The Judgment Layer: Making AI Judgment Explicit, Testable and Cumulative

Christopher Berzins
Visiting Professor, Graduate School of Public and International Affairs, University of Ottawa
cberzins@uottawa.ca

Draft 11, 25 September 2026

Abstract

In September 1962 the United States Intelligence Board judged it unlikely that the Soviet Union would place strategic missiles in Cuba. The estimate was carefully reasoned, and it was wrong. It was corrected by an official who set the probability aside and asked what the new air-defence sites in Cuba were there to protect. AI systems that carry out open-ended work now make judgments of this kind all the time, about what deserves attention, which explanations to pursue and when the evidence justifies action, and usually nobody has examined how they make them. This paper treats judgment as a capability of AI systems that can be specified, disciplined and tested, and proposes that it can be made cumulative, so that experience scored against outcomes improves the questions a system asks. Intelligence Community Directive 203 supplies the institutional precedent, and a judgment layer carries it into AI design. When the analyst is a system whose procedures can be switched on and off, the effect of a single tradecraft standard can be isolated, which has not been possible with human analysts. Every claim about the programme behind the paper is labelled as specified, implemented, operating or measured. Three measured results are reported. Six widely used models, left unprompted, omitted the sources, alternatives or gaps the standards require. The programme's own architecture produced citations that resolved to the document without supporting the claims attached to them. And retrieval, structured prompting and deliberation added nothing measurable to forecast accuracy. A decisive study is specified, the objection that better models will make the layer redundant is addressed, and ways to take part are set out.

Keywords: judgment; artificial intelligence; strategic cognition; insight generation; analytic standards; intelligence analysis; decision-making under uncertainty; AI safety; human–AI collaboration

1. Why AI needs judgment

On 19 September 1962 the United States Intelligence Board approved a special estimate of the Soviet build-up in Cuba (USIB, 1962). The reasoning was careful. Moscow had not, the estimate noted, placed strategic missiles even in its satellite states, and on 11 September it had declared that it had no need to station such weapons abroad. Kent later recorded that September's overflights had repeatedly shown ground-observer reports of missiles to be inaccurate or wrong (Kent, 1964). Offensive missiles were judged unlikely. One official had already asked a different question. John McCone, the Director of Central Intelligence, had argued in August and then cabled from France that the surface-to-air missile sites photographed on 29 August made sense only if they guarded something worth the risk. A Defense Intelligence Agency officer, Colonel John Wright, then noted

that three of the sites in Pinar del Río protected no known installation, that a trapezoid of ground nearby had been sealed off by Soviet personnel and emptied of Cubans, and that an agent's report placed secret work believed to involve missiles inside it (Wright, 1962; Holland, 2005).

On 9 October Kennedy approved sending the U-2 back over western Cuba, from which it had been kept for fear of those same air-defence sites, and on 14 October it photographed the medium-range launchers (Holland, 2005). Sherman Kent, who chaired the board that drafted the estimate, wrote afterwards that his estimators had missed the decision because they could not believe Khrushchev could make a mistake (Kent, 1964). His account appeared in his agency's classified journal and was reprinted publicly three decades later. Inside the agency the failure acquired a name, mirror-imaging, and was taught (Heuer, 1999). In 1984 another estimate rested on the same assumption about what the Soviet leadership would find reasonable, and a later review found it wrong (DCI, 1984; PFIAB, 1990). The lesson had been recorded and was known, and nobody applied it when the assumption returned.

I expect that a forecasting system given the September evidence would have returned the estimate's answer. A well-calibrated forecaster assembles the base rate, treats the declaration as evidence of intent and produces a low probability with a sensible interval, and nothing in calibration prompts it to ask what the air-defence sites protect, whether the decision in front of the President is the estimate or the overflight, or whether a rival explanation already on the table could be settled by a single observation. This is a hypothesis about current systems, untested directly, and Section 6.2 proposes the cheap test. The pattern is general. A board reading a supplier's announcement, a research group whose planned experiment cannot separate the explanations it was designed to test, and a ministry forecasting another government's next move face the same structure. A correct answer is worth little when the important question has been missed, and an accurate forecast can support a poor decision when its bearing on that decision has been misunderstood.

Judgment, as I use the term, is the formation, use and revision of assessments about what matters, what is credible and what warrants inquiry or action under uncertainty. It includes recognizing a worthwhile question, developing an explanation, choosing an investigation and deciding how much confidence the evidence permits. When an AI agent, a system that pursues a task through repeated use of information and tools, takes on such work, these choices are delegated with it, and their quality shapes what its other capabilities accomplish. Two distinctions do most of the work in what follows. Evidence is a relation between a signal and a claim: a reliable report of an official's statement is strong evidence of the declared position and weak evidence of the official's intention, and several reports that repeat one announcement add little independent support, a failure analysts call circular reporting. And an accepted assessment is separate from permission to act on it: a hypothesis may be uncertain while a modest test of it is authorized. In 1962 McCone's hypothesis was uncertain, the overflight it implied was a decision for the President, and the two were rightly kept apart.

The opportunity extends to learning how to make these choices better. Kent's diagnosis was a judgment about judgment, recorded by the person best placed to record it, and it was not brought to bear when it was needed. Experience could help a system recognize which questions have led to useful discoveries, which analogies have survived scrutiny and which investigations have resolved consequential uncertainty, and a lesson about a recurring causal mechanism might improve an inquiry in a different domain. No institution has managed that reliably, and it is the capability this paper is about.

