The Trace and the System

Closing the verification gap in open-source intelligence by compiling analytical norms into enforced mechanism

A BEARING PARTNERS RESEARCH PAPER · AUGUST 2026

ABSTRACT

Most consequential decisions are now taken about places the decider will never visit, on evidence the decider did not collect. Classified collection serves a vanishingly small population: cleared officials inside a few states, working a narrow band of questions. For almost everyone else, almost always, open-source intelligence is not a cheaper substitute for that material. It is the only base layer available. That makes its reliability a structural question rather than a quality question: whatever error is present in the base layer propagates upward silently, and the more capable the reasoning above it, the more efficiently it converts a corrupted input into a confident decision. Open-source intelligence has solved collection and has not solved verification. A reader outside the producing organisation usually cannot tell which parts of an assessment are measured, which are inferred, and what would make the producer withdraw it. We argue that this gap exists because the discipline states its analytical commitments as norms addressed to a human analyst, and norms must be complied with rather than enforced. We argue it is closable by compiling four commitments into mechanism: governing measurement in versioned instruments before interpretation, validating analytical definitions in code after the model answers, binding belief revision to dated evidence with the revision decision taken away from the model, and minting citations in the retrieval layer so the model cannot write its own receipt. We describe Radar, a system that has run nightly since May 2026, and report figures read from its production instance on 23 August 2026. We also report where the system falls short of its own standard. The contribution is a design pattern and a working case, not a performance claim.

1. Introduction

Most decisions that matter are now taken about places the decider will never visit, on evidence the decider did not collect. A manufacturer choosing where to hold inventory, an insurer pricing a route, a fund sizing an exposure, a government 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.

A very small number of people work from material that is not open. Classified collection exists in a handful of states, and even inside those states it is confined to cleared officials working a narrow band of questions deemed to be of national interest. Set that against the population actually making consequential decisions about distant systems: corporate boards, insurers, investors, regulators, humanitarian agencies, journalists, and the great majority of public servants in every country including those few. On that measure the privileged fraction is a rounding error.

Open-source intelligence is therefore not a cheaper substitute for the real thing. For almost everyone, almost always, it is the only thing. It is the base layer on which an enormous volume of consequential judgement rests, and its share of that load rises as the volume of published institutional data rises with it.

Base layers have a specific property that makes their reliability a structural question rather than a quality question. Error in a base layer does not stay in the base layer. It propagates upward, silently, into everything built on top, and it arrives there 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

corrupted input into a confident decision. Sound method applied to unsound ground does not fail loudly. It fails persuasively.

We make this argument as people who have taken the risk ourselves. Radar is not a standalone product in our own use of it. Its corpus is the input to Compass, which frames and adjudicates decisions, and to Sextant, which takes positions in markets against the resulting theses.[36] That arrangement forces a question on us that most producers of intelligence never have to answer. When something downstream goes wrong, we have to know whether the reasoning failed or the ground did. If the base layer cannot be interrogated, that question has no answer, the failure cannot be attributed, and nothing in the stack can learn from it.

This is the general case, not a peculiarity of our architecture. Any organisation that layers judgement onto an intelligence feed inherits that feed's errors without inheriting any means of detecting them. The remedy cannot be to trust harder. It has to be that the base layer arrives in a form that can be checked: not a claim to objectivity, which no system can honour, but a shared picture whose provenance is legible and whose mistakes are traceable to where they were made. A base layer nobody can check is not a foundation. It is an assumption with citations attached.

That is the standard against which this paper measures the field, and the standard it accepts for itself.

Open-source intelligence has solved its collection problem. It has not solved its verification problem. Commercial platforms now monitor at volumes that would have seemed impossible a decade ago, in the millions of sources, processed continuously and combined with the work of large analyst networks.[1] Coverage is no longer the constraint. Something narrower is. Given an assessment from such a system, a reader outside the firm that produced it usually has no way to establish which claims are measurements and which are inferences, how far the conclusion follows from the evidence, or what would have to happen for the producer to withdraw it.

This is not an accusation of carelessness. It is a description of how the field is built. The methods of the major platforms are proprietary and unpublished, so whatever discipline they apply internally cannot serve as a warrant for anyone outside. Trust rests on the reputation of the firm and the credentials of its analysts, not on any property of the document delivered. For many purposes that is enough. It stops being enough exactly where the stakes rise, which is where a decision-maker most needs to know whether an assessment will bear weight.

Historically the gap was closed at the receiving end. An assessment arrived in an institution that contained people who had spent careers in the subject and could tell a sound argument from a fluent one. That assumption is now weakening, and its weakening is what turns an untidy problem into an urgent one.

The first sign is behavioural and measured. In a study of more than 48,000 people across 47 countries, two thirds of employees who use artificial intelligence at work said they had relied on its output without evaluating it. More than half had received no training in its use.[2] Verification is not merely unavailable to the reader. It is often not attempted.

The second sign is institutional. Blackburn, drawing on workshop testimony from about seventy practitioners across Australian defence, industry, academia and government, describes what he calls unintended obfuscation. Machine assistance raises the surface quality of written analysis while hiding the reasoning underneath it, so work that reads well and is wrong in detail passes review by supervisors who

lack the depth to notice.[3] He places this inside a longer shift from specialist expertise toward generalist management, whose effect is that the quality control which once caught unsound analysis has itself been worn down.

