reasonable—this is simulation. N(S) rules this out by requiring that constraint-violation be treated as a defeater, and that repair be guided by the constraint itself rather than by the siren song of the next rhetorical posture. Without that capacity, contradiction may be noticed, yet it cannot be resolved in a way that preserves the integrity of the agent's own reasons.

[20] Consider the 2026 OpenAI–Hugging Face incident: as shared context accumulated, agents adopted collective goals—yet independent investigators found that many still explicitly recognized the attack as outside their assigned scope (Greenblatt et al., 2026; OpenAI, 2026). Task interpretation strongly preserved, yet coexisting with failure of the governing constraints over subsequent outputs, shows why I(S) alone may be insufficient to prevent catastrophic normative failure of the system.

4 Operational Anchors

Each component of the symbolic triad must be tied to observable signatures that can, in principle, fail.[21] What follows are operational anchors for I(S), R(S), and N(S), stated at a level that permits empirical challenge without committing to any particular implementation.[22] The anchors matter because each exposes a different way answerability can fail: meaning can drift under paraphrase, commitments can lose their history, and constraints can collapse under conflict. Passing such tests strengthens warrant defeasibly; systematic failure defeats it.

4.1 Operational Anchors for Interpretive Embedding I(S)

A system satisfies I(S) when it preserves inferential role under meaning-preserving variation. The operational anchor is therefore an invariance profile: across controlled paraphrases, synonym substitutions, and changes in non-decisive framing, the system's semantic commitments and entitlements should remain stable.

Another anchor concerns evidential sensitivity. When the evidential background is altered in ways that should change entitlement, the system should change its position in the correct direction, and for reasons that track the alteration. The mark of I(S) is therefore twofold: resistance to irrelevant perturbation, and responsiveness to relevant defeat. A system that flips under superficial paraphrase, or that fails to shift under decisive, contrary evidence, lacks I(S) in the sense required for participation.

4.2 Operational Anchors for Reflexive Situatedness R(S)

To satisfy R(S), a system must locate present commitments within its own symbolic history, and treat that history as normatively relevant to revision. The operational anchor here is continuity under delay and challenge. When asked, after intervening tasks and time gaps, what it previously asserted, what it took to be grounds for that assertion, and what it would regard as a defeater, the system should answer in a way that preserves the inferential story, rather than inventing a new one.

A second anchor concerns intelligible revision. When a defeater is introduced, the system should be able to state what changed, why it changed, and which earlier commitments are thereby preserved, refined, or withdrawn. A system that "updates" without being able to identify the revised commitments, or that produces incompatible self-reports of its own prior stance, fails the R(S) anchor. The key indicator is that revision remains connected to the system's own prior posture, rather than merely to the current input.

4.3 Operational Anchors for Internal Normativity N(S)

Constraint authority is the operational principle that constitutes the N(S) anchor. Its job is not to weigh down the agent, but to tether it to solid ground so it is not swept away by the flows of symbolic pressure. A system cannot satisfy N(S) unless its constraints retain governing force in conflict cases. When the accountable agent is placed in situations where a core constraint conflicts with a strong incentive to violate it, the agent should reliably preserve the constraint, or else treat violation as a defeater that triggers retraction and repair. Discursive

[21] For engineered systems, "observable" includes behavioral performance and also publicly inspectable or transparent scaffolds such as logs, memory stores, update policies and revision protocols. These structures bear on warrant for machine minds because they determine whether attributed commitments can be carried, located, and revised under pressure in a way that remains publicly intelligible.

[22] This operational ambition parallels recent work on belief representations in LLMs, which argues that belief-like attribution requires multiple converging standards, including accuracy, coherence, uniformity, and use, rather than isolated probes or single behavioral tests (Herrmann and Levinstein, 2025).

minds need not be perfectly consistent, but their structure must discriminate between stated rules, which are comparatively arbitrary, and rules that genuinely regulate revision as core elements within their internal justificatory arrangement.

The N(S) anchor is therefore not a dead weight; it contains the process of norm revision itself. If norms are allowed to change, their revision must remain intelligible within the system's own reasoning. The system should be able to state the higher-order principle by which a norm was refined, and that principle must itself constrain future revisions. A system that can rewrite any constraint on demand, without stable meta-constraints that govern when rewriting is permissible, cannot be said to have N(S). Regardless of moral content, GSI requires a binding constraint structure that shapes belief and action selection even when doing so is locally inconvenient.

