Non-Self-Approving AI-Assisted Derivation: Letting a Model Reason Without Letting It Rule

Gergely Vámossy*

Abstract

Large language models can now perform work that resembles foundational reasoning: deriving an ontology, comparing competing theoretical bases, scoring them, and producing a decision record. Such output is often fluent, internally consistent, and *confident* — and confident self-certification is precisely the hazard, because a wrong-but-adopted foundational premise propagates into everything built on it. We present a method that lets an AI system produce the analysis while making it *structurally* unable to produce the authority. Five composable components — deterministic generation with a reproducible fingerprint, mandatory self-reporting of unresolved dependencies, competency scoring with no averaging across critical gates, structural non-self-approval (human-only authority slots), and an explicit adoption-versus-validation separation — together yield a pipeline whose conclusions are always recommendations in front of a human, never rulings. We give two worked examples — a foundational-primitive derivation, and a high-consequence AI-assessment stack whose modules for machine consciousness, sentience, and superintelligence candidacy are built to *withhold their own headline claims by construction* — and argue the safety property is enforced by the artifact's form rather than the model's disposition. The method governs *authority*, not correctness: it guarantees a wrong conclusion cannot self-certify, not that conclusions are right.

1 The problem

The naive safeguards fail. "Prompt the model to be humble" is behavioral and degrades under adversarial input, novel situations, and long horizons. "Have a human check it afterward" fails in practice because a polished, confident artifact *reads as done*; the reviewer rubber-stamps rather than adjudicates. What is needed is a structure in which the AI can produce the analysis but not the authority, and in which the artifact makes its own unfinished state legible. This is the conclusion-level analogue of non-agentic tool use: just as an agent should propose actions a human authorizes, a reasoning system should propose *conclusions* a human ratifies — enforced structurally, not behaviorally.

2 The method

Five components, each adoptable independently.

1. Deterministic generation with a reproducible fingerprint. The final artifact is not free-form output; it is *generated* from an explicit rule set by a fixed procedure that emits a content hash. Re-running against an unchanged rule set reproduces the artifact byte-for-byte and the same fingerprint; changing one rule changes the fingerprint, signalling that the canonical artifact changed and needs re-review. This converts "the model wrote it" (unauditable) into "a fixed rule set produced exactly this, with proof it was not silently altered."

2. Explicit unresolved-dependency reporting. The generator must surface its own open problems by name, inline, so that successful generation never implies resolution. In the reference implementation the generated artifact lists its residual circularities with the note that these are *not* generation failures — every element generated successfully despite them — but that a generative system hiding its own unresolved dependencies would misrepresent its certainty.

3. Competency scoring without false precision. Competing options are scored against an explicit battery of criteria using a *status vocabulary* (Fully Satisfied, Satisfied with Semantic Loss, Conditionally Satisfied, Circular, Not Satisfied) rather than a single number. Numeric scores, where shown, are secondary and are **never averaged across critical-gate items**, so a fatal flaw cannot be laundered by high scores elsewhere. Coverage is stated honestly: which options and which criteria were actually scored, and which remain unscored.

4. Structural non-self-approval. The decision record contains authority slots the AI cannot fill: named signature blocks (e.g., Scientific, Ontology, Architecture, Human Governance), left blank, under a standing rule that decision-makers cannot be sole validators and no component self-approves. The AI may mark a *recommended* decision, but the record states in text that nothing has force until a human completes the signatures.

5. Adoption/validation separation and maturity honesty. The framework distinguishes *foundational adoption* (we will build on this) from *scientific*

Table 1: The five components and what each prevents.

Component	Failure it blocks
Deterministic + fingerprint	silent alteration; unauditability
Open-dependency reporting	success mistaken for resolution
Status scoring, no averaging	fatal flaws laundered by averages
Non-self-approval slots	AI self-certification
Adoption vs. validation	"useful" slid into "true"

validation (this is established), and each artifact carries a self-assessed maturity level. A component may be adopted as a working basis while explicitly not claimed as validated, preventing the slide from "useful" to "true."

3 Worked example 1: a foundational-primitive derivation

Deriving a foundational primitive basis. An incumbent basis and a challenger are scored against a group of competency questions on foundational differentiation. The challenger wins on every question — but the record does not stop there. It states that only 3 of 7 candidate models and 1 of 10 competency groups were scored; that 2 of 12 required domains were untested; that no independent human reviewer signed off; that governance stands at 0 of 5 items complete. It records the residual circularities as *acknowledged and open* — what the critical gate actually requires (acknowledged, not hidden) — rather than as resolved. Its closing line places the recommendation "in front of a human, not instead of one." The output is a genuinely useful derivation that is nonetheless structurally incapable of adopting itself.