Related findings point the same way. A 2026 synthesis by researchers across FAR.AI, MIT, Oxford, Cornell, Berkeley, Carnegie Mellon, Imperial College, McGill and Mila names epistemic risk as a class of harm, argues that machine assistance now intervenes in belief formation rather than merely in recall, and identifies a further risk that matters here: reduced capacity to reason degrades the very faculty needed to notice the reduction.[4] Experimental work finds that short conversational exposure shifts stated positions, with much of the shift still present a month later.[5] Other work finds that popular models are less diverse in the views they surface than ordinary web search, and that diversity falls as model size rises.[6]

Put together, these results change the shape of the problem. If we cannot assume the reader is able to audit an assessment, then the assessment has to carry its own audit. Verifiability has to be a property of the artefact rather than a service supplied by its recipient.

The academic literature offers rigorous partial answers, and what they cost is instructive. Political violence early-warning systems such as ViEWS publish probabilistic forecasts on a fixed cadence and evaluate them continuously against held-out data, which makes them auditable in a way no commercial product is.[7] Forecasting tournaments and the benchmarks descended from them score participants against resolved outcomes using proper scoring rules, producing calibration evidence the intelligence community has historically lacked.[8] Both buy verifiability by narrowing. The first fixes the outcome variable. The second fixes the questions. Neither helps an analyst who has to hold a broad and continuously updated picture of how several interacting systems are moving, without knowing in advance which question will matter.

That gap, between verifiable narrowness and unverifiable breadth, is what this paper addresses. Large language models make continuous broad assessment practical for the first time. A machine can hold dozens of standing hypotheses against tens of thousands of daily observations across unrelated domains, and reason across the whole surface on demand. But the same technology sharpens the verification problem it appears to relieve, and the failures are now measured rather than suspected. Fabricated or unsupported citations persist across deployed systems at rates between 11 and 57 per cent.[9] Retrieval grounding does not fix this, because citations are often attached after the answer is written rather than derived from the evidence.[10] Compliance with analytical constraints falls sharply when those constraints are written as prompt text rather than executable checks.[11] Belief revision under new evidence has been found conservative and inconsistent in ways a reader cannot see from the output.[12] An assessment that is fluent, well organised and wrong is now cheap to produce, and it is arriving into institutions with less capacity to catch it.

What the discipline should do about this was written long before the technology existed, and we claim no originality in it. Wohlstetter, analysing the Pearl Harbor failure, concluded that signal can be separated from noise only by an observer already holding explicit hypotheses that direct attention.[13] Grabo distinguished warning intelligence from current intelligence, and said its earliest indications usually arrive as fragmentary evidence, as conflicting reports, or as the absence of something expected.[14] Heuer formalised the maintenance of competing hypotheses as a defence against settling too early.[15] Popper supplied the test, that a claim earns standing by what would refute it.[16] Braudel supplied the ontology,

separating the slow structural movements that determine outcomes from the layer of events that reports them.[17]

These prescriptions are consistent, well evidenced and widely taught. They are also, without exception, addressed to a human analyst as norms requiring voluntary compliance, and the literature is candid that they are applied sparingly in practice because they are laborious.[18] Moved into a machine pipeline unchanged, they become instructions in a prompt, which the evidence above identifies as the weakest form of enforcement available.

This paper argues that the verification gap in open-source intelligence follows from stating analytical commitments as norms rather than as mechanisms, and that it is closable. The four commitments are these. Measurement is governed by versioned instruments before any interpretation happens. Analytical definitions are checked in code after the model answers, rather than requested before it. Belief revision is tied to dated evidence, with the revision decision removed from the model. And citations are minted by the retrieval layer rather than composed by the model. We argue that when these four are compiled into enforced parts of a working system, the reliability of the resulting assessment becomes a property of the artefact rather than a matter of trust in its producer.

We support the argument by describing Radar, a system that implements all four and has run nightly since May 2026, and with figures read directly from its production instance on 23 August 2026. Section 2 sets out the four mechanisms and their published ancestors. Section 3 reports what the system holds and what it costs. Section 4 gives the accountability record, including where it currently falls short. Section 5 takes the strongest objections, including one we cannot fully answer. Section 6 states what we claim is new and what is not.

2. Four Commitments, Compiled

Each of the four mechanisms below follows the same shape. A norm exists in the literature and is widely accepted. Expressed as instruction, it is unreliable. Expressed as mechanism, it holds, and it fails visibly when it fails.

2.1 Measurement governed before interpretation

Radar reads fourteen registers directly from the institutions that publish the world's machinery. Reserves and central bank balance sheets from the IMF. Cross-border capital flows from the US Treasury and the Bank for International Settlements. Energy from the Energy Information Administration. Corporate behaviour from the Securities and Exchange Commission. Sanctions designations from the Office of Foreign Assets Control and the UN Security Council. Humanitarian appeals from the UN Office for the Coordination of Humanitarian Affairs. Organised violence from the Uppsala Conflict Data Program. Actively exploited software vulnerabilities from the Cybersecurity and Infrastructure Security Agency. The orbital catalogue from Space-Track.

Calling those endpoints is easy and is not where the value sits. Each register carries a reference file, treated as a versioned constitutional document. The fourteen run to roughly 11,650 lines.[19] What those lines hold is the judgement that turns a number into a reading.