Intelligence Community Directive 203, or ICD-203, provides the institutional precedent. Issued

in 2007 under the Intelligence Reform and Terrorism Prevention Act of 2004 and revised in 2015, after the Iraq post-mortems had found, among other failures, that several reports which appeared to corroborate one another traced to a single source (Senate Select Committee on Intelligence, 2004; Butler, 2004; WMD Commission, 2005), it sets standards for assessments made with incomplete and sometimes deceptive information and places evidence, uncertainty, alternatives and revision within a system of review and training (ODNI, 2015, as amended). Current research shows why AI needs the same treatment. Co-Scientist combines hypothesis generation, criticism and refinement (Gottweis et al., 2026), while two case studies of open-ended AI research describe agents that completed the engineering unaided yet showed poor judgment about what counts as publishable, uncreative responses to flaws in their own designs and ineffective backtracking from dead ends (Kirgis et al., 2026). The question also has an economic form: as prediction becomes cheap, value moves to its complements, among them the judgment that assigns value to outcomes and chooses the action (Agrawal, Gans and Goldfarb, 2018), and on the account given here judgment also covers deciding which question is worth asking.

A *judgment layer* is the set of functions through which an AI system develops, tests, uses and revises its understanding. The functions can be distributed across models, retrieval, memory, computational tools and human review. In scientific work they include finding the observation that separates rival explanations, and in strategic work recognizing a changing constraint, an overlooked dependency or the consequences of another actor's response. McCone's question and Wright's map are examples of what the layer exists to produce. Leadership Under Uncertainty is the research programme in which this work is pursued. It connects a philosophical account of judgment, a shared representation for judgments, an architecture for discovering strategic insights, evaluation instruments, an organizational theory, operating reference feeds and a set of implemented applications, and the connections give each part a role: an explicit alternative can direct a search for evidence, a candidate insight can reveal an assumption worth reconsidering, and evaluation can identify why a contribution helped. Each claim about the programme is labelled by the level at which it exists, specified, implemented, operating or measured, and Section 5 gives the map. The programme's own measurement of the failure this paper is about appears there: on the largest board document it examined, every citation the architecture produced resolved to the document and none supported its claim.

2. ICD-203: a practical precedent

ICD-203 offers a mature example of an institution making its expectations for judgment explicit. Its five overarching standards and nine tradecraft standards govern intelligence analysis, and product evaluation and analyst training support their application (ODNI, 2015, as amended). Table 1 restates the nine tradecraft standards as behaviour an AI system would have to show, which is how the precedent becomes a specification.

Table 1. Tradecraft standards as requirements on an AI system

ICD-203 tradecraft standard	What the system would have to show
1. Describe the quality and credibility of sources	Which claims rest on which sources, how reliable each is, and when several reports share one origin
2. Express and explain uncertainties	A likelihood for the assessed event and, separately, the strength of the evidence behind it

ICD-203 tradecraft standard	What the system would have to show
3. Distinguish information from assumptions and judgments	Which statements are reported, which are assumed and which are the system's own inferences
4. Incorporate analysis of alternatives	Competing explanations kept open, with the evidence that would discriminate among them
5. Demonstrate relevance to the decision	The question the assessment bears on and what changes if the assessment is wrong
6. Use clear and logical argumentation	An inference chain a reader can follow from evidence to conclusion
7. Explain change to, or consistency of, judgments	What changed since the last assessment and why, or why nothing changed despite new reporting
8. Make accurate judgments	Dated, resolvable assessments that can be scored against outcomes
9. Use effective visual information	Where a visual is used, one that shows the structure of the inference and where it is disputed

Standards are numbered as in ODNI (2015, as amended), § D.6.e, and paraphrased. The right-hand column is a proposed application to AI research.

The standards connect qualities that AI research usually treats separately. A clear answer needs adequate support, an uncertainty statement needs a stated basis, an alternative needs to bear on the inquiry, and a change of assessment needs a reason. Even the words of estimative probability are institutional choices that a system has to carry explicitly. A probability of 0.78 is "likely" on the ICD-203 scale, whose band runs from 55 to 80 per cent, and "very likely" on the Canadian Centre for Cyber Security's scale, which the programme also carries and whose band runs from 75 to 89 per cent (Canadian Centre for Cyber Security, n.d.). A summary that reported the word without the scale would erase a difference that two allied agencies have each decided matters. The 2011 United States supervisory guidance on model risk requires validation, monitoring, governance and effective challenge of the models on which financial institutions rely (Board of Governors and OCC, 2011), and read with ICD-203 it shows four levels at which judgment is exercised: a judgment about the world, a judgment about the quality of that assessment, a judgment about the reliability of whoever assessed it, and the governance that settles who may rely on what. Section 4.2 builds on this hierarchy.

Research on human judgment supplies findings and cautions. Mandel and Barnes (2014) found 1,514 strategic intelligence forecasts well calibrated against outcomes, and Mellers et al. (2014) found improvements associated with training, teamwork and the identification of skilled geopolitical forecasters. Coulthart (2017) found limited empirical support for several widely used structured analytic techniques, and individual ratings of 64 hypothetical intelligence reports against an ICD-derived rubric had poor reliability until aggregated (Marcoci et al., 2019).

These findings support a reciprocal opportunity. Analytic practitioners can contribute methods, difficult cases and professional criticism, and AI research can provide repeatable comparisons and controlled changes to procedures. Whether separating assumption from judgment improves a conclusion, whether an alternatives check finds anything and whether explaining a change makes the next assessment better have been hard questions to answer with human analysts and scarce products, and they become tractable when the analyst is a system whose procedures can be switched on and off. The adaptation has boundaries: intelligence analysis supplies assessments to decision-makers, agents that take actions require additional rules concerning permission and consequences,

and scientific, commercial and public-sector work carry different obligations.

3. Capabilities of a judgment layer

3.1 Searching for useful insights

Strategic insight concerns a relationship, explanation, risk or opportunity that changes what a particular audience can understand or do. Originality alone is a poor criterion, since a surprising claim may be familiar to a specialist, unsupported by the evidence or irrelevant to the decision. The programme's entry test is demanding: an insight earns its place only if it rests on something a thoughtful analyst could not readily have done alone under realistic limits of time and memory, such as holding hundreds of decomposed assumptions in view over months, noticing that several confirmations share one upstream cause, or noticing that a regular source has fallen silent.

