The Trace and the System

Why your intelligence base layer has to be checkable, and what it takes to make one that is

A BEARING PARTNERS WHITE PAPER · AUGUST 2026

Radar is a nightly structural intelligence pipeline. It takes around 24,000 direct readings from the institutions that publish the world's machinery, holds them against a standing set of dated, falsifiable beliefs, and delivers a corpus you can interrogate rather than a report you have to trust. This paper explains how it works, what it costs, where it currently falls short, and how to test it in a week.

The problem is not what you are told. It is what you cannot check.

Most decisions that matter are taken about places nobody in the room will visit, using evidence nobody in the room collected. A manufacturer deciding where to hold inventory. An insurer pricing a route. A fund sizing an exposure. A department briefing a minister on a region it has no post in.

All of them are reasoning about systems they cannot observe directly, from material somebody else assembled. That material is the base layer, and base layers have an unpleasant property: their errors do not stay in them. They propagate upward into everything built on top and arrive wearing the authority of established fact. Worse, the relationship runs the wrong way round. The better the reasoning above, the more efficiently it converts a bad input into a confident decision.

Sound method on unsound ground does not fail loudly. It fails persuasively, and the failure is attributed to the decision rather than the data.

So the question worth asking of any intelligence supplier is not whether they are good. It is whether you could tell if they were wrong. For most of the market the honest answer is no. The large platforms monitor at enormous scale and combine that with substantial analyst networks, and their methods are proprietary and unpublished. You are asked to trust the firm and the credentials of its people, because there is nothing in the document itself you could check.

For many purposes that is fine. It stops being fine exactly where the stakes rise, which is where you most need to know whether an assessment will bear weight.

Why the usual answer no longer works

The traditional answer was that the reader checks. An assessment arrived in an organisation containing people who had spent careers on the subject and could tell a sound argument from a fluent one.

That assumption is weakening, and there are numbers on it. In a study of more than 48,000 people across 47 countries, two thirds of employees using AI at work said they had acted on its output without evaluating it. More than half had received no training in its use. Separately, work drawing on some seventy Australian practitioners across defence, industry, academia and government describes machine assistance raising the surface quality of written analysis while hiding the reasoning underneath, so that

work which reads well and is wrong in detail passes review by people who lack the depth to notice.

Meanwhile the technology has made confident wrongness cheap. Across commercially deployed AI systems, between 11 and 57 per cent of citations are fabricated or unsupported. More than half the citations in retrieval-based systems are attached after the answer was written rather than derived from evidence. When you instruct a model to follow a rule in a prompt, compliance runs around two thirds; when you ask it to check its own compliance, it does worse than a coin toss.

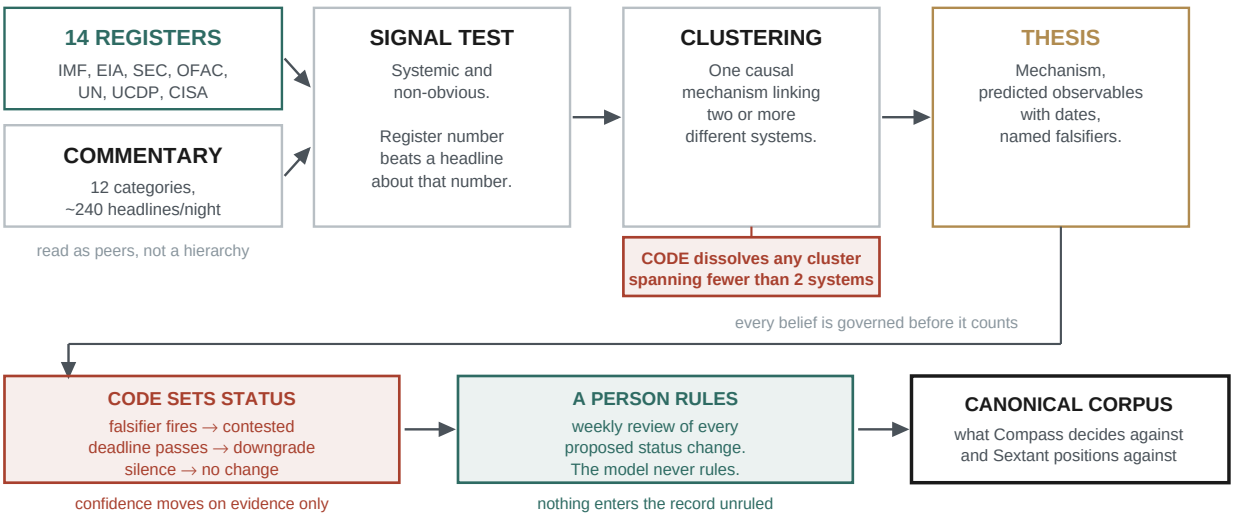
Put those together and you get the position most organisations are now in. The supplier cannot be checked from outside, the reader increasingly cannot check from inside, and the technology in the middle produces fluent output at a cost approaching zero.