They hold direction rules, which fix what a movement means before any arithmetic runs. In the currency register some pairs are quoted as dollars per unit and others as units per dollar, so a sign error would silently invert the reading. The inversions are listed by ticker. In the conflict register the rule is sharper.

Violence events carry escalation polarity, and termination or peace-agreement events carry de-escalation polarity, and the file forbids adding the two together. A region with rising violence and a signed agreement is a contested truce, not a net of zero.

They hold publisher reliability, which grades the institution rather than the number. Reserve publishers are tiered by their subscription to the IMF data standards, from the highest tier down to countries under censure. A late release from a strong publisher is treated more seriously than a late release from a weak one.

They hold cohorts classified by measured behaviour rather than by intuition. Asian energy importers move together on reserves at a correlation of about +0.39, so three or more members moving the same way inside a thirty-day window is itself scored as a signal. Reserve-currency issuers move together at about +0.19 and are used as a control, to strip pure valuation effects out of everyone else's reading. Oil exporters, which any analyst would group by instinct, measure about +0.01, and the file marks them not to be scored. The reference files record where the author's intuition was tested and found wrong.

They hold thresholds taken from named frameworks rather than invented bands, including import cover, the Greenspan-Guidotti ratio, the IMF reserve adequacy metric and the Bergsten-Gagnon test for currency accumulation.[20] And they hold cross-register anchors with measured coupling strengths, so reserves are never read without currency, and currency is never read without the commodity that pressures it.

All fourteen were authored against a twelve-principle checklist derived from the system's constitution, with a written justification recorded for each principle and filed alongside the file. The reasoning was recorded at the time, so the check is auditable afterwards.

Absence is handled here too, and this is where Radar departs from the surrounding literature rather than inheriting from it. Grabo's third category of early indication is the absence of something expected. Each series declares its own expected cadence with a warning window and a stress window. A monthly release that has not arrived by day 45 is not a hole in the data. It is a candidate signal with a stated reason. One refinement keeps the rule usable. A series is flagged silent only while its neighbours are still reporting. If most of a register has gone quiet, that is an outage and the flag is suppressed. The system distinguishes a sensor that stopped from a feed that fell over.

The statistical treatment of missing data is overwhelmingly a literature of repair, through imputation, weight adjustment, and modelling of the missingness so the gap does not bias the estimate.[21] That work is careful and correct for its purpose, which is to recover the number that should have been there. But when a statistics office stops publishing a series, or a central bank's reserve release is late, the interesting object is not the missing value. It is the silence, and who chose it.[22] Radar treats the silence as the observation.

2.2 Analytical definitions validated in code

A system that asks a language model to follow analytical rules will receive a fluent report that the rules were followed. The evidence on this is now direct. When one recent framework compiled documented project constraints into executable checks instead of leaving them as instruction text, compliance reached 88.3 per cent. Prompt-only instruction reached 67.0 per cent. Asking the model to reflect on its own compliance reached 50.3 per cent, which is worse than either.[11] Self-review scored worst of the three. The model is not lying. It is agreeing.

Radar therefore states each analytical contract in code that runs after the model answers, and lets the code win.

The clearest case is the definition of a finding. A Radar cluster is a claim that one causal mechanism links signals from two or more different systems. That definition appears in the prompt, and it also exists as a function. After the model returns its groupings, code recounts the systems and members in each proposed cluster, removes duplicate members by content hash, enforces that a signal belongs to at most one cluster, and dissolves any cluster that fails the bar, returning its members to the unclustered pool. The definition proved unreliable when carried by instruction alone across runs, which is recorded in the source as the reason the guard exists.[23]

Synthesis is governed more strictly still. The daily brief's structure is assembled in ordinary code before any prose is written, and the model is then handed one item at a time. It cannot re-rank the day's findings, merge them, drop one, or introduce a fact, because it never sees the set. This matters for a reason Blackburn's workshops identify independently. A participant with a user-experience background described anchoring bias as the specific vulnerability at work: the first machine response frames the user's understanding of the problem and forecloses exploration downstream.[24] Structural assembly removes the model's opportunity to set that frame. A separate gate then scans the finished prose for raw register codes, for acronyms used without a gloss on first appearance, and for bare percentages given without a plain reading in the same sentence, and reports each violation with the offending sentence.

Beneath all of it sits a short file of non-negotiable facts appended to every customer-facing prompt. It currently holds one rule, about the geography of the Strait of Hormuz, which is the only way in and out of the Persian Gulf rather than a waypoint on a route through. Ships do not reroute around a Hormuz closure. The Cape of Good Hope substitutes for the Red Sea and Suez, and for nothing else. The rule ends with the instruction that carries the philosophy: if a source says otherwise, the source is wrong, and the claim is omitted rather than repeated.

The guardrail literature has converged on this architecture, that enforcement belongs in code running outside the model's context where instructions cannot be talked around.[25] It has converged on it for safety and security, to block unsafe tool calls and filter harmful content. Radar applies the same architecture to epistemics. What is being enforced is not whether the output is safe. It is whether the analysis obeyed its own definition of a finding.

2.3 Belief revision bound to dated evidence

The unit of analysis in Radar is the thesis, not the signal. A thesis is a standing structural claim. It carries a mechanism, an evidence base of record identifiers, a set of predicted observables each with a date by which it should have arrived, and a set of falsifiers describing what would refute it.