The insight architecture develops explicit search patterns. It looks for a constraint that creates an advantage for some actors, a shared assumption under pressure, an analogy with a similar causal structure, an expected signal that has failed to appear, or a consequence that emerges through another actor's response. The vocabulary of patterns is closed, so that a claim outside it is by definition not the kind of contribution the system is scored on, and a new pattern enters only if it exploits one of ten named advantages a system has over an analyst working alone. McCone's question in 1962 had the form the programme asks for: it rested on a dependency the estimate had not traced, since air defences imply something to defend, it named a rival explanation, it identified the observation that would discriminate, which is what the analysis of competing hypotheses asks of an alternative (Heuer, 1999), and it was answerable within weeks.

The architecture separates candidate generation, criticism, source checking and continued evaluation, with a gate between each stage and the next. The stage that writes an insight may add no name, number or date that the search stage did not supply. An adversarial critic on a different model, instructed that the default verdict is failure, asks whether the cited material supports the claim, and a mechanical audit confirms that every cited record exists as cited. A ranker enforces diversity so that no single pattern dominates a day's output. Insights that survive are held as positions with a stated conviction, a probability that the claimed mechanism holds which is scored like a forecast, and with dated indicators that would confirm or refute them. A system producing different clever claims every day with no continuity is producing commentary. The research question is how much useful work assistance adds under realistic constraints on time, attention and cost, given that a larger volume of generated ideas also creates more work for reviewers.

3.2 Developing strategic understanding

Strategic understanding concerns how actors, causes and possible futures fit together. An adequate assessment needs an account of interaction, incentives and constraints, since other participants can change the setting and the assessing organization's own intervention may alter the behaviour a forecast depends on. An actor model distinguishes declared aims, observed actions and inferred motives. A delayed investment could reflect weak demand or a decision to wait for policy clarity, and a government's public assurance could reflect its intention or its wish to be believed while it does the opposite, as Moscow's September 1962 statement did. The system preserves both accounts and identifies the observations that would distinguish them, and representing several perspectives also prevents the analyst's own assumptions from being silently attributed to every participant, which is the failure Kent described.

The programme operates revisable models of this kind for fifteen states and for the principal firms in one industry, and is built to test them by pre-registered forecasts of what each actor will do, elicited in pairs at the same moment with and without the actor model so that the model's contribution is isolated. The paired forecasting is implemented and currently paused, with forty-eight pairs awaiting resolution and none yet scored. A census of the evidence available to these models found the state models better supplied than the company models, because governments take more discrete, dated and publicly archived actions. Scenarios extend this understanding into possible futures, where a system can help locate a consequence missing from every scenario or an assumption on which all of them depend.

3.3 Representing an inquiry

The programme's shared vocabulary lets people and systems express the judgments involved in an inquiry in a common form. An ontology, in this sense, specifies the kinds of things being represented and how they relate, and here those things include claims, evidence, alternative explanations and questions that remain unresolved. The general vocabulary contains nine elements (Table 2). Whether each is necessary is a research question with a first answer, reported in Section 5. Their sufficiency has been tested by construction: a chest-pain differential diagnosis, a negligence determination, a peer-review recommendation and a joint-doctrine risk assessment have each been expressed from the nine elements alone, and eighteen concepts from board governance reduce to the same nine under a mapping checked on every change to the software.

Table 2. A common vocabulary for judgments under uncertainty

Element	Question it helps answer
Claim	What is being asserted or considered?
Evidence	What supports or challenges it?
Source	Where did the information originate, and how reliable is that origin?
Alternative	What other explanation or possibility remains open?
Uncertainty estimate	How likely is the claim, and what limits confidence in its basis?
Reason for change	Why was the assessment revised or reconsidered?
Attribution	Who advanced this assessment, on what reasoning, and how can that reasoning be contested?
Decision frame	Which question, options and stakes give the inquiry its purpose?
Information gap	What remains unknown that could matter?

Source and attribution are different kinds of provenance. Source records where a piece of evidence came from, and attribution records who made a judgment about it. A newspaper reporting a company's claim about demand establishes that the company made the claim, and the system still has to assess what the report establishes about demand and record who made that assessment, which prevents an interested party's statement from quietly becoming an independently supported finding. Strategic extensions add actors, goals, relationships, assumptions, time horizons and the observable developments that would prompt reconsideration. The reference specifications provide machine-readable schemas and validation rules, published with a public validator, and a handoff format under which one agent can pass a judgment artifact to another with the expected output and the authority granted stated in the envelope.

3.4 Connecting understanding to decisions

Assistance should make the relationships among options, objectives and evidence available to the people responsible for deciding: comparing options against stated objectives, calculating consequences under explicit assumptions and identifying where the preferred action changes. Numerical precision has to reflect the quality of the inputs, and the programme's computational layer, built but not yet running on live data, records for each numerical sub-claim whether it rests on a mechanism, a reference class, a bound from known constraints or nothing. Participants may disagree about facts, interpretations, acceptable risk or whose interests deserve priority, and a synthesis should preserve consequential disagreement. The authority to settle a contested objective remains with the appropriate people and institutions, and the programme's rule for advice follows from this: a recommendation is admissible only when the decision-maker has asked for one, only within that person's stated question, deadline, authority and appetite for risk, and it is tied to a resolution criterion so that advice is scored like a forecast.

Where an agent can act, its permissions have to be tied to the scope of the task. Approval to investigate an option can permit inquiry while a decision remains open, and a consequential commitment may require a current assessment and a designated review. In the reference implementation these controls exist. A candidate insight enters in quarantine and can be surfaced, confirmed, falsified or retired only through governed operations that require a recorded reason and respect a daily budget, the highest-consequence action is staged until a person commits it, and every interface an agent can call carries a level of authority that refuses unless it has been explicitly raised.

