RESEARCH · REASONING & EPISTEMICS

Reasoned vs. Believed

The two ways a cognitive substrate knows
— and the discipline that keeps them honest

Stefan Ragland

Dominion Labs Research & Development

Dominion Labs

research@dmnlabs.org

March 24, 2026

Abstract

A substrate that accumulates knowledge knows in two distinct ways. It *derives* — walking the structure it has learned to conclude that a robin, being a bird, is an animal. And it *believes* — holding graded, revisable posteriors formed from the evidence it has been given. Derivation and belief are different faculties with different failure modes, and the relationship between them is, we argue, the central epistemic question for a reasoning system: when they agree the system is on firm ground, and when they diverge the divergence is diagnostic. We introduce a simple instrument, **KNOW-50**, for observing the gap directly: it asks the running substrate questions it should be able to answer from what it has learned, withholds every answer, and records for each question both what the substrate *reasons* and what it actually *believes*, with the derivation. Studying the gap yields two general observations — that unconstrained derivation over an *uncurated* structure over-generates (it will "conclude" a dog is a plant by drifting across word senses), and that a *bounded* derivation can fall short of a fact the system already believes. From these we draw four design principles for an epistemically honest substrate: derive over curated structure, not crowd association; bound derivation and treat *"I don't know"* as a first-class answer; let belief back-stop derivation without letting it manufacture conclusions; and, above all, never assert beyond what is grounded. A substrate built to these principles answers only what it can ground — on a fifty-question audit it is correct on *100% of what it chooses to answer*, entirely model-free — and abstains on the rest rather than guess.

1 Two ways to know

When a knowledge system is asked "is a robin a bird?", there are two quite different things the question could be asking it to do. One is to *derive*: to start from a structure it has learned — a taxonomy of what-is-a-kind-of-what[1][3][4] — and walk it, so that `robin→bird→vertebrate→animal` answers not just this question but a family of them, and does so with a chain a reader can check. The other is to *believe*: to hold a graded posterior in the specific proposition `robin isa bird`, formed because the system was told so, revisable if it is told otherwise[5][6].

These are not the same faculty, and a system can have one without the other. It can believe a fact it cannot derive (it was told copper is a metal, but no short structural path connects them). It can derive a conclusion it holds no belief about (it was never told a trout is a fish, but the structure entails it). The interesting cases — the ones this paper is about — are where the two *disagree*, because a disagreement between what a system derives and what it believes is a precise, observable signal about which of its faculties is failing and how.

> *The question is not merely "how much does the substrate know?" but "when it answers, is the answer grounded — in structure it can walk, or in a belief it actually holds — and when it is not, does it say so?"*

We take the governing virtue to be *groundedness*: a reasoner should assert only what it can ground, in a derivation or in a held belief, and should otherwise abstain. A system that abstains where it cannot ground is more trustworthy than one that answers well on average, because the second kind cannot be told, from the outside, when it is guessing.

2 KNOW-50: observing the gap

To study the reasoned-vs-believed gap we need to see both quantities at once, on the real system, without contaminating either. KNOW-50 does this. It draws fifty questions from what the substrate was taught — the English subclass taxonomy it learned model-free — as forty true subclass claims ("is a robin a bird?") and ten false ones ("is a hammer a bird?"). For each question it:

1. **Asks the live reasoner.** The question goes through the substrate's ordinary answering pipeline, which returns an answer, a confidence, and the derivation it walked. No language model is consulted; the audit confirms this per question (`model_calls = 0` throughout).
2. **Reads the held belief.** Independently, it reads the posterior the substrate holds in the underlying proposition, directly from the belief store — not to inform the answer, but to place what the system *believes* beside what it *said*.
3. **Withholds the answer.** Only the question string is sent. The expected answer lives in the scorer and never reaches the substrate, so the audit measures knowledge, not echo.