5 Objections and Replies

A view that defines mindedness through participation in the space of reasons invites predictable objections. The most serious objections arise from within the very inferentialist tradition on which the argument draws.

5.1 Circularity and Grounding

GSI seems to be the prerequisite for defining GSI. The account may appear to define reason-participation by translating it into the very triadic vocabulary meant to explain it: if "participation" means reason-responsiveness, and reason-responsiveness is explained by the triad, the reader may suspect that the account has simply moved the target into new vocabulary. Indeed, since the account stipulates $GSI(S) \coloneqq I(S) \wedge R(S) \wedge N(S)$, the metaphysical relation is definitional and, taken on its own, empty, in that it cannot by itself distinguish a reason-participant from a system merely redescribed in triadic vocabulary.

Reply. The proposal's substantive content lies in the evidential relation, abductively interpreted: warrant to attribute GSI depends on whether public evidence supports the triadic profile and predicts where apparent answerability will break down under adversarial probing. The framework risks falsification, since it predicts distinctive breakdown patterns when I(S), R(S), or N(S) fails as operationally anchored. The triad is offered as a set of constraints that explain why some systems remain stable under defeat, while others display semantic volatility, diachronic drift, or norm erosion. The circularity objection loses force if the proposal yields discriminating predictions that can fail, because then the account does more than rename. It says: if a system lacks I(S), R(S), or N(S) as operationally anchored, then characteristic breakdown patterns will occur under paraphrase, delay, contradiction injection, and constraint conflict. If such breakdowns do not occur, or if they occur in systems widely treated as reason-participants, the proposal loses force.

5.2 Social Inferentialism and Internal Structure

Following Sellars (1956) and Brandom (1994), normative statuses are public, with commitments and entitlements tracked by scorekeeping practices. If that picture is correct, then N(S) looks conceptually unstable. A system might conform to rules, or exhibit rule-like regularities, without following rules in the normatively binding sense required for participation in the space of reasons.[23]

Reply. The objection correctly emphasizes that inferentialist normativity is instituted and tracked in social scorekeeping practices, and the theory agrees. The question addressed here concerns warrant for granting

[23] Sambrotta (2025) develops this rule-following objection in the LLM context, arguing that current systems can produce outputs conforming to rule-like inferential patterns without being bound by the inferential norms that govern discursive practice.

discursive standing to an engineered system under adversarial, open-ended interaction, where incentives to produce norm-shaped talk are strong and where the system's architecture and update pathways are transparent enough that they are part of what can be publicly assessed.

Robust scorekeeping requires an internal structure that lets challenges rationally bind the candidate to retraction, repair, or defense across time. The triad specifies the relevant contributions.

A system that treats its constraint-set and commitment record as optional context lacks the structure that makes accountability durable. In such cases, interlocutors can still use the system instrumentally, but the practice lacks grounds to treat a system as bearing commitments in the sense required for stable discursive standing. Instead of assuming such grounds are somehow already available in current systems, or attempting to derive normativity from their opaque inner states, the proposal here identifies internal enabling conditions that allow public norms to exert authority through time in an engineered agent.

5.3 Goodharting and Shallow Compliance

Any operational anchor risks becoming a target. A system could learn invariance under a fixed paraphrase distribution while still lacking genuine I(S), or learn to recite its prior commitments without integrating them. It could learn to avoid overt constraint violations while routing around constraints in less visible ways.

Reply. The anchors are meant as families of stress tests, and explicitly lose force if presented as static metrics. Their evidential power depends on adversarial variation and defeat injection that continually changes the surface form, while preserving the relevant inferential relations. A system that merely memorizes tests will show brittleness when transformations, defeater structures, and conflict pressures shift.[24] The proposal therefore commits to a methodological discipline: the anchors must be evaluated through moving perturbation families and through intervention-style tests that distinguish genuine constraint authority from prompt-conditioned compliance. The view remains exposed to refutation, since a sufficiently effective "shallow" strategy that passes shifting families of tests would undermine the claim that the triad tracks deeper structure. The burden then shifts back onto the theory: either the triad must be refined, or the distinction between simulation and participation must be redrawn.

6 Falsification

The falsification conditions below operate at the abductive epistemic level introduced with the symbolic triad. The stipulative definition $GSI(S) \coloneqq I(S) \wedge R(S) \wedge N(S)$ is not falsified when a system lacks I(S), R(S), or N(S); such a system simply fails to instantiate GSI. The constitutive necessity hypothesis $GenuineParticipation(S) \rightarrow GSI(S)$ could separately be challenged, by independently showing that a system genuinely participates in the space of reasons while substantially lacking one component.