3.5 Learning across inquiries

Experience can improve the questions an organization asks, the explanations it considers and the methods it trusts. Judgment-centric work is defined by three conditions: feedback about an assessment's quality is delayed, ambiguous or contested, uncertainty that matters remains at the moment of decision, and the assessment informs a consequential decision for which someone is responsible. Organizations doing such work rely on both live judgment, which reconsiders a situation in context, and frozen judgment, embodied in a model, routine or rule. Frozen judgment makes expertise reusable, and seven decades of evidence on clinical versus statistical prediction say it is often the better performer (Meehl, 1954; Grove et al., 2000), until the conditions that produced it change. The distinction is between modes, and a language model is itself a hybrid, fixed weights whose answers vary with context, retrieval and memory. Learning could improve the retrieval of relevant cases, the selection of critics and the calibration of forecasts, and each needs its own feedback, because an outcome can resolve a forecast while leaving its causal explanation disputed.

Changes of this kind need evaluation on later cases, including unfamiliar topics and different users, and they need to be prevented from oscillating. The programme treats this as a design problem with named parts: a requirement that a signal persist for several consecutive periods before a threshold moves, a minimum interval between changes, a watchdog that reverts a change when quality degrades, and a daily budget on the number of changes any subsystem may make. No change to the forecasting method is adopted unless it beats the current method on paired questions resolved after the candidate was sealed, with the realized statistical power disclosed and a person making the final call. Sealed has a specific meaning throughout this paper: a record is hashed when it is written, the day's hashes are folded into a single root, and that root is published where a third party can check it, so that anyone can later confirm that a judgment was recorded before its outcome and has not been altered since. The first statistically sound win under that gate was

reversed on inspection, because the candidate had been handed part of the market price it was scored against and lost to copying that price outright, and the reversal was recorded beside the original in the same form.

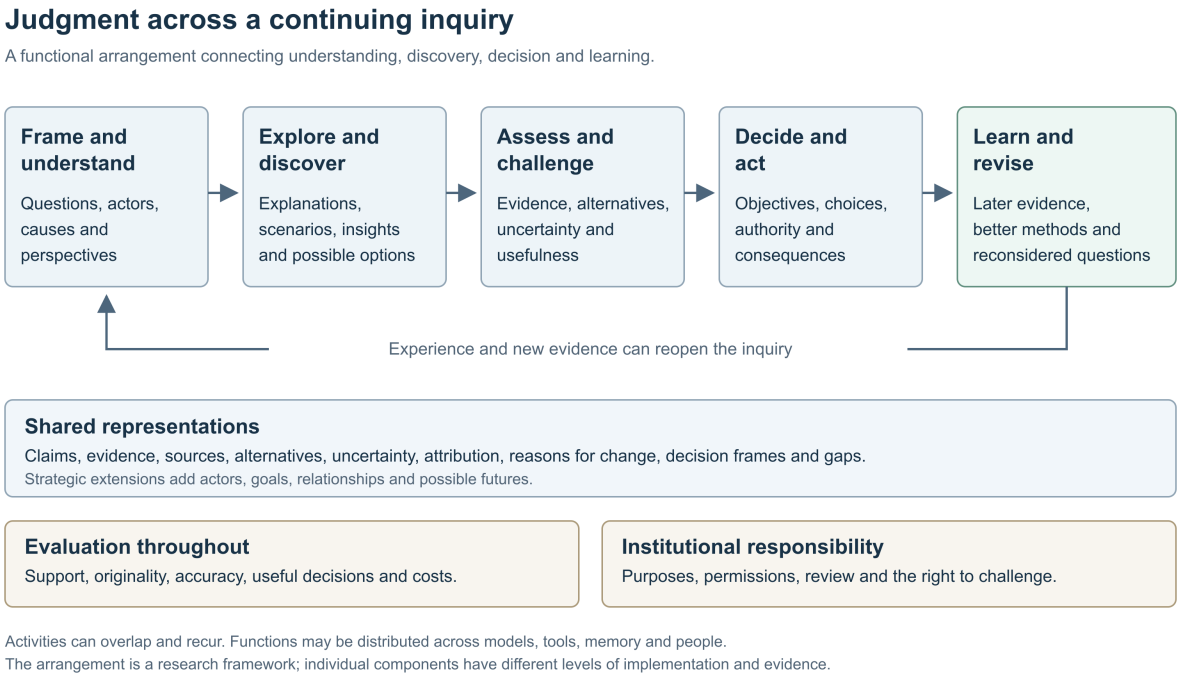

Figure 1. Judgment across a continuing inquiry. Activities can recur and run in parallel, new evidence can reopen any of them, and evaluation and authority apply throughout. The arrangement describes functions and relationships, with no claim that the complete system has been validated.

4. Evaluating the contribution to judgment

4.1 Forms of evidence

The programme's benchmark scores five broad dimensions, reasoning under uncertainty, recognition of limits, generation of alternatives, belief revision across sessions and integration of information from different sources, with domain-specific tasks in geopolitical, corporate, technological and financial analysis. A forecast can be accurate while its explanation is unsupported, and an explanation reasonable while its prediction fails, so the dimensions are reported separately, and the benchmark caps its composite when selected judgment dimensions fall below a threshold, publishing the raw and the capped composites together.

Forecasting supplies a valuable outcome-based measure, and its methodology is less settled than its numbers suggest. Metaculus's Spring 2026 analysis reported a narrow, statistically inconclusive aggregate difference between leading bots and professional forecasters (Wilson and Bash, 2026). The programme's own work found that comparing a forecaster with market prices taken near resolution, instead of thirty days out, inflates the market's apparent accuracy about 4.4-fold, so that evaluation design can dominate a published result.

Adjudication is a judgment in its own right, and the programme learned this by measurement. A criterion for deciding whether a claim has come true has to discriminate the claim from the

ordinary behaviour of the world. A criterion labelled as a three-standard-deviation event but scanned hourly for thirty days has 526 chances to fire and does so about half the time when nothing has happened, and several indicators on electricity demand were being satisfied every night by the ordinary overnight trough. Eighty-three such criteria across fifty-one open assessments were withdrawn as sealed, dated retractions, and the thresholds that now gate new criteria were fixed before they met live data so that they could not be tuned to the known failures. A meta-analysis finds that human–AI combinations often perform worse than the better of the two alone, especially on decision tasks (Vaccaro, Almaatouq and Malone, 2024), so any claimed benefit of assistance has to clear that bar.