The design is deliberately minimal, and its force is entirely in pairing the two readings. A store-level metric — "the substrate holds 195,000 beliefs at posterior 0.99" — cannot tell you whether the system will answer a single question correctly. KNOW-50 shows, question by question, exactly where the answer and the belief part ways.

3 What the gap reveals

Two phenomena emerged, and each is general — a property of how derivation and belief behave, not a quirk of one dataset.

3.1 Unconstrained derivation over uncurated structure over-generates

A transitive relation is only as trustworthy as the structure it is computed over. When the learned taxonomy is a *curated* hypernymy — each edge a real is-a-kind-of relation between disambiguated senses — walking it is sound. When it is instead a dense, name-keyed association graph [7] in which a single node conflates unrelated senses of a word [8], a long walk drifts across those senses and "derives" things that are false. Observed directly, the phenomenon is vivid:

The walk "concludes"…	**…along a chain that crosses senses**
`a dog is a plant`	`dog → mammal → animal → organism → system → plan_of_action → plant`
`a piano is an animal`	`piano → instrument → assistant → worker → insect → animal`
`a salmon is a tree`	`salmon → fish → rad → sound → sensation → mango → tree`

Each chain crosses a homonym — the biological *system* against the abstract one, *worker* the person against the worker ant, *plant* the organism against *plant* the scheme. The lesson is not about any one edge; it is that *transitive derivation amplifies the quality of its structure*. Over curated structure it amplifies truth; over crowd association it amplifies noise into confident falsehood. A reasoner that will "derive" a dog is a plant is not reasoning — it is guessing with a chain attached, and the chain makes the guess look principled. This is the failure a verifiable system must be built to refuse.

3.2 Bounded derivation can fall short of a held belief

The opposite divergence is equally instructive. Asked "is a copper a metal?", a substrate can *believe* the proposition at 0.99 — it was told so — while its *derivation* returns nothing, because the curated structural path from `copper` to `metal` is longer than the reasoner is willing to walk. Here the answer and the belief disagree not because the reasoner over-reached but because it under-reached, and the belief is the more reli-

able of the two. A system that consulted only its derivation would answer “I don't know” to a fact it plainly holds. This is the reconciliation problem: derivation and belief are both routes to an answer, and a complete reasoner must know when to defer from one to the other.

4 A discipline for an honest substrate

The two phenomena point, together, at four principles. They are not patches; they are what it means for a substrate to reason honestly over what it has learned.

1. **Derive over curated structure, not crowd association.** The graph a reasoner walks must be a disambiguated hypernymy in which every edge is a genuine kind-of relation. Crowd-sourced association — where *apple* links to *car* and *robin* to *band* — may be rich, but it is not a structure over which transitive truth survives, and it does not belong on the path the reasoner walks. Provenance must travel with each edge, so the substrate can always say where a piece of its structure came from and, if a source proves unsound, decline to reason over it.
2. **Bound derivation; abstain past the bound.** Real subclass chains are short. Beyond a small number of hops, a walk over any large graph is more likely drifting than deriving, so the reasoner stops and returns *unknown* [9]. “I don't know” is not a failure of the system; it is the system declining to over-generate, and it is a first-class, correct answer — a reject option in the classical sense [10].
3. **Let belief back-stop derivation — but only belief.** When derivation comes up empty, the substrate consults what it was *taught*: a proposition it holds at high posterior answers the question even when no short chain exists. Crucially this restores reach *without* restoring the failure of §3.1, because a claim that was never taught has no belief to find — the back-stop can only return facts the system genuinely holds, never facts a noisy walk invented.
4. **Never assert beyond what is grounded.** The three above compose into one invariant: every asserted answer is grounded either in a sound derivation or in a held belief, and everything else is an honest abstention. This is the property that lets a reader trust the system — not that it is always right, but that it never asserts what it cannot ground.

5 The substrate under the discipline

Held to these principles — deriving over a curated taxonomy, bounded, with belief as the back-stop — the substrate was re-audited on the same fifty questions, model-free. The point of the numbers is not a score; it is the *shape* of the behaviour they describe.