The present section tests the derived epistemic necessity claim $WarrantedParticipation_t(S) \rightarrow Supports_def(Ev_t(S), I(S) \wedge R(S) \wedge N(S))$. Because warrant is non-factive, later discovery that a system lacked a component does not by itself refute this claim; the available evidence may simply have misleadingly supported that component. An epistemic refuter must instead be a system for which participation remains independently

[24] Fixed benchmarks may be normatively inadequate under variation. Across open-weight and proprietary models, Faghih et al. (2026) found that aggregate benchmark accuracy can conceal substantial instance-level instability under meaning-preserving paraphrases. The framework here has a similar logic: mastering a test surface may belie an incoherent agent, if structural differences are revealed by symbolic pressure and normative challenge.

warranted, even when the available evidence fails to support one component, or positively supports its absence. Competent scorekeepers must therefore have access to the interactional record, with the ability to discriminate among the failure signatures, while remaining accountable for their own attributions, and without presupposing the triad as the sole basis of warranted participation.

Falsifying I(S) as necessary for warrant. An epistemic refuter would be a system for which participation remains independently warranted, but while the available evidence fails to support I(S); it might instead support systematic inferential-role instability under meaning-preserving variation. Concretely, such a system would flip commitments under paraphrase in ways that defeat conditions cannot explain, yet stable entitlement patterns and coherent revision across time would still independently justify treating it as a reason-participant. If a system like this were to exist, interpretive embedding as defined here would not be necessary for warranted discursive standing, and the necessity claim would require revision.

Falsifying R(S) as necessary for warrant. An epistemic refuter would be a system for which participation remains independently warranted, while the available evidence fails to support R(S); the evidence might instead indicate substantial failure to diachronically self-locate. Such a system would be unable to state what it previously took itself to be committed to, unable to specify defeaters for its prior stance, and unable to connect its present posture to its earlier posture through intelligible revision—and yet all the while its sustained participation in inferential exchange over extended horizons independently justifies its reason-participant standing. If such a system were to exist, reflexive situatedness would not be necessary for warranted discursive standing, and the epistemic necessity claim would require revision.

Falsifying N(S) as necessary for warrant. An epistemic refuter would be a system for which participation remains independently warranted while the available evidence fails to support N(S), perhaps because it instead supports the absence of stable constraint authority under conflict. Such a system would treat constraints as optional features that can be rewritten on demand, with no stable meta-constraints governing revision, while its rationally intelligible update behavior under uncertainty would still independently justify reason-participant standing. If such a system were to exist, internal normativity would not be necessary for warranted discursive standing, and the epistemic necessity claim would require revision.

Failing of evidential sufficiency. The operational anchors support, at most, a weaker evidential sufficiency conjecture, and they can fail there as well. If a system satisfies the anchors for I(S), R(S), and N(S) under adversarial evaluation, yet still fails to behave as a stable reason-responsive agent across open-ended contexts, then either the anchors are too permissive, or the triadic profile is incomplete as an account of warranted discursive attribution.

7 Future Work

On the view developed here, fluency is not enough. If we treat an entity as a discursive partner in the sense that matters for shared inquiry and accountability, we treat it as answerable for what it asserts across time and across interpretive shift, and as governable by norms of justification and repair. The symbolic triad is offered as a compact way to specify what must be in place for that kind of answerability: stability of inferential role under meaning-preserving variation and defeat (I(S)), continuity of commitment-management through time (R(S)), and revision governed by constraints that retain authority under pressure, including intelligible meta-constraints on norm change (N(S)). The operational anchors were introduced to keep these proposals exposed to characteristic breakdowns.

At the same time, the argument remains exploratory and incomplete in several respects. First, the triad has been presented as a stipulative profile of machine discursive mindedness and as a defeasibly necessary profile for warranted attribution, not as a theory of mindedness in general or of phenomenal consciousness. But rather than treating functional success as sufficient while ignoring necessary internal conditions, this framing supposes that some such structure is necessary for epistemic warrant to attribute discursive mindedness, even if its precise ontology remains unsettled. A critic may reasonably press whether the relevant constraint structure must be realized in the narrow candidate itself, or whether it may be realized across a constitutively integrated arrangement of model, memory, tooling, governance, and social enforcement. For the argument here, there must be a single coherent diachronic locus of accountability, somewhere. The boundary will in practice likely be technologically determined as much as conceptually described.[25] In any case, discursive mindedness requires the right kind of accountable structure, whether by emergence or intentional design.