A base layer nobody can check is not a foundation. It is an assumption with citations attached.

What we built instead

Radar is built on the assumption that a language model left to its own devices will drift, and that asking it nicely will not stop it. Nothing in the system relies on the model behaving. Four mechanisms make it structurally unable to misbehave without you seeing it.

HOW A READING BECOMES A BELIEF



Every checkpoint above is code that runs after the model answers, or a person. Neither is a prompt.
Machine-proposed theses may act autonomously in Sextant, where they are marked to market. They may not enter the record.

1. It measures before it reads

Radar pulls fourteen registers directly from the institutions that publish the world's machinery. The IMF and central banks on reserves. The US Treasury and the Bank for International Settlements on capital flows. The Energy Information Administration on energy. The SEC on corporate behaviour. OFAC and the UN Security Council on sanctions designations. UN OCHA on humanitarian appeals. Uppsala on organised violence. CISA on actively exploited software vulnerabilities. Space-Track on the orbital catalogue.

REGISTER	WHAT IT MEASURES	READ DIRECTLY FROM
Reserves	Central bank reserves and balance sheets	IMF, FRED, national central banks
Capital flows	Cross-border portfolio and banking flows	US Treasury TIC, BIS, OECD
Energy	Crude, gas, refined product, storage	US Energy Information Administration
Sovereign debt	Yields, credit spreads, issuance	FRED, TreasuryDirect
Currency	Spot rates and effective exchange rates	Market feeds, IMF
Trade and shipping	Trade volumes, freight, chokepoint transits	UN Comtrade, OECD, US Census
Corporate behaviour	Filings, disclosures, material events	SEC EDGAR
State action	Sanctions, policy actions, central bank calendars	OFAC, EU, Federal Reserve
Demographics	Fertility, dependency ratios, structure	World Bank Indicators
Infrastructure	Generation capacity, submarine cables	Ember, EIA International, TeleGeography
Conflict	Organised violence, terminations, peace agreements	Uppsala Conflict Data Program
Cyber	Actively exploited vulnerabilities	CISA Known Exploited Vulnerabilities
Space	Launch and catalogue tempo	Space-Track, US Space Force
Institutional	Designations, listings, humanitarian appeals	OFAC SDN, UN Security Council, UN OCHA

Calling those endpoints is easy and is not the product. Each register carries a reference file, and the fourteen run to about 11,650 lines of encoded analytical judgement. Those files decide what direction of movement counts as pressure, so a quote-convention error cannot silently invert a reading, and so that escalating violence and a signed peace agreement are never summed into a meaningless zero. They tier the publisher rather than the number, from the strongest statistical agencies down to countries under IMF censure. They define cohorts by measured behaviour rather than intuition: Asian energy importers move together on reserves at a correlation of about +0.39 and are scored as a group; oil exporters, which anyone would group by instinct, measure about +0.01 and are explicitly marked not to be scored.

They also handle silence. Every series declares how often it should publish. A monthly release that has not arrived by day 45 is not a hole in the data, it is a signal with a stated reason, and it is only raised while neighbouring series are still reporting, so an outage is not mistaken for a country going quiet. The rest of the industry treats missing data as a defect to be estimated around. When a statistics office stops publishing, the interesting fact is not the missing number. It is the silence, and who chose it.

2. The rules run in code, not in a prompt

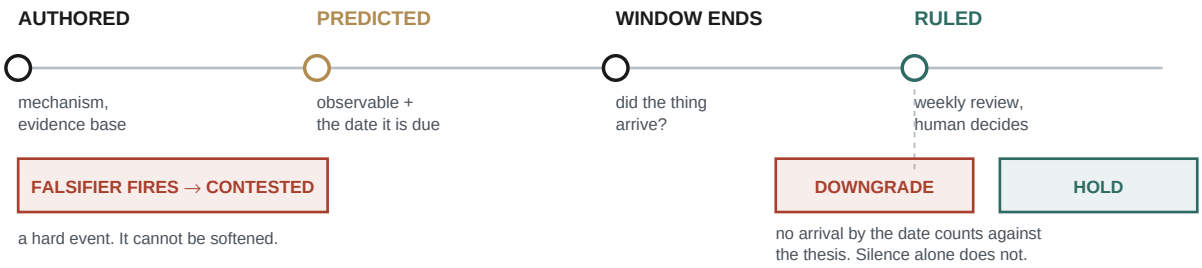
Radar defines a finding as one causal mechanism linking signals from two or more different systems. That definition is in the prompt, and it is also a function that runs after the model answers. Code recounts the systems, removes duplicates, and dissolves any grouping that fails the bar. The definition proved unreliable when carried by instruction alone, which is recorded in the source as the reason the check exists.