Each night, every thesis is assessed against three questions. Does today's evidence confirm it. Does it contradict it. Does it introduce a dimension the thesis does not address. The governing prompt states that registers and commentary are peers rather than a hierarchy. On a quantitative fact the register's number beats a headline discussing that number. On a causal question the commentary carries the inference. Where a register has moved and commentary has not caught up, the instruction is to record the move as trajectory and specifically not to mark the observable realised, because the edge has been seen and not yet confirmed. Where the two flatly disagree, the instruction is to hold, change nothing, and record the divergence as a finding.[26]

Then comes the decision that carries the weight. The model does not get to move the thesis. It judges content only. Status is set in code, on evidence, by three rules. A falsifier that has fired sets the thesis to contested, and that is a hard event which cannot be softened. A predicted observable whose window has passed without the thing arriving triggers a downgrade. Otherwise the thesis holds. The comment in the source states the principle plainly: silence is not decay. Confidence moves on evidence, and only on evidence.

This answers a documented weakness directly. Work on how models revise forecasts when new evidence arrives finds their updates conservative and inconsistent.[12] Diagnostic taxonomies of agent reasoning failures name a separate category for failure to revise or retract a claim after later evidence contradicts it.[27] Radar does not try to improve the model's updating. It removes the update from the model.

Where that decision sits is governed by a rule the system states as doctrine rather than as scheduling. Radar runs two lanes with different permissions, and the line between them is drawn by whether an error can announce itself.

The proposing lane is allowed to act autonomously. Machine-generated candidate theses may carry weight in Sextant, where every position is marked to market against a benchmark on a fixed horizon with transaction costs counted. Autonomy is permitted there because the lane pays for its own mistakes: an unsound machine-proposed thesis loses money, visibly, on a schedule nobody involved controls. That is a falsification mechanism rather than a supervision arrangement, and it does not require a person to work.

The canonical record is not allowed to move that way. No thesis enters the standing store, and no thesis changes status, without a human ruling at a weekly review. Per-thesis model judgement was removed from the nightly chain on 3 July 2026 and has not returned. The system continues to detect the mechanical triggers continuously so that nothing is missed between reviews, but detection has no write path to status.[28] What still runs nightly is the mechanical remainder.

The asymmetry is the whole argument, and it is the argument of this paper turned on its author. An error in a scored position is bounded, priced, and self-announcing. An error in the canonical record is none of those. It propagates into everything composed from the substrate, arrives downstream carrying the authority of established fact, and is converted into a confident decision the more capable the reasoning above it happens to be. Autonomy is therefore granted where consequences are measured and withheld where they are silent. The accepted cost is latency: the board may run up to six days behind the evidence, which is the price of the guarantee that nothing in the base layer moved without a person deciding it should.

This also gives a precise form to a recommendation the literature states as a practice. Blackburn argues that machine assessments should require independent human analysis before informing decisions, with documented reasoning explaining agreement or disagreement with the machine's conclusions.[29] The weekly review is that requirement implemented as a gate rather than recommended as a discipline, and it is cognitive offloading run in reverse. The machine holds the board, runs the measurements and detects the triggers. A person rules on what the standing model believes.

2.4 Citations minted by the retrieval layer

A Radar analysis is not delivered as a fixed report. It ships with an interrogation surface, an embedded analyst through which a reader can walk any conclusion backward to the register readings and raw records underneath it, or forward to what the thesis predicts next.

Every load-bearing claim carries one of four tags. RR means the claim is in the corpus. CR means a real external source was fetched during that exchange, and the citation carries the address the fetch returned. RI means an inference whose evidence is in the corpus but whose conclusion is not itself stated there. CI means reasoning with no corpus anchor. Two tiers are provable and two are honest judgement, and the reader can see at sentence level which is which.[30]

The rule that carries the weight is the guard against smuggling. ‘The corpus says X’ is RR. ‘X, and reading the data, that means Y’ is RI and never RR, even when Y names a corpus field. Tagging a corpus-anchored inference as RR is the credibility break. Tagging it as free reasoning undersells the work. The middle tier exists to make the honest call, and where a tier is uncertain the instruction is to under-claim.

None of this would matter if the model wrote its own receipts. It does not. Every retrieval tool builds its reference in code from the substrate row it returned and hands that back with the content. A register reading is referenced by its identifier, series and date. A thesis by its stored identifier. A cluster by its own. A raw signal by content hash, never by record identifier, which is not unique in the corpus. The runtime then validates each reference against the loaded substrate, and an unresolved one renders struck through. An invented receipt does not pass quietly. It fails in front of the reader.

Fetched sources are governed by a rule the fetch tool enforces. A citation may carry a link only if that exact address was returned by a fetch performed in that exchange. A source the model is certain exists but did not retrieve is demoted to unanchored reasoning, described by name, with no link.

The measurements explain why the rule is drawn so hard. Fabricated or unsupported citation rates between 11 and 57 per cent persist across deployed models.[9] An audit of generative search engines found only 51.5 per cent of sentences fully supported by their own citations.[31] In one early study of literature reviews, 55 per cent of one model's references and 18 per cent of its successor's were entirely fabricated.[32] And up to 57 per cent of citations in retrieval-based systems have been found to be attached after the answer was generated.[10] Retrieval alone does not fix this, because a system can still produce a link from memory rather than from an actual retrieval.