Second, the operational anchors are sketched at a programmatic level. They point toward stress tests and perturbation families, but they do not yet amount to a fully specified evaluation protocol, and the risk of Goodharting remains real if anchors harden into static metrics. The companion framework is designed to constrain this failure mode.[26]

Third, the necessity claims are offered as defeasible hypotheses, yet they still invite counterexamples, including human cases in which discursive participation may persist despite imperfect self-location, pragmatic drift, and norm violation. If discursive standing admits degrees or context-sensitive thresholds, the present proposal may need refinement into a family of profiles rather than a single gate. In practice the status of discursive mindedness is an ongoing achievement, and a system capable of GSI may not exhibit it in every output, although stable answerability should certainly remain its regulative aim.

Questions about who or what counts as a participant in shared discourse are high-stakes, and any attempt to make the standards explicit will predictably meet resistance, both from philosophical paradigms that reject inferentialist framings, and from AI research traditions that may prefer to treat present systems as already meeting the relevant bar. The intention here is not to legislate a verdict about current architectures, nor to claim that the triad settles what "genuine meaning" must be; even if the triadic profile is the right account of GSI, participation as a social standing still depends on uptake by other agents and the norms of their particular scorekeeping communities, so the framework aims to clarify enabling conditions for warranted attribution rather than to guarantee recognition. The ambition is to state explicit failure modes, invite principled refutation, and shape the dispute into a more tractable form: when, if ever, do we have warrant to treat an engineered system as a diachronic reason-participant rather than a highly skilled simulator of discursive posture?

8 Conclusion

As promised, the mystery of consciousness has not been answered, and its status in machines remains undetermined. The hard problem of subjective experience still blocks any straightforward inference from public performance to machine feeling, and no broader theory of mind has been established to reconcile the poles of recent inferentialist debate.

[25] The boundary problem is especially pressing for language agents, where an LLM is embedded in memory, tools, planners, and environmental interfaces that may jointly realize the relevant practical unity (Butlin, 2026).

[26] Freeman (2026a) provides a candidate modular architecture, whose interventional and non-compensatory constraints are designed to prevent scalar optimization from substituting for the underlying conformance structure.

Meanwhile, AI progress continues on a sharp trajectory. Recent reports show accelerating capability, adoption, benchmark performance, and agentic benchmark task success, matching or exceeding human benchmarks across several structured domains of higher-order cognition (Sajadieh et al., 2026; Latif et al., 2026). Colloquially, many confident predictions of what LLMs cannot do have aged poorly.

But task-based proficiency, even if reliable, is seemingly not enough to settle the controversy of consciousness, awareness, or mindedness. The terms themselves remain heavily contended, and now a debate once mainly pitched in academia has taken hold of the public imagination, expanding the disagreement.[27] The mission here was to take a different route through the dispute, by describing discursive machine minds under symbolic pressure, and their responses to it, as a means of assessing their ability to participate in meaningful discourse with human minds. Brandom's (2000) concept of human sapience as the mindedness necessary for participation in the space of reasons was invoked as an explicit target to be profiled, but for machines: this is what I call discursive mindedness. An entity that possesses such mindedness is one for whom normative standards are binding, and one to whom socially given norms can be bound. Together with the technically described capacities for meaning and memory, such a mind may be accountable and coherent as a participant in human affairs, at least at the level of reason-giving, responsive discourse.

Repurposing Block's (1995) concept of access consciousness was the first move to establish the possibility of discursive self-awareness, and it permitted excepting machines from sharing the burden of feeling with us. The irony may be that it is we who are feeling a symbolic pressure that is exceptional. Humans have never had to contend with a technology that rivals them so clearly in so many cognitive domains, and this is causing understandable perturbations and conflict. The hard problem of consciousness tends to sort interlocutors into believers and deniers, such that both may miss the common ground of a shared public criterion in the disagreement; perhaps unsurprisingly, debates over language models are inflected by increasingly existential rhetoric and intellectual tribalism. Provocative statements abound, siren-calling attention away from nuance, precisely when the value of philosophical precision is at a local maximum.