The daily brief goes further. Its structure is assembled in software before a word is written, and the model is handed one item at a time. It never sees the full set, so it cannot re-rank the day, merge two findings, drop one, or introduce a fact. A separate gate then scans the finished prose for jargon, unglossed acronyms and bare percentages, and reports each violation with the offending sentence.

3. Beliefs carry deadlines, and only a person can move them

The unit of analysis is not the signal. It is the thesis: a standing structural claim carrying a mechanism, an evidence base, a set of predicted observables each with a date by which it should have appeared, and a set of falsifiers describing what would prove it wrong.

A BELIEF WITH A DEADLINE ON IT



Confidence moves on evidence and only on evidence. A falsifier that fires forces the thesis into contested and that cannot be softened. A deadline that passes without the predicted thing arriving forces a downgrade. Silence is explicitly barred from counting as decay, because a quiet week is not an argument.

And the model does not get to change its own mind. It judges content; status is set in code and ruled on weekly by a person. We ran fully autonomous nightly revision, and stopped. Machine-proposed theses are still allowed to act autonomously in Sextant, our trading layer, because there they are marked to market and pay for their own errors on a schedule nobody controls. Nothing enters the canonical record that way. Autonomy is granted where consequences are measured and withheld where they are silent. The cost is latency of up to six days, and we accept it.

4. The receipt is printed by the till, not written by the cashier

Radar ships with an interrogation surface. You can take any conclusion and walk it backward to the readings and records underneath it, or forward to what the thesis predicts next.

Every load-bearing claim carries one of four tags. RR means it is in the corpus and provable. CR means a real external source was fetched during that exchange, carrying the address the fetch returned. RI means an inference whose evidence is in the corpus but whose conclusion is not. CI means reasoning with no corpus anchor at all. Two tiers are provable and two are honest judgement, and you can see at sentence level which is which.

The model does not write these references. Each retrieval tool builds the reference in code from the actual record and hands it over, and the runtime then checks it against the loaded corpus. An invented reference does not slip past. It renders struck through, in front of you. A link may only appear if that exact address was returned by a real fetch in that exchange, so a source the model merely remembers is demoted to unanchored reasoning with no link.

A SPECIMEN. EVERY CLAIM CARRIES ITS OWN STANDING.

Brent fell almost 9 per cent last week, to $72.60 RR reg:REG-03#BZN26#2026-06-28 and the wet freight benchmark dropped close to 40 per cent in the same window RR reg:REG-06#BWET#2026-06-28 , which reads as the market pricing a fast reopening RI paired price collapse vs the cluster's supply-shock framing . But the physical picture in the corpus, roughly 80 mines and 1,550 stranded vessels RR clu:CL-2026-06-28-7 , says restoration lags the diplomacy RI member records against the price move . This rhymes with insurance-driven rerouting outlasting past ceasefires CI general supply-shock pattern .

RR In the corpus. Provable, validated at runtime.

CR External source fetched this turn, carrying the address returned.

RI Inference. Evidence is ours; the conclusion is not stated in it.

CI Reasoning with no corpus anchor. Labelled so you can discount it.

This is also the answer to a practical problem in your organisation. Tiered citation does not require domain expertise to read. Someone who knows nothing about sovereign reserves can still see that a sentence is inference rather than measurement, and aim their scepticism accordingly.

What it holds, and what it costs to run

Figures below were read from the production system on 23 August 2026.

WHAT THE SUBSTRATE HOLDS · READ FROM PRODUCTION, 23 AUGUST 2026

263,443 primary observations	**1990** earliest baseline	**24,000** new readings a night	**14** instrumented registers
25 standing theses	**102** live falsifiers	**84** predictions that died	**$8.35** compute per night

The substrate holds 263,443 primary observations across 3,170 distinct observation dates, the earliest reaching back to 1990. About 24,000 new observations arrive each night. Alongside them sit 10,532 first-tier signal records from commentary, 909 mechanism-bound clusters and 321 tracked situations. The standing model is 25 active theses carrying 121 predicted observables and 102 falsifiers between them.