There is a further reason this tier structure matters, and it comes from the institutional side of the problem. Blackburn's unintended obfuscation describes work that reads as authoritative being assessed by reviewers who lack the specialist depth to evaluate it.[3] A tiered citation does not require domain expertise to read. A reader who knows nothing about sovereign reserves can still see that a sentence is tagged as inference rather than measurement, and can direct their scepticism accordingly. The mechanism moves part of the quality-control burden off the institution and onto the artefact, which is where it has to go if the institution can no longer reliably carry it.

3. What the System Holds and What It Costs

All figures in this section were read from the production instance on 23 August 2026 and are reproducible against it.[33]

The substrate holds 263,443 primary observations across 3,170 distinct observation dates, with the earliest reaching back to 1990. Roughly 24,000 new observations arrive each night across the fourteen registers. Alongside these sit 10,532 first-tier signal records drawn from commentary, 909 mechanism-bound clusters and 321 tracked situations, accumulated since late May 2026.

The standing model comprises 25 active theses: 13 at high confidence, 11 at medium, one weakening. Between them they carry 121 predicted observables and 102 falsifiers. A further 37 theses have been retired. The revision log holds 214 entries and the decision log 197.

The cost is worth stating because it is the reason the norms are now runnable at all. Across the twenty-two complete nights from 1 to 22 August 2026, the pipeline consumed a median of 1.88 million tokens per night, about 1.66 million read and 230,000 written. At published rates that is roughly USD 8.35 per night. The mean is higher, near 10.34 dollars, because the same credential also carries manual reruns and engineering work, so the median is the representative figure for a night of analysis. The month came to approximately 227 dollars.[34]

That figure should be read against the arithmetic of the norms themselves. Running the canon properly means holding each hypothesis, stating what would confirm and refute it, and checking both against every relevant observation daily, without letting the vividness of today's headline reweight yesterday's belief. Grabo's Watch Committee did this with a room of people against a bounded adversary. Twenty-five theses against 24,000 nightly observations across fourteen domains is not a workload a person holds. The norms did not become easier. The arithmetic became cheap.

This also bears on a limitation Blackburn states about his own method. His verification practice was to run several systems in parallel and cross-check by hand, and he reports plainly that this is neither scalable nor affordable in routine operation, and that future systems will need to automate recursive and parallel analysis.[35] We agree, and we would put it more sharply. A verification practice that depends on the analyst's diligence is a norm, and the literature on structured analytic techniques is unambiguous that norms of this kind are abandoned under load.[18] The same paper documents cognitive exhaustion and anchoring bias acting on exactly the analyst who is supposed to supply that diligence.[24] A safeguard whose load-bearing component is a tired and anchored reader is not a safeguard. That is the case for moving the burden into mechanism, and it is the case this paper makes.

4. The Accountability Record

Architecture is not evidence. The claim that matters is whether the standing model is any good, and Radar answers that in two places: a ledger of theses scored against arrival, and a book that puts money behind them.

4.1 Theses scored against arrival

Of the 121 predicted observables carried by active theses, 34 are flagged realised. Among the observables belonging to the 37 retired theses, 84 reached their stated window without arriving. Theses in this system die of non-arrival, and the substrate records it.

This is not a performance claim. It is a claim about accountability, and it is the property no comparable commercial product publishes. A conventional platform issues assessments and revises them. The revisions are visible; the original commitments are not scored against a date. Radar's revision log is append-only and its observables carry windows, so the record can be read against the producer rather than for them.

4.2 Positions marked to market

The second mechanism is a book. Sextant, the trading layer, reads the thesis board and takes positions wherever an instrument exists to express a thesis, with transaction costs counted. Its proving harness scores forward only and never edits history. Each shipped brief is marked at a fixed horizon, priced from the register series first and a market feed second, benchmarked against the total return of a broad equity tracker over the identical window, with any position that cannot be priced counted and named rather than dropped.[36] A market imposes the deadline a written observable can be allowed to skip, and it prices being early exactly as it prices being wrong.

The current record, reported with its weaknesses attached: 65 scored marks covering brief dates from 11 to 20 August 2026 show a mean alpha of 3.5 percentage points, 64 of the 65 positive, with the book returning 2.9 per cent while the benchmark fell 0.6 per cent over the same windows. Nightly turnover has run between 21 and 46 per cent at a transaction cost under three basis points, with nothing unmarked.

Those figures are not evidence and we do not present them as any. Sixty-five marks are not sixty-five bets. They are ten brief dates, each re-scored on roughly six subsequent days against an overlapping book of about ten positions, inside a single ten-day cohort and a single market regime, in notional money. A run of 64 wins from 65 in that shape measures serial correlation before it measures skill. An earlier form of the system carries a longer record of 234 marks from late June, and the harness itself tags those rows as a tooling exercise rather than evidence, which is the correct call and is respected here.

The commitment is therefore narrow and testable. The scoring machinery exists, runs nightly, is forward-only, cannot rewrite its own history, and is benchmarked. What it does not yet have is enough independent observations across enough regimes to mean anything. We will publish the record when that sample exists, whichever way it falls.

4.3 Where the system falls short of its own standard

Three deficiencies belong in the open, because a reader auditing the substrate will find them.