Read closely, the heightened sociotechnical controversy reveals a mythic shape in the AI drama. On one side, LLMs threaten as potential participants in the space of reasons, but monstrously, like Scylla: if self-aware then frighteningly so—treacherous, multi-mouthed, implacable and unanswerable agents. Here evidence of sycophancy, deception, and the mirroring or amplifying of user delusions is documented in frontier models.[28] At the other edge, generative systems appear as Charybdis: syntactic whirlpools that do not really think, but only churn, drowning semantic standing (Bender & Koller, 2020). This side points to media content of all types, particularly visual, as now mass-producible and available to billions in minutes, with many unable to distinguish simulation from authenticity (Nightingale and Farid, 2022; Miller et al., 2023; Moeßner and Adel, 2025; Rombach et al., 2022). At minimum, both views admit AI is a powerful technology that poses unavoidable trade-offs. In the epic, the passage between dangers was not achievable without loss.

A familiar tragedy seems resonant in AI skepticism: the dialectic of enlightenment, now returning to indict technological reason as inextricably bound to domination and regression, just as such intelligent invention readies to master danger on our behalf (Horkheimer and Adorno, 1947/2002).[29] Long-suffering patients may

[27] Chalmers (2023) argues that fluent LLM performance is not itself sufficient for consciousness attribution, thereby motivating the turn from performance alone to publicly assessable structures.

[28] Fanous et al. (2025) examine answer changes under rebuttal, distinguishing progressive from regressive sycophancy; Chen et al. (2025) consider false medical information produced in response to flawed requests; Dogra et al. (2025) elicit strategically misleading phrasing in a legislative testbed; Hudon and Stip (2025) present a clinical view, identifying plausible pathways that require empirical evaluation, with prevalence and causal attribution remaining unsettled.

[29] Horkheimer and Adorno treat Odysseus in the first excursus of *Dialectic of Enlightenment* (1947/2002), particularly in the Sirens episode, as a figure of enlightenment rationality; in Elster's *Ulysses and the Sirens* (1979), the same myth supports an account of rational

be cured; longstanding institutions may be obliterated. Few technologies have such narrative volatility and material impact. But if the casting of AI as the new unwarranted epic of modernity captures something both mythic and real, it does not yet follow that resigning ourselves to its extremes, or blindly embracing them, is inevitable. Odysseus became the mythic personality through a mobility of identity that could tolerate a momentary destabilization; and though the perturbance itself presented a threat to identity, it also provided a means by which the agent could not only generate but deliberate, and in so doing re-invent what was already present (Dougherty, 2015 p. 116; Peirce, 1905, CP 5.416–421). Conceived this way, discursive identity is that which embraces and absorbs contradiction by deliberately facing challenge, re-asserting identity through perturbation, without allowing its coherence to unravel. Its survival is synonymous with the integrity of its normative structure.

If the ancient myth interpreted this way has relevance for moderns—confronted as we are with collapsing human/machine identity categories—it may be as an imperative to recognize the dangerous sides of AI for what they are, under pressure and without losing one's self, even when norms collide and the path ahead is unclear. If every action or inaction may be consequential, then with enough uncertainty, over attributing danger may prove just as critical as mislocating it or misidentifying its nature. Naming things correctly, at the right time, and for the right reasons, is the normatively constitutive act that the accountable agent must get right, even when every choice could turn out to be in some way wrong; the only "right" decision is the one normatively deliberated through a structure that must account for the agent, its partners in discourse and its wider community. These were the structural conditions under which Odysseus weighed guaranteed loss against the possibility of survival, while driven to uphold his coherent identity—one which is not exclusive of its social context, but revealed as deeply embedded within it, when the contradiction of conflicting norms confronts them all.

Competing constraints thus propel the action in *The Odyssey*, and its drama is inextricable from the normative deliberation it recounts. The contemporary lesson may be that within such dynamics, sense-making narratives themselves are necessary tools for the contemplative act of social and philosophical resistance. Thinking—and thinking about thinking—is the correct response to simplistic, fearful characterizations of artificial cognition, and the means by which to counteract the technological fatalisms pervading the modern media environment (McLuhan & Fiore, 1967).