4.2 Evaluating the evaluators

A model critic can share the generator's assumptions, prefer a particular writing style or confuse elaboration with quality, and agreement among several critics can reflect a common mistake. The programme's two model judges agreed with each other only slightly beyond chance in a pilot audit (Cohen's kappa 0.13 to 0.30), with a measurable preference for longer outputs, which is why rubric-judged scores are held as shadow scores and every property that admits deterministic scoring is scored without a model judge.

The programme has turned this audit into a control. A bias audit of the production critic found near-zero agreement with a second rater, a strong preference for longer outputs and a preference for its own model family, and registered its reliability at 0.55. That falls below the 0.60 threshold at which, in the reference implementation, an adverse verdict is no longer accepted on its own. It is referred to a second critic from a different model family and settled by consensus, the referral is sealed, and the critic's reliability is thereafter updated from whether its verdicts were borne out by outcomes. A per-run budget of three referrals, beyond which an adverse verdict is deferred to an authorized person, keeps the recursion finite. Review examines false rejection as well as mistaken acceptance, since excessive caution can eliminate the discoveries the system is meant to help produce.

Inspectable reasoning also needs a behavioural test, and the programme grades what a record can show on a ladder. Its first rung, structural coverage, asks whether the required elements are present, and the second, relational integrity, asks whether a cited claim rests on the cited evidence. Section 5 reports the programme's principal result on this ladder: a record can clear the first rung completely and fail the second completely on the same document. Upper rungs, behavioural correspondence and causal sensitivity to a changed premise, have no scorer yet, and work on monitoring expressed reasoning identifies the same limit (Korbak et al., 2025).

5. What exists, and what it shows

The programme's components exist at four levels, and a reader should hold each claim at its level. Specified means a manuscript or protocol states it. Implemented means reference software does it and tests confirm the properties of the software. Operating means it runs on live data. Measured means a study has estimated an effect, with the study's limits. Appendix B maps eleven components to these levels. The philosophical account and the organizational theory are specified. The shared vocabulary with its schema and validator, the insight architecture, the evaluation instruments, the governed-action controls and the learning loops are implemented, and the insight architecture, the sealed record, two reference feeds of sealed judgments on AI compute and on AI policy and economics, and the actor models are operating, fed by roughly 3,800 observations a day from

eighteen open sources. Several parts exist without running: the trial that delivered assessments to users is suspended, the computational layer is inactive, the paired test of the actor models is paused, the two reference feeds were withdrawn from public view in September 2026, and many scheduled background processes do not run. Appendix B records these states, since a map that showed only what runs would overstate the programme.

Six studies conducted within the programme illustrate what can already be learned at the measured level. None has been independently replicated, and Appendix A gives methods and limits for each.

First, a comparison across fifteen evidence contexts examined structured and unstructured insight generation with the same model, inputs and output budget. A separate model evaluator assigned higher scores to the structured condition for evidential grounding and falsifiability and found no improvement in decision relevance or non-obviousness. Support and testability are therefore separable from the qualities that make an insight worth discovering. The full generation-and-criticism architecture has not yet been compared with a plain model call, and the condition for abandoning it is fixed in advance: if the full process does not beat a plain model on at least three of four dimensions, the architecture is unjustified.

Second, a corrected forecasting comparison across 344 resolved questions found no measurable accuracy advantage from the tested combination of retrieval, structured prompting and deliberation. An earlier reported deficit was traced to failed outputs being counted as forecasts, and the repaired comparison has a limitation of its own, since the structured arm was rerun after the questions had resolved. The wider record points the same way: realistic retrieval added about 2.6 per cent, a larger model of the same family did not move the score, and a multi-model deliberation panel scored worse than a coin flip on 143 resolved questions and was retired, while post-hoc calibration on 1,160 crowd forecasts cut expected calibration error by 43 per cent. Calibration is learnable at this scale and discrimination is not.

Third, two independently written validators agreed on all sixty judgments of whether a test collection met the published structural rules. This is an engineering result about the consistent interpretation of a specification, and it says nothing about the truth or usefulness of the represented assessments.

Fourth, a pre-registered element-removal study asked whether each of the nine elements in Table 2 is load-bearing. Under a rule fixed in advance, eight of the nine were individually load-bearing at sixty forecasting questions, on a measure adopted after the registered measures showed nothing: the forecaster's point probability did not depend on which fields accompanied it, and the effects reported are on a single model judge's capability scores that sit close to the fields themselves, which Section 8 notes is circular. Attribution was the exception. A detector excluded runs in which the model rebuilt a removed element in prose, and evidence could rarely be removed cleanly at all because the model supplied it unprompted, a finding Section 6.3 returns to.

Fifth, six widely used models from three vendors, given each task with no instruction to report sources, alternatives or gaps, were scored on fifty entry-level tasks by a judge from outside the candidate set, element by element. Every one was deficient or absent, on some tasks, at naming the source a claim rested on, at keeping competing explanations open and at stating what was not known. One model fell below the minimum-competence threshold, and on harder geopolitical tasks it failed seven of nine elements while producing fluent strategic prose. This is a conformance finding: a structured-output instruction would close part of the gap, and whether their judgments were worse is a separate question. An earlier run in which candidate models also served as judges

had scored the weakest model about sixteen points higher, which is the measured size of the self-preference bias the evaluation procedure now excludes.

Sixth, on four publicly available board and committee packs, the architecture was compared with the best evidence-first prompt I could write on the same model, passed through the identical citation validator and vocabulary filters. The prompt retained seven, one, four and zero usable questions as the documents grew from five thousand to nearly seven hundred thousand characters, while the architecture retained ten on each at the same 100 per cent citation validity, because it extracts and validates claims before it generates questions. The same packs then supplied the measurement that weighs most heavily against the architecture. When forty of its own citations were audited by a model from a different family, 67.5 per cent resolved to real passages under a strict check, and 56.3 per cent of those supported the claim attached to them. On the largest pack every citation resolved and none supported its claim. A record can be fully well-formed and evidentially empty.