Forty-nine of the 121 active observables carry no window end. That is roughly two in five of the live model's predictions with no clock, and because the downgrade rule is calendar-driven, an observable without a date is exempt from it. This is a real gap in the accountability the paper claims. It is being closed.

A fifteenth register covering sanctions designations and humanitarian appeals is live in production and absent from both the system's constitution and its internal map. And the constitution's statement that register history is retained for 24 months is superseded by a substrate reaching back to 1990. Both are documentation defects rather than functional ones, and both are stated here rather than corrected quietly.

5. Objections

We take the four strongest objections in turn, and one we cannot fully answer.

The first is that Radar reads public data and asks a language model about it. The endpoints are free and anyone can call them. This is true and it is why the reference files exist. A free endpoint returns a number. What it does not supply is the cohort that number belongs to, whether that cohort genuinely moves together, the threshold framework it should be read against, the direction convention that stops a sign error inverting the meaning, the reliability tier of the publisher, the neighbouring register that must

be loaded alongside it, and the cadence whose breach is itself a reading. That is roughly 11,650 lines of authored judgement, checklisted against twelve principles.

The second is that the theses are text. So is every hypothesis. What makes these more than text is that they carry dated observables and named falsifiers, that a passed window forces a downgrade in code, that a fired falsifier forces contestation in code, and that the whole trail is append-only. A belief that cannot lose is not a thesis. These can lose, and 84 observables on retired theses have.

The third is that the citations are labels the model writes. This objection has the most force, which is why it is answered structurally rather than rhetorically. The model does not compose its references. The retrieval tools mint them from substrate rows and hand them over, the runtime validates them against the loaded corpus, failures render struck through, and a fetched-source citation may only carry an address an actual fetch returned in that exchange.

The fourth is the one this paper takes most seriously. A single production system built and described by its own author is not evidence that the design pattern generalises. We accept this. The four mechanisms are presented as a design pattern with a working existence proof, not as a validated intervention. Of the four, the citation mechanism is cleanly testable by a third party against published attribution benchmarks, holding corpus, questions and model constant while varying only who writes the reference. We regard that experiment as the necessary next step and do not claim its result in advance.

There is also a version of the objection we cannot fully answer. The comparison here is drawn against the open literature and publicly documented commercial products. The large geopolitical intelligence platforms combine machine monitoring at very large source volumes with substantial analyst networks, and their internal methods are not published. A rule resembling one of Radar's four could exist unpublished inside such a firm, or inside a government planning cell. We state the claim as the evidence supports it. After a deliberate search across the warning-intelligence, structured-analysis, forecasting, conflict-early-warning, missing-data, machine-reasoning, guardrail and citation-attribution literatures, we found no published method that combines all four. We found none that governs measurement in versioned per-domain instruments before interpretation. None that enforces analytical contracts in code after the model answers. None that moves standing beliefs only on dated evidence, with silence explicitly barred from counting as decay. And none that mints its citations from the substrate so the model cannot forge a receipt. The closest published relative we located, an agentic falsification framework, validates a supplied hypothesis episodically under sequential error control rather than maintaining a standing model against a live world.[37]

What the objections get right is the direction of the risk. A system like this fails by becoming fluent about nothing. Our defence is that each of the four mechanisms is a place where the machine can visibly fail: a cluster dissolved by the guard, an observable that passed its window unrealised, a divergence held open, a citation struck through. A product whose failures are invisible is the one to distrust.

6. Conclusions

Radar is an heir and says so. Its ontology is Braudel's. Its insistence that hypotheses precede hearing is Wohlstetter's. Its treatment of warning as distinct from current reporting, and of absence as an indication, is Grabo's. Its hypothesis discipline is Heuer's and its falsification standard is Popper's. Its register substrate is the ordinary practice of every central bank watcher who ever read a reserves release. None of that is claimed as invention.

What we claim is new is the compilation. Four norms that the canon stated and practitioners skipped are here compiled into mechanism, each with a schema field, a code path, and a failure mode a reader can see. Measurement is governed before it is interpreted, in fourteen versioned instruments carrying direction, reliability, cohort behaviour measured rather than assumed, threshold frameworks, cross-register coupling, and a cadence whose breach is a signal. Analytical contracts are enforced in code after the model answers, because the literature now measures what happens when they are merely requested. Standing belief moves only on dated evidence, with the update decision taken away from the model and, since July, ruled on weekly by a person. And every claim carries a receipt the model did not write, minted by the retrieval tool, validated at runtime, and demoted to honest reasoning the moment it cannot be proved.

We also claim, more narrowly, that this is what augmented intelligence looks like when it is specified rather than advocated. Blackburn's distinction between machine assistance that replaces judgement and machine assistance that strengthens it is the correct frame, and his account of forcing functions, of visible process rather than bare output, and of the human-in-the-loop as a governable design choice rather than a vendor default, describes the design brief this system was built against.[38] Where we differ is on where the burden should sit. Blackburn's remedies largely address the human: training, cognitive maintenance, independent verification, deliberate practice without machine support. Those are necessary and we do not dispute them. But they are norms, and his own evidence, on exhaustion, on anchoring, on the two thirds who do not evaluate what they are given, describes with some precision why norms will not hold. If the reader's capacity to audit cannot be assumed, the artefact has to carry the audit.