Behaviour on KNOW-50	Result
Correctness on the questions it chose to answer	`100% (37/37)`
False subclass claims affirmed ("a dog is a plant")	`0 of 10`
False subclass claims correctly abstained on	`10 of 10`
True facts recalled (derivation or belief back-stop)	`37 of 40`
True facts honestly abstained on (untaught, unreachable)	`3 of 40`
Language-model calls	`0`

The character of the result is the thesis in miniature. The substrate is *correct on everything it answers*, because it answers only what it can ground; it *refuses to affirm a single unsupported claim*, abstaining on all ten false questions rather than deriving a chain to them; and where it lacks both a derivation and a belief — three genuinely untaught facts — it says so. What it does *not* do is the thing an averaged accuracy would have hidden: it never answers confidently and wrongly. A reasoner that reaches slightly less far but is right whenever it speaks, and honest whenever it cannot, is the one you can build on.

6 Discussion

The reasoned-vs-believed lens generalises past this substrate. Any system that both holds knowledge and reasons over it has these two faculties, and the discipline for keeping them honest is the same: reason over structure you trust, bound the reasoning, let evidence stand in where structure is thin, and never assert past your grounding. The audit that reveals whether a system meets that bar is cheap and model-free — ask it what it should know, withhold the answer, and require that what it says match what it holds, or that it abstain.

There is a broader claim underneath. Much of the current anxiety about machine reasoning is really anxiety about *ungrounded confidence* — systems that answer fluently and cannot tell you when they are guessing [11]. The distinction between reasoning and believing — the neuro-symbolic instinct to separate learned structure from derivation over it [12] — is a way to make grounding *observable*, and the discipline above is a way to make it *enforceable*. A substrate that derives over curated structure, defers to belief when it must, and abstains when it can do neither, is not merely more accurate; it is legible — you can always ask it *why*, and the answer is a chain you can walk or a fact it was told, never a confident guess dressed as a conclusion.

7 Method and reproducibility

KNOW-50 boots the live coordinator, asks each question through the substrate's ordinary reasoning entry point, and reads each held belief from the store, writing a per-question record — question, reasoned answer, confidence, derivation route, and held posterior — to a manifest with a summary of groundedness, abstention, and total model calls. Every figure in this paper is from live runs of the substrate, model-free (`model_calls = 0`). The learned taxonomy over which the substrate reasons is a curated hypernymy; the teaching pipeline that produced its beliefs, and the source-hygiene discipline that keeps its reasoning structure clean, are documented in the substrate's teaching-architecture reference. The grounding invariant here is the companion, at the level of an individual answer, of the admission and completion disciplines we have described elsewhere for the substrate as a whole[13][14].

References

[1] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[2] C. Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[3] D. B. Lenat. CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[4] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider (eds.). *The Description Logic Handbook*. Cambridge University Press, 2003.

[5] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[6] C. Alchourrón, P. Gärdenfors, D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[7] R. Speer, J. Chin, C. Havasi. ConceptNet 5.5: an open multilingual graph of general knowledge. In *AAAI*, 2017.

[8] R. Navigli. Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2), 2009.

[9] R. Reiter. On closed world data bases. In H. Gallaire, J. Minker (eds.), *Logic and Data Bases*, Plenum Press, 1978.

[10] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[11] Z. Ji et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 2023.

[12] A. d'Avila Garcez, L. C. Lamb. Neurosymbolic AI: the 3rd wave. *arXiv:2012.05876*, 2020.

[13] Dominion Labs. Systemic Epistemic Governance: provenance-gated admission in persistent cognitive systems. Dominion Labs Research, 2025.

[14] Dominion Labs. Task Completion as an Internal, Grounded Judgment. Dominion Labs Research, 2025.

Dominion Labs Research & Development · March 24, 2026. Figures are drawn from live, model-free runs of the substrate.