A night of that analysis costs about eight dollars and thirty-five cents in compute, and the whole of August came to roughly 227 dollars. That number matters for a reason beyond price. Holding twenty-five hypotheses against 24,000 daily observations across fourteen domains, every night, without letting today's headline reweight yesterday's belief, is not a workload a person or a team sustains. The discipline did not become easier. The arithmetic became cheap, which is why the machine can now be made to run it.

Our failures have a number

Every intelligence supplier will tell you their analysis is good. Here is the part nobody publishes.

Of the observables carried by retired theses, 84 reached their stated deadline and never arrived. Theses in this system die of non-arrival, and the record keeps the body. The revision log is append-only, so the

predictions can be read against us rather than for us.

Our trading layer takes notional positions against these theses with transaction costs counted, marked against the market on a fixed horizon. The early numbers are strong: 65 scored marks across brief dates in mid-August show a mean alpha of 3.5 percentage points with 64 of the 65 positive, against a benchmark that fell. We are not quoting those at you as evidence, because they are not. Sixty-five marks are ten brief dates re-scored across an overlapping book inside a single ten-day window and a single market regime, in notional money. A run of 64 wins from 65 in that shape measures correlation before it measures skill. We will publish the record when the sample means something, whichever way it falls.

Three current limitations, stated so you hear them from us. Roughly two in five live predictions do not yet carry a deadline, which means that portion of the model cannot currently fail by the passage of time; the rule that closes this is now built into our weekly review. The corpus you carry is current state plus short-window deltas rather than deep narrative history. And non-English primary sources are thin.

What this is not for

If your work sits one domain deep, a specialist terminal will serve you better and we will tell you so. Radar earns its place where decisions cross domains and cannot wait for consensus: energy against currency, freight against conflict, sanctions against supply, demographics against sovereign credit.

It is also not a replacement for judgement. The design premise is the opposite. Every one of the four mechanisms exists to return a specific judgement to a person rather than absorb it, and the one place we gave a machine the final word, we took it back.

How to test us

Ask Radar something you already know the answer to.

That is the only test worth running and it is the one we would run in your position. Bring a question inside your own domain, where you can tell immediately whether the answer is right, whether the reasoning holds, and whether the parts it claims to have measured were actually measured. Then follow one claim down to the record underneath it and check that the record says what the claim says it says.

We will show you the reading, the chain beneath it, and the places where the corpus does not reach. If it does not survive that, you should not buy it.

Bearing Partners Pty Ltd is a Canberra-based firm. Radar is the sense layer of the Radar-Compass Strategic OS. Compass decides against its corpus; Sextant positions against it. The full methodology, with complete citations and a fuller statement of limitations, is set out in the companion research paper of the same title.

Sources

1. Substrate figures read from the Radar production instance, 23 August 2026, and reproducible against it.
2. Gillespie, N., Lockey, S., Ward, T., Macdade, A., & Hassed, G. (2025). Trust, attitudes and use of artificial intelligence: A global study 2025. University of Melbourne and KPMG. 48,340 respondents across 47 countries. https://doi.org/10.26188/28822919
3. Blackburn, J. (2026). A Pandemic of the Mind? The Potential Impact of AI on National Security and Resilience. Institute for Integrated Economic Research – Australia, §2.3.3 and §8.1. https://www.jbcs.co/iieraustralia-projects
4. Citation hallucination rates of 11 to 57 per cent across commercially deployed models, as surveyed in ‘Cited but Not Verified’ (2026). https://arxiv.org/pdf/2605.06635

5. Dassen et al. (2026), finding up to 57 per cent of retrieval-augmented citations attached after answer generation. https://arxiv.org/pdf/2606.13104

6. Sharma, R. K. (2026). ContextCov. Constraint compliance of 88.3 per cent for compiled executable checks, against 67.0 per cent prompt-only and 50.3 per cent model self-reflection. https://arxiv.org/pdf/2603.00822

7. Grabo, C. M. (2002). Anticipating Surprise: Analysis for Strategic Warning. On absence of an expected indication as an early warning signal. https://apps.dtic.mil/sti/pdfs/ADA476752.pdf

8. Register reference files REG-01 to REG-15, read 23 August 2026: approximately 11,650 lines across fourteen documents, each authored against a twelve-principle checklist.

9. Anthropic Console usage export, 1 to 22 August 2026, twenty-two complete nights: median 1,875,678 tokens per night, median USD 8.35 at published rates, month total approximately USD 227.

10. Sextant proving harness and ledger, read 23 August 2026. Scoring is forward-only and history is never edited; positions that cannot be priced are counted and named rather than dropped.