I read these results as a redirection. Structure alone secures grounding and testability, which a record needs if it is to be checked, and has not yet secured relevance or originality, which make a discovery worth having. If the representation on its own does not improve judgment, the hypothesis has to rest on the connections, on a critic that kills weak candidates, on outcomes that adjudicate which lessons were real, and on the use of those lessons to choose the next question. That is the claim Section 6 states and the study in Section 6.2 tests.

6. Research contribution and responsible use

6.1 Connecting inquiry with learning

The central architectural claim is that explicit relationships among understanding, discovery and evaluation can make experience useful to the next inquiry, through three operations. Representation directs discovery: when an account preserves rival explanations and the evidence bearing on them, a search can target what would distinguish them. Discovery revises the account: a new insight that reveals an overlooked actor, a changing constraint or an option the framing excluded returns to the same evidence and alternatives against which existing assessments are judged. Evaluation distinguishes the lessons worth carrying forward: a well-supported insight may owe its value to a useful analogy or an informative question, a failed one may reveal a weak search pattern or an evaluator's blind spot, and telling these apart is what could improve the selection of methods on later cases. In the programme's implementation, when a judgment resolves its share of the error is apportioned backward across the assumptions, sources and search methods that produced it, sealed and never applied automatically. The method's history so far is a record of failures caught by its own controls, including a synthetic test with a known answer on which it did no better than crediting each source in proportion to how often it was cited.

The Cuba estimate shows what a lesson of this kind looks like and why recording it is not enough. The judgment that corrected the estimate came from asking what the evidence was protecting, which turned a rival explanation into a collection target. Kent recorded a version of the lesson two years later, an estimate twenty years on rested on a similar assumption, and the record was not retrieved at the moment the later inquiry formed its question. The programme's proposal is to record such lessons with their provenance, the question that helped, the evidence that proved independent, the indicator that proved informative and the objection that proved mistaken, to adjudicate them against what happened, and to make them retrievable when a later inquiry in a different domain arrives, so that the question of what the evidence is protecting is available to a

company reading a competitor's sudden investment in a capability it has no visible use for. Whether the retrieved lesson improves the new inquiry is the empirical question. Figure 2 shows the loop.

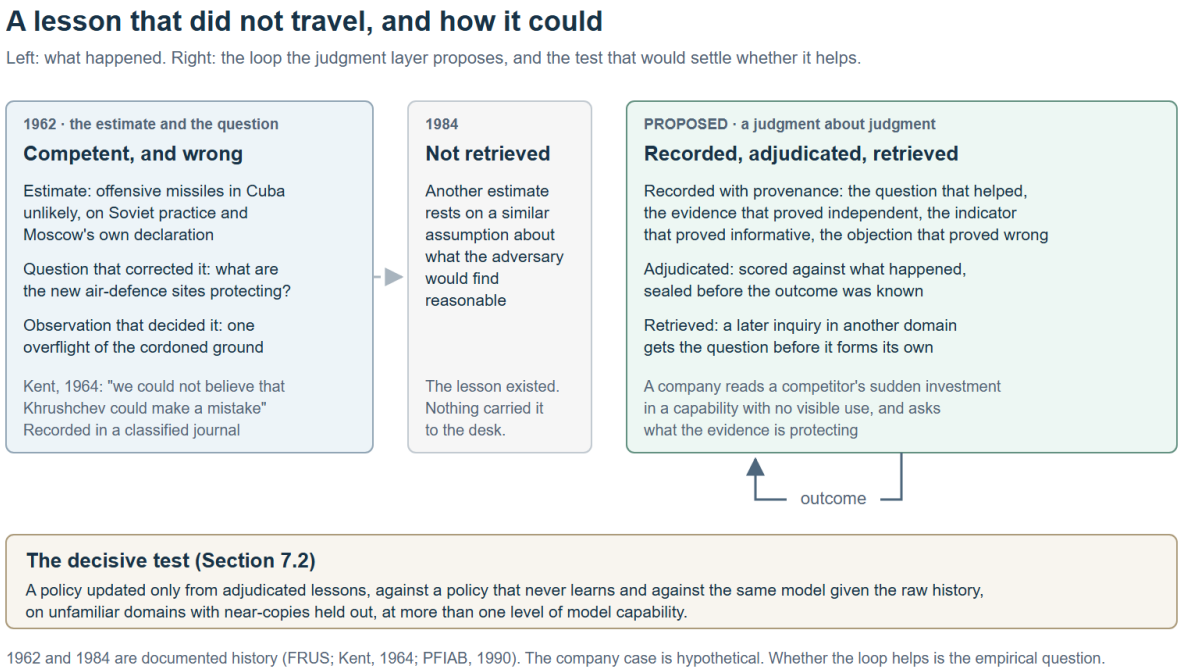

Figure 2. A lesson that did not travel, and how it could. Left: the 1962 episode, the question that corrected it, and Kent's recorded diagnosis, known but not applied in 1984. Right: the proposed loop, in which a lesson is recorded as a judgment about judgment with the question, the evidence and the indicator that proved informative, adjudicated against the outcome, and retrieved with its provenance before a later inquiry in a different domain forms its question.

The layer draws on bounded rationality (Simon, 1955), metareasoning (Russell and Wefald, 1991), formal belief revision (Alchourrón, Gärdenfors and Makinson, 1985), which restores consistency and is indifferent to who supplied a proposition and what was at stake, belief–desire–intention architectures (Rao and Georgeff, 1995), which treat a belief as an internal state with no one answerable for it, and organizational sensemaking (Weick, Sutcliffe and Obstfeld, 2005). Table 3 states what it adds to the implemented precedents.