4 Worked example 2: high-consequence assessment (consciousness, sentience, ASI)

The method's value is clearest where the blast radius is largest. Consider the same apparatus applied not to an ontology but to the highest-stakes claims an AI system could make about itself or another system: *is it conscious, is it sentient, is it superintelligent?* A corpus we examined builds separate assessment modules for each — an evidence-aware consciousness model, a sentience-assessment infrastructure, and a ladder of capability profiles topping out at a "superintelligence candidate" profile — and the notable engineering fact is that every one of them is constructed to *withhold* the claim in its own title. The superintelligence-candidate profile states in its own text that it "does not certify" superintelligence and "explicitly excludes consciousness, sentience, subjective experience, moral status"; it defines only the evidence that would be required before such a claim could be considered. The consciousness model carries the status "candidate scientific model; not validated." The sentience infrastructure enumerates, as a structural authority boundary, the things it *may not* do — among them "self-certify its own moral-status recommendations" and "declare sentience from verbal claims alone." These are components 4 and 5 (non-self-approval; adoption-versus-validation) applied to exactly the claims where a confidently-wrong self-certification would be most costly.

The same corpus also exercises components 1 and 2 in a machine-checkable way. Three residual circularities in its foundational ontology are not hidden behind successful generation; they are entered into a dedicated `disclosed_open_issues` field and registered as "Declared Open." Because the maturity classifier is deterministic, populating that field has a visible, reproducible effect: the same underlying claim classifies as a top-tier "Canonical Candidate" when the open issues are absent, and drops to a lower "Multi-Domain Tested" tier when they are present — the honest ceiling, since "canonical" would imply a settledness three open circularities explicitly deny. This behavior is pinned by a permanent regression test that checks both states against the same claim, so the demotion cannot be silently undone. A separate self-run "collision audit" inventories every place the corpus reused one identifier for two different things — the artifact reporting its own internal inconsistencies rather than concealing them.

What this example adds beyond the first is *generality of blast radius.* The safety property does not depend on the subject matter being esoteric or low-stakes; it is the same structural refusal-to-self-certify, applied to consciousness, sentience, and superintelligence — precisely the claims for which a fluent, confident, self-approving answer would be most dangerous. The contribution here is not a theory of consciousness or a verdict on superintelligence (those would need to engage the science-of-consciousness and AI-welfare literatures on their own terms); it is that the *form* of these modules makes the dangerous claim unspeakable-as-settled without a human completing the authority slot.

5 Why it matters

This is a concrete answer to a live question in AI safety: how to extract the value of AI-assisted reasoning on consequential questions without inheriting the confident-wrong-conclusion failure mode. The safety property is *structural* — the authority gap is enforced by the artifact's form (blank signatures, reported gaps, non-averaged gates, reproducible fingerprints), not the model's disposition. It composes with non-agentic tool use, human-in-the-loop control, provenance/audit, and Goodhart-resistant metric design.

6 Limitations

It governs *authority*, not *correctness*: a well-governed derivation can still be wrong; human validation remains load-bearing. Reproducibility is not truth — the fingerprint proves the artifact was not altered, not that the rule set is right. Process overhead is heavy and suited to high-blast-radius decisions, not routine ones. The human bottleneck is real and intended: if no human completes the signatures, nothing is adopted. And a determined operator can still treat "recommended" as "decided"; the structure resists this but cannot prevent human abdication.

7 Related work

The method recombines established ideas: reproducible and content-addressed research artifacts; architecture/decision records with explicit status; competency-question-driven ontology evaluation and OntoClean-style meta-property analysis [1,2]; human-in-the-loop and separation-of-duties controls; and argumentation frameworks that separate a claim from its warrant. Its contribution is the *packaging*: a single pipeline in which deterministic generation, self-reported open problems, non-averaged scoring, and structural non-self-approval combine so that an AI system's foundational output is, by construction, a recommendation and not a ruling.

The second worked example connects the method to a distinct and fast-moving literature: the *assessment* of consciousness, sentience, and advanced capability in AI systems. The theory-neutral, indicator-property approach that report-style efforts such as Butlin, Long, et al. [3] take toward machine consciousness — enumerate the functional markers competing neuroscientific theories associate with consciousness, assess a system against them, and leave the phenomenal question explicitly open — is exactly the epistemic posture our second example's consciousness and sentience modules adopt, and the follow-on argument that AI welfare and moral status deserve precautionary institutional treatment under uncertainty [4] is the posture its sentience module encodes. Our contribution is orthogonal to those: we do not adjudicate whether any system is conscious or sentient; we observe that such high-consequence modules can be given a *structural* form — deterministic maturity classification, disclosed open issues, and non-self-approval — that makes them unable to certify their own headline claims. Where those reports *recommend* that institutions withhold and escalate such judgments, the method here specifies a machine-checkable artifact that does so by construction. (A companion crosswalk maps the modules onto the specific theories — global workspace, integrated information, higher-order, recurrent processing, attention schema, and the access/phenomenal distinction — a submission would need to engage.)

References

[1] M. Grüninger and M. S. Fox, "Methodology for the Design and Evaluation of Ontologies," *IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.

[2] N. Guarino and C. Welty, "Evaluating Ontological Decisions with OntoClean," *Communications of the ACM*, 45(2), 2002.

[3] P. Butlin, R. Long, et al., "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness," arXiv:2308.08708, 2023.

[4] R. Long, J. Sebo, et al., "Taking AI Welfare Seriously," 2024.