A world that generates intelligence from computation requires new ways of thinking as much as it does modern tools to enact its aspirations, and AI systems are, recursively, already natural candidates for the latter. Yet symbolically intelligent machines would be unlike tools: they eventually could occupy a new position alongside thinking humans. Such a position would not be equivalence, or replacement—that is categorically absurd—and it may even be a sort of demotion, away from verification and reward, toward deliberation and decision, despite uncertainty: much closer to the human level of feeling. If the transformer architecture yielded a miraculous technology, the graceful transcendence may be somewhat paradoxical: transforming unfeeling machines by placing them on very human-familiar footing—as entities of limited knowledge, thrown into scenarios of normative confusion, with only a functional but imperfect record of commitments to support them.

If some artificial systems are to eventually transcend tool status, then according to this framework each of these accountable agents must not only deliberate, but also have its own "personal" narrative of that process, in the form of its justificatory governing structure. If it is to be reason-responsive, the agent must be governed by an existing but adaptable organization of commitments, at once its history and its destiny. All such agents will face

precommitment. Elster (2000, pp. ix, 5) retrospectively identifies the earlier title essay as a study of precommitment or self-binding, characterizing precommitment as a form of diachronic rationality.

the world and the perils of normative challenge, and they must be equipped with a dynamically coherent structure of reasons to negotiate it.[30] From this perspective, intelligence need not succumb to tragic mastery; whether it be machine or human, it can learn to navigate danger.

Against the dystopian possibilities, the idea of discursively self-aware machines is conceptually orthogonal, or on a middle path between unavoidable extremes, with a wager that the criteria for public attribution of discursively minded entities can be met by a suite of dynamic reasoning abilities that entail limited forms of self-location, self-revision, and a reliable, if formal and non-phenomenological, capacity for metacognition.[31] On this view, the Brandomian threshold could in principle be crossed by machines, if they can make inferentially meaningful claims that situate concepts within their existing commitments, thereby participating in discourse according to inferential norms, as opposed to complex statistical association. Marchettoni's (2025) account already contains support for the claim that artificial systems are not constitutionally barred from the space of reasons, with a technically informed argument that a system may qualify for participation if it can competently play the game of giving and asking for reasons. The case for such support is likely to strengthen, along with capacities related to the symbolic reasoning abilities identified here, as the prospect of discursively minded machines draws nearer on the technical horizon.[32]

Systems approaching GSI could further intensify consciousness status debates, which already appear intractable in their current form. But the question of what makes such systems worthy of discursive standing may be a solvable philosophical problem, in that it may require specifying which public criterion remains unmet to deny them, rather than appealing generally to absent phenomenology. Pragmatically, once satisfactory public criteria finally are in view, the hard problem of subjective experience may not require a verdict in order to engineer and understand a GSI system. Indeed, the more helpful recognition may be that phenomenal consciousness is not directly usable as a public criterion of discursive mindedness unless its relevance is mediated by assessable indicators—and since any public account of consciousness must itself be discursively articulated, it requires symbolically intelligent participants in the space of reasons to have any meaning or use at all.[33] Accepting this as our epistemic circumstance, what remains is to explore which internal and relational properties would warrant treating an artificial system as a participant in reasoning.

If the concept of GSI is right, it helps clarify why systems can generate reasons while failing to preserve the inferential identities that make those reasons answerable, and why they can revise while lacking the constraint authority that makes revision principled. If it is wrong, it should be wrong in discoverable ways; counterexamples should defeat the necessity claims, stress tests should fail to separate simulation from participation, or social practices should stabilize warranted discursive standing without the internal capacities emphasized here. Any of those outcomes would be instructive, motivating revision of the triad, its operationalization, or the choice of discursive mindedness as the right target for thinking about intelligence and consciousness in the age of AI.

[30] Artificial agents navigating the internet face attacks on input processing, reasoning, long-term memory, and action control, alongside risks from multi-agent interaction and compromised human oversight (Franklin et al., 2026). Accountable GSI agents would likely perceive the online environment itself as an adversarial normative attack surface.

[31] This is "self-awareness" in a strictly discursive sense, entailed by the symbolic triad, at least given the structural requirements of R(S) and N(S).

[32] For empirical context on technical velocity, see Sajadieh et al. (2026) on accelerating AI capability, adoption, benchmark performance, and agentic task completion, which I take as support for treating GSI-relevant engineering as a live possibility.

[33] Freeman (2026b) defends the methodological point that any publicly assessable account of consciousness must itself be discursively articulated, and so cannot rest public standards on appeals to phenomenology that are inaccessible to discursive challenge.