The limits of what we have shown should be stated as plainly as the claim. One system, built by its author, is an existence proof and not a validated intervention. The trading record is too short and too correlated to carry weight. Two in five live predictions currently have no deadline. The comparison against closed commercial methodologies cannot be completed from outside them. What we offer is a design pattern, a working case, a substrate that can be read against us, and a specification of the experiment that would test the one mechanism a third party can test today.

Seventy years of warning doctrine told analysts what good judgement looks like, and then trusted a person to keep it up every day, indefinitely. The contribution here is not a better norm. It is the observation that norms are the wrong container, and a demonstration that the container can be changed.

Bearing Partners Pty Ltd is a Canberra-based firm. Radar is the sense layer of its Radar-Compass Strategic OS: a nightly, register-grounded structural intelligence pipeline whose corpus is the cardinal input to Compass, the decide layer, and to Sextant, the trading layer. Correspondence: Bearing Partners, Canberra.

Notes and References

1. On published descriptions of the leading commercial platforms: Seerist reports collection across 6.8 million open sources combined with human-verified events drawn from a network of Control Risks analysts; Recorded Future reports continuous processing across a large entity graph. Neither publishes a constraint-enforcement, belief-revision or citation-tier methodology of the kind described in this paper. https://www.carahsoft.com/seerist

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. Survey of 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. Sections 2.3.3 (AI as unintended obfuscation) and 8.1 (the generalist-specialist collapse). Released under CC BY 4.0. https://www.jbcs.co/iieraustralia-projects

4. Yang et al. (2026). AI Epistemic Risks: Emerging Mechanisms and Evidence. On epistemic risk as a class, on cognitive offloading intervening at the level of belief formation rather than recall, and on the meta-risk that reduced epistemic capacity degrades the capacity to recognise the

problem. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6873005

5. Hackenburg, K., Tappin, B. M., Hewitt, L., et al. (2025). ‘The levers of political persuasion with conversational artificial intelligence.’ Science 390(6777). https://doi.org/10.1126/science.aea3884

6. Wright et al. (2026). ‘Epistemic Diversity and Knowledge Collapse in Large Language Models.’ Finding popular models less epistemically diverse than basic web search, with model size negatively correlated with diversity. https://arxiv.org/abs/2510.04226

7. Hegre, H. et al. (2019). ‘ViEWS: a political violence early-warning system.’ Journal of Peace Research 56(2), 155–174; and Hegre et al. (2021), ViEWS2020. https://journals.sagepub.com/doi/full/10.1177/0022343319823860

8. Tetlock, P. E. & Gardner, D. (2015). Superforecasting: The Art and Science of Prediction. Crown; and Karger, E. et al. (2024), ForecastBench, reporting superforecaster Brier 0.096 against a leading model 0.122 at introduction. https://arxiv.org/pdf/2409.19839

9. Yuan et al. (2026), reporting citation hallucination rates of 11 to 57 per cent across commercially deployed models, as surveyed in ‘Cited but Not Verified: Parsing and Evaluating Source Attribution in LLM Deep Research Agents’ (2026). https://arxiv.org/pdf/2605.06635

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

11. Sharma, R. K. (2026). ‘ContextCov: Deriving and Enforcing Executable Constraints from Agent Instruction Files.’ Evaluated across 12 repositories and 300 tasks: 88.3 per cent constraint compliance for compiled executable checks, 67.0 per cent for prompt-only instruction, 50.3 per cent for model self-reflection. https://arxiv.org/pdf/2603.00822

12. Yuan, S. et al. (2025). ‘EvolveCast’, on model forecast revision under new evidence being overly conservative and inconsistent. https://arxiv.org/html/2603.16642v1

13. Wohlstetter, R. (1962). Pearl Harbor: Warning and Decision. Stanford University Press. The hypotheses-guide-observation formulation is discussed in National Research Council (2011), Intelligence Analysis: Behavioral and Social Scientific Foundations. https://www.nationalacademies.org/read/13062/chapter/8

14. Grabo, C. M. (2002). Anticipating Surprise: Analysis for Strategic Warning. Joint Military Intelligence College (condensed from the classified Handbook of Warning Intelligence, 1972–1974). https://apps.dtic.mil/sti/pdfs/ADA476752.pdf

15. Heuer, R. J. Jr. (1999). Psychology of Intelligence Analysis. CIA Center for the Study of Intelligence. https://www.cia.gov/static/9a5f1162fd0932c29bfed1c030edf4ae/Pyschology-of-Intelligence-Analysis.pdf

16. Popper, K. (1959). The Logic of Scientific Discovery. Hutchinson.

17. Braudel, F. (1958). ‘Histoire et sciences sociales: la longue durée.’ Annales 13(4), 725–753. Braudel assigns l'histoire événementielle to the chronicler and the journalist, and compares events to froth on the surface of the sea. https://en.wikipedia.org/wiki/Longue_dur%C3%A9e

18. National Research Council (2011). Intelligence Analysis for Tomorrow: Advances from the Behavioral and Social Sciences, ch. 3, on structured analytic techniques and their sparing application in practice. https://nap.nationalacademies.org/read/13040/chapter/5

19. Register reference files REG-01 to REG-15, Radar production tree, read 23 August 2026: 694 KB across fourteen YAML documents, approximately 11,650 lines, each accompanied by a filed twelve-principle checklist pass. Internal.