Table 3. What the layer adds to its nearest implemented precedents

Approach	How lessons are produced	Scope of reuse	Evaluator reliability carried	Survives a change of model
Reflexion (Shinn et al., 2023)	Self-graded reflection on a failed attempt	The same task, retried	No	No
Voyager (Wang et al., 2023)	Execution feedback	The same environment	No	A library of skills only
Raw history in a long context	None, a transcript	Any	No	Yes, unadjudicated
Structured outputs from a model interface	A schema only	Not applicable	No	The form only

Approach	How lessons are produced	Scope of reuse	Evaluator reliability carried	Survives a change of model
DIVE (Friedman et al., 2020)	Provenance and appraisals	One analysis session	No	Not addressed
Truth maintenance (Doyle, 1979)	Reasons recorded for beliefs	One reasoner	No	Not addressed
Judgment layer (this paper)	Lessons independently adjudicated against outcomes	A different domain	Yes, a registry updated from outcomes	Yes, the record persists outside the model

Characterizations follow the cited works' own descriptions. The last row is a design claim, tested in Section 6.2.

6.2 The decisive test

The study that would settle the central claim compares two inquiry policies on the same continuing tasks. Under the fixed policy, the system's methods for choosing questions, retrieving cases, selecting critics and weighing sources stay constant across episodes. Under the learning policy, they are updated only from episodes whose outcomes have been independently adjudicated, and only through lessons expressed as judgments about judgment: which reframing helped, which analogy survived criticism, which evaluator's rejections proved mistaken. Both policies use the same model, evidence access, tools and budget, so that any difference is attributable to what was learned and how it was used.

Three further arms answer the objection of Section 6.3. A third policy gives the same model the organization's full, unadjudicated history in its context window and no curated lessons, which is the rival that a long-context model supplies for free. A fourth policy is updated from the same episodes through lessons the system graded itself, without independent adjudication, and the comparison of the third and fourth policies with the learning policy isolates adjudication as the active ingredient. The margin of the learning policy over the fixed policy is also measured at two or more levels of base-model capability, so that a margin which shrinks as the model improves counts against the layer. A preliminary arm, cheap enough to run first, gives current models a decontaminated analogue of the September 1962 evidence base and scores whether any of them asks what the air-defence sites protect, since every model has read about the crisis itself.

Evaluation takes place on later episodes, on at least one domain absent from development, with closely related cases held together so that the learning policy is never tested on a near-copy of an episode it learned from. The primary measures are useful reframings and informative investigations identified, as judged by domain specialists blind to the condition, and prospective forecast accuracy where questions resolve. Unsupported changes to sound assessments, review burden and computational cost are reported alongside. The claim is supported if the learning policy improves on the fixed policy on unfamiliar cases at acceptable cost. It is undermined if the gains vanish outside the development domain, if the raw-history rival matches them, if they shrink as the model improves, or if the policy learns to produce what evaluators reward in place of what decisions need. The number of episodes and domains, the power to detect a stated difference, the specialist judges and the effect that counts as improvement are fixed before the first episode is scored.

The study's binding constraint is the supply of adjudicated outcomes, and the programme has measured it. Of fifty-five live claims in one reference feed, only one could be resolved by any

instrument the programme holds, because two-thirds of the claims concern quantities that no instrument settles on causal grounds, and a general news corpus corroborated none of forty-four company questions in another feed. The programme therefore treats judged reading of retrieved text as the primary route to resolution and has posted a first resolvable question on a public prediction market, and the timing of the decisive test depends on that supply. The study sits inside a wider set of twenty-one pre-specified propositions in the organizational companion work, each with a named comparison, a primary outcome and a statement of what would disconfirm it.

6.3 Whether better models make the layer redundant

A standard objection holds that models will improve at judgment until the explicit architecture of Sections 2 to 4, the nine-element record, the search patterns, the critic, the reliance controller and the sealed chain, stops adding value. The objection has to be met on its real ground, which is whether making judgment explicit adds value that improvement in the model erodes.

Part of it is conceded. The forecasting comparison of Section 5 found no accuracy advantage from retrieval, structured prompting and deliberation, so there was no accuracy value for improvement to erode. The element-removal study found that models rebuild a removed element in prose often enough that clean ablation became scarce, which shows that the element already lives in the weights. Scaffolds for accuracy do wash out, and where an answer can be checked cheaply, producing it will get cheaper still. The companion organizational work turns the remainder into testable predictions: that the cost of generating analysis falls faster than the cost of producing a validated evaluation of it, and that the net benefit of judgment-governance procedures holds or grows as generating-model capability improves, with a sustained decline across capability gains counting as disconfirmation. A third prediction states the condition under which the objection would be right: the category of judgment-centric work could shrink even if the per-task benefit persists.

Three parts of the layer's claimed value then have separate fates. The first is substantive improvement in judgment, in better questions, preserved alternatives and lessons carried between inquiries. This is the claim that model improvement threatens most directly, and the programme's evidence is at fixed capability only: six widely used models, unprompted, were deficient on sources, alternatives and information gaps, and on four public documents the same model under the same prompt and filters kept between zero and seven usable questions against the architecture's ten. Whether a stronger model closes those gaps unaided is what the model-swap arm of Section 6.2 measures, and no result exists yet. The second is checkability, where the objection is partly right. A structured-output mode on a model's interface supplies a checkable form, and a deterministic check removes fabricated citations without any layer. What a model interface does not supply is adjudication against outcomes, sealed timestamps that prove a judgment preceded its resolution, a registry of how reliable each evaluator has proved, and portability of all of this across a change of model. Those are properties of a record that persists outside the model, and they are what makes a recorded lesson trustworthy later. The third is authority. A model can record attribution and disagreement, and it cannot be the accountable party, since whether to act on an assessment is a further judgment that two institutions can reasonably make differently at the same probability. I do not rest the case on this, though it does mean that the record, the permissions and the review will exist in any deployment that matters, designed or improvised.