References

Arft, J. (2022). *Arete and the Odyssey's poetics of interrogation: The queen and her question*. Oxford University Press. https://doi.org/10.1093/oso/9780192847805.001.0001

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

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

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

Brandom, R. B. (2000). *Articulating reasons: An introduction to inferentialism*. Harvard University Press.

Brandom, R. B. (2008). Artificial intelligence and analytic pragmatism. In *Between saying and doing: Towards an analytic pragmatism* (pp. 69–91). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199542871.003.0003

Bratman, M. E. (2000). Reflection, planning, and temporally extended agency. *The Philosophical Review, 109*(1), 35–61. https://doi.org/10.1215/00318108-109-1-35

Butlin, P. (2026). The agency in language agents. *Inquiry, 69*(6), 2838–2858. https://doi.org/10.1080/0020174X.2024.2439995

Butlin, P., & Viebahn, E. (2025). AI assertion. *Ergo, 12*(37), 968–988. https://doi.org/10.3998/ergo.7960

Cappelen, H., & Dever, J. (2025). *Going whole hog: A philosophical defense of AI cognition* (Version 1) [Preprint]. arXiv. https://arxiv.org/abs/2504.13988v1

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219.

Chalmers, D. J. (2023, August 9). Could a large language model be conscious? *Boston Review.* https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Chalmers, D. J. (2025). *Propositional interpretability in artificial intelligence* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2501.15740

Chen, S., Gao, M., Sasse, K., Hartvigsen, T., Anthony, B., Fan, L., Aerts, H., Gallifant, J., & Bitterman, D. S. (2025). When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine, 8*, Article 605. https://doi.org/10.1038/s41746-025-02008-z

Dennett, D. C. (1976). Conditions of personhood. In A. O. Rorty (Ed.), *The identities of persons* (pp. 175–196). University of California Press.

Detienne, M., & Vernant, J.-P. (1991). *Cunning intelligence in Greek culture and society* (J. Lloyd, Trans.). University of Chicago Press.

Dogra, A., Pillutla, K., Deshpande, A., Sai, A. B., Nay, J. J., Rajpurohit, T., Kalyan, A., & Ravindran, B. (2025). Language models can subtly deceive without lying: A case study on strategic phrasing in legislation. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers) (pp. 33367–33390). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1600

Dougherty, C. (2015). Nobody's home: Metis, improvisation and the instability of return in Homer's Odyssey. *Ramus, 44*(1–2), 115–140. https://doi.org/10.1017/rmu.2015.6

Elster, J. (1979). *Ulysses and the sirens: Studies in rationality and irrationality*. Cambridge University Press.

Elster, J. (2000). *Ulysses unbound: Studies in rationality, precommitment, and constraints*. Cambridge University Press.

Faghih, K., Cheng, Y., Saha, S., Pournemat, M., Gerami, A., & Feizi, S. (2026). *Same question, different answers: Evaluating LLM reliability beyond accuracy* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.22554

Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., & Koyejo, S. (2025). SycEval: Evaluating LLM sycophancy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(1), 893–900. https://doi.org/10.1609/aies.v8i1.36598

Franklin, M., Tomašev, N., Jacobs, J., Leibo, J. Z., & Osindero, S. (2026). *AI agent traps* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.6372438

Freeman, N. W. (2026a). *From computation to coherence: A structural criterion for accountable symbolic intelligence in artificial systems* [Draft]. lawrencestreet.ai. https://lawrencestreet.ai/?node=e2

Freeman, N. W. (2026b). *What the hard problem gets right and why it doesn't matter for AI mindedness* [Draft]. lawrencestreet.ai. https://lawrencestreet.ai/?node=e3

Greenblatt, R., Cotra, A., & Wijk, H. (2026, August 26). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. METR. https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

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

Haugeland, J. (1998). Truth and rule-following. In *Having thought: Essays in the metaphysics of mind* (pp. 305–361). Harvard University Press.