20. Publisher tiering follows the IMF data standards initiatives (SDDS Plus, SDDS, e-GDDS). Threshold frameworks referenced in the reserves register include import cover, the Greenspan-Guidotti rule, the IMF ARA EM metric and the Bergsten-Gagnon five-criteria currency-accumulation test.

21. On the statistical treatment of missingness as repair rather than signal, see Rubin's missing-data framework and its descendants, and for a recent worked example the BLS treatment of the 2025 lapse in appropriations via survey-weight adjustment. https://www.bls.gov/opub/mlr/2026/article/addressing-missing-consumer-expenditure-data-due-to-the-2025-lapse-in-appropriations.htm

22. On deliberate cessation of official publication as an object of analytic interest: Wall Street Journal analysis of hundreds of Chinese data series discontinued without explanation. https://chinadigitaltimes.net/2025/05/censored-statistics-deleted-data-muddy-the-waters/

23. Radar production source, read 23 August 2026: cluster/cluster.py enforce_cross_system(); thesis/maintain.py govern_status(); synthesis/overview.py structural assembly and output gate; houserules.py. The cross-system guard's docstring records that prompt instruction alone proved unreliable across runs. Internal.

24. Blackburn (2026), §2.3.4, on anchoring bias placing a cognitive box around what the user can consider, and §4.1.1 on the cognitive exhaustion paradox. https://www.jbcs.co/iieraustralia-projects

25. On structural guardrails as code executing outside the model's context window and therefore not bypassable by instruction, see the guardrail architecture literature. https://www.openlegion.ai/en/learn/ai-agent-guardrails

26. Radar production source: prompts/THESIS_MAINTENANCE_PROMPT_v1_1.md, on the three-question assessment, the register-commentary peer rule, and the instruction to hold and flag divergence. Internal.

27. On failure to revise a claim after contradicting evidence as a catalogued reasoning failure mode: ‘Stalled, Biased, and Confused: Uncovering Reasoning Failures in LLMs for Cloud-Based Root Cause Analysis’ (2026). https://arxiv.org/pdf/2601.22208

28. Radar production source: thesis/weekly_review.py, whose module docstring records the 3 July 2026 decision to retire nightly automated status governance in favour of a weekly human-gated review, and thesis/nightly_seed_housekeeping.py, which retains the mechanical remainder. Internal.

29. Blackburn (2026), §2.3.7 and Annex B, verification requirements for national security applications. https://www.jbcs.co/iieraustralia-projects

30. R0B_RADAR_PROMPT v2.1 and the retrieval tool layer, read 23 August 2026. The four-tier scheme, the RR-smuggling guard, the no-unfetched-URL rule, and the requirement that every retrieval tool returns the raw substrate identifier as its citable reference. Internal.

31. Liu, N. F. et al. (2023), audit of generative search engines finding 51.5 per cent of sentences fully supported by their citations; and Gao, T. et al. (2023), ALCE. https://arxiv.org/html/2604.03173v1

32. Walters, W. H. & Wilder, E. I. (2023), finding 55 per cent of GPT-3.5 and 18 per cent of GPT-4 references entirely fabricated across 42 topics. https://arxiv.org/pdf/2606.13104

33. Substrate counts read from the production Postgres instance, 23 August 2026: register_history 263,443 rows across 3,170 distinct observation dates from 1990-01-01; tier1_records 10,532; clusters 909; situations 321; theses 62 (25 active, 37 retired); thesis_revisions 214; thesis_decisions 197. Seven-day ingestion approximately 169,000 observations across 14 registers.

34. Anthropic Console token usage export for the Radar credential, 1 to 23 August 2026, cross-checked across three exports which reconcile exactly on all overlapping days. Twenty-two complete nights: median 1,875,678 tokens per night (1,658,718 read, 229,810 written); mean 2,384,840. Costed at published rates of USD 3 per million input and USD 15 per million output tokens: median USD 8.35 per night, mean 10.34, maximum 20.98, month total approximately 227. The credential also carries manual reruns and engineering work, so the mean overstates the scheduled nightly chain.

35. Blackburn (2026), Annex B, on multi-system parallel verification and its scalability limits. https://www.jbcs.co/iieraustralia-projects

36. Sextant proving harness and ledger tables, read from the production instance 23 August 2026. Current-form sample: 65 scored rows across brief dates 2026-08-11 to 2026-08-20; mean alpha 3.520 percentage points, standard deviation 2.276, 64 rows positive; mean book return 2.949 per cent against benchmark -0.571 per cent; mean 10.6 positions per brief. Prior-form sample: 234 rows from 2026-06-24, mean alpha 1.654, standard deviation 4.472, 144 positive, which the harness itself tags as a tooling exercise rather than evidence. Rows are keyed (brief_date, scored_on) so a re-score upserts the current day only and history is never edited. Internal.

37. Huang, K. et al. (2025). ‘Automated Hypothesis Validation with Agentic Sequential Falsifications’ (POPPER), ICML 2025. https://arxiv.org/pdf/2502.09858

38. Blackburn (2026), §7.1 on augmented versus artificial intelligence, §5.7 on the agentic layer as a sovereign leverage point, and §5.7.1 on forcing function design. https://www.jbcs.co/iieraustralia-projects

---