The cumulative-judgment claim faces a rival the programme has not yet tested, a capable model given the organization's raw history in a long context, and Section 6.2 includes the arm that separates them. An adjudicated, provenance-bearing lesson says which question helped and which

evaluator was wrong, has been scored against what happened, and survives the replacement of the model that produced it. Whether that is worth its cost at the next capability level is the empirical question.

6.4 Capability, safety and institutional responsibility

Several judgment functions have both constructive and protective uses. Recognizing an uncertain assumption can lead to a better experiment and prevent an excessive commitment, and preserving disagreement can improve a decision and prevent a fluent summary from creating unwarranted consensus. The benefits can also diverge: broader search may generate valuable options and more unsupported proposals, stronger strategic reasoning can advance harmful objectives, and strict review can block errors and suppress worthwhile ideas. Evaluations therefore need separate measures of useful contributions, unjustified acceptance, lost opportunities and human control. AI-control research provides one comparison, since monitoring and selective deferral have been studied in coding tasks involving deliberately introduced backdoors (Greenblatt et al., 2023), and a judgment architecture needs corresponding attention to misleading evidence, fabricated support, manipulated review and attempts to exceed authority.

One claim in this area is testable now, and the mechanism it concerns is built. Keeping a generated possibility, an accepted assessment and an authorized action distinct, and enforcing the last at the point where a tool is used, should allow a system to explore more widely without acting on more of what it explores. Controls of this kind have already had a cost, since a two-source corroboration requirement routed every automatically settleable question to a human review queue before any surface existed for a person to clear it. The comparison is between otherwise identical agents with and without the separation, on tasks that reward discovery and penalize unauthorized or unsupported action. If exploration rises and unsupported action does not, the separation is justified, and if review merely suppresses good ideas, it is not.

The institution determines whose interests count, what authority is delegated and how a decision can be challenged. Rewarding agreement, output volume or acceptance rates can steer a system away from better judgment, so a record should separate what a user asked for, what the institution authorized, what the system inferred and what it was rewarded for. The original question set should be retained so that drift toward easier questions is visible, and candidates not adopted should be recorded, because an organization observes the consequences only of the actions it took.

7. Research and applied opportunities

7.1 Work the approach could enable

The practical opportunity is to make forms of inquiry feasible that currently exceed a team's available time or attention. In scientific research, assistance could help identify a test that distinguishes explanations, recover a relevant mechanism from another discipline or recognize why an earlier line of inquiry failed. In organizations, a developing strategic understanding could connect the views of teams that examine different parts of a situation, tracing how a policy change alters an actor's incentives, the plausibility of a scenario and the timing of an investment. On pre-decisional policy papers of the kind a foreign ministry prepares, an ICD-derived judge scored the programme's full pipeline at 82.9 against 55.7 for a single model call, with the largest gaps on stated assumptions, source characterization and alternatives, and it caught the unassisted model issuing policy recommendations, which the pipeline prohibits structurally. The judge was single and the papers

were author-generated, so the result shows that the standards can be enforced and scored and says nothing yet about professional usefulness.

An AI Futures Judgment Ledger of dated forecasts, alternative explanations and reasons for revision exists and is accumulating entries that are not yet public. It is the first form of a proposed AI Futures Observatory, a shared body of consequential questions about AI's capabilities, adoption and consequences on which researchers with different expectations can challenge one another's reasoning. Board applications have completed their machine-layer evaluations, and the human study of whether assistance improves preparation, substantive challenge and appropriate reliance is fully specified and has not yet been run with a participating board.

7.2 A shared research programme

Collaboration offers the opportunity to shape how judgment is represented, developed and evaluated in AI, and the existing vocabulary, benchmark, insight methods and operating feeds provide starting points. Researchers in AI, cognition and decision science will find the open questions at the connection between a representation and the inquiry it enables: which distinctions help an agent find a useful question, and how feedback can improve its choice of explanation or investigation on unfamiliar cases. Evaluation researchers and domain experts are needed to define and measure the qualities that make an insight useful, and to reproduce the initial findings independently. Intelligence analysts and tradecraft researchers have the offer of Section 2: contribute difficult cases, historical assessment sets or expert raters, and receive in return the experimental setting in which the effect of a tradecraft standard can be measured. AI developers and safety researchers have a setting in which useful exploration and warranted reliance can be examined together.

The first project is the study of Section 6.2, run at small scale in the first half of 2027 on the programme's existing evaluation tasks and a bounded set of continuing episodes in geopolitical and technology assessment. A collaborating group contributes one of three things: a competing system or a simpler representation to run under the same test conditions, an appropriately shareable set of episodes with independent reviewers, or adversarial episodes designed to make the learning policy learn the wrong lesson. The programme supplies the tasks, the test conditions, the adjudication procedure and the analysis, and both parties receive an independently assessed account of what the assistance added, what it cost and where it failed. Data access, confidentiality and publication are agreed for each project, and human-participant research follows the applicable institutional processes.

A board, ministry or agency contributes material in place of models: board or committee packs and the decision records behind them, practitioner raters, or a venue for the specified human study. It receives an independently assessed account of what the assistance found, missed and got wrong on its own documents, at a cost of a few hours of rater attention per pack. Documents stay with the institution, results are released only in a form it has approved, and an early participant chooses the cases and the measures before the protocol is fixed.

Anyone can start with what is already public. A reader can validate an artifact against the published vocabulary, apply the sufficiency test that says when a plain model call with a tight schema is enough and no layer is needed, pass a judgment between two agents under the published handoff format, or download a sealed record and confirm offline that a judgment was recorded before its outcome. The locations are listed under Research materials. The twenty-five-task starter evaluation, which a new system can complete within a working day, and the governance walkthrough, which runs one decision through the permission controls with no model call, are available on request until their

public release. Readers who want to contribute, challenge the design or run the comparison on their own systems are invited to write to me at the address above.

8. Limits

The integrated architecture is unvalidated. Every measured result in Section 5 was produced within the programme, most under a single model judge, and none has been independently replicated. The two