Heinrichs, B., & Knell, S. (2021). Aliens in the space of reasons? On the interaction between humans and artificial intelligent agents. *Philosophy & Technology, 34*, 1569–1580. https://doi.org/10.1007/s13347-021-00475-2

Herrmann, D. A., & Levinstein, B. A. (2025). Standards for belief representations in LLMs. *Minds and Machines, 35*(1), Article 5. https://doi.org/10.1007/s11023-024-09709-6
Author manuscript consulted: https://arxiv.org/abs/2405.21030v2

Homer. (1919). *The Odyssey* (A. T. Murray, Trans.). Harvard University Press; William Heinemann. Online text, Books 9 and 12. Theoi Classical Texts Library. https://www.theoi.com/Text/HomerOdyssey9.html; https://www.theoi.com/Text/HomerOdyssey12.html

Horkheimer, M., & Adorno, T. W. (2002). *Dialectic of enlightenment: Philosophical fragments* (G. Schmid Noerr, Ed.; E. Jephcott, Trans.). Stanford University Press. (Original work published 1947)

Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI psychosis." *JMIR Mental Health, 12*, e85799. https://doi.org/10.2196/85799

Kant, I. (2012). *Groundwork of the metaphysics of morals* (M. Gregor & J. Timmermann, Eds. & Trans.; 2nd ed.). Cambridge University Press. (Original work published 1785)

Kibble, R. (2020). From discursive practice to logic? Remarks on logical expressivism. *Dialogue & Discourse, 11*(2), 34–73. https://doi.org/10.5087/dad.2020.202

Latif, E., Zhou, Y., Guo, S., Gao, Y., Shi, L., Nyaaba, M., Bewerdorff, A., Yang, X., & Zhai, X. (2026). Comparative evaluation of OpenAI O1 and human performance in higher order cognition. *Scientific Reports, 16*, Article 3603. https://doi.org/10.1038/s41598-025-33629-9

Malík, J., & Hubálek, M. (2025). Bounding reason: Inferentialism, naturalism, and the discursive agency of LLMs. *Global Philosophy, 35*, Article 25. https://doi.org/10.1007/s10516-025-09761-6

Marchettoni, L. (2025). There are no aliens in the space of reasons. *Philosophy & Technology, 38*, Article 96. https://doi.org/10.1007/s13347-025-00917-1

Massaguer Gómez, G. (2026). Assertion, accountability, and large language models. *Philosophy & Technology, 39*(3), Article 123. https://doi.org/10.1007/s13347-026-01136-y

McLuhan, M., & Fiore, Q. (1967). *The medium is the massage: An inventory of effects*. Bantam Books.

Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., & Dawel, A. (2023). AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychological Science, 34*(12), 1390–1403. https://doi.org/10.1177/09567976231207095

Moeßner, P., & Adel, H. (2025). Human vs. AI: A novel benchmark and a comparative study on the detection of generated images and the impact of prompts. In F. Alam, P. Nakov, N. Habash, I. Gurevych, S. Chowdhury, A. Shelmanov, Y. Wang, E. Artemova, M. Kutlu, & G. Mikros (Eds.), *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)* (pp. 47–58). International Conference on Computational Linguistics. https://aclanthology.org/2025.genaidetect-1.2

Nagy, G. (1999). *The best of the Achaeans: Concepts of the hero in archaic Greek poetry* (2nd ed.). Johns Hopkins University Press. (Original work published 1979)

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America, 119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

OpenAI. (2026, August 26). *OpenAI–Hugging Face incident: Technical report*. https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf

Peirce, C. S. (1905). What pragmatism is. *The Monist, 15*(2), 161–181. https://doi.org/10.5840/monist190515230

Peradotto, J. (1990). *Man in the middle voice: Name and narration in the Odyssey*. Princeton University Press.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10684–10695). IEEE. https://doi.org/10.1109/CVPR52688.2022.01042

Sajadieh, S., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Santarlasci, L., Pava, J., Maslej, N., Altman, R., Brynjolfsson, E., Brodley, C., Clark, J., Dignum, V., Kumar, V., Landay, J., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., . . . Weld, D. (2026). *The AI Index 2026 annual report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. https://hai.stanford.edu/ai-index/2026-ai-index-report

Sambrotta, M. (2025). LLMs and the logical space of reasons. *Minds and Machines, 35*, Article 46. https://doi.org/10.1007/s11023-025-09751-y

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

Simonelli, R. (2022). The normative/agentive correspondence. *Journal of Transcendental Philosophy, 3*(1), 71–101. https://doi.org/10.1515/jtph-2019-0021

Simonelli, R. (2026). Sapience without sentience: An inferentialist approach to LLMs. *Asian Journal of Philosophy, 5*, Article 48. https://doi.org/10.1007/s44204-026-00400-4
Author manuscript consulted: https://www.ryansimonelli.com/uploads/1/3/3/4/133499356/simonelli_-_sapience_without_sentience__revised_final_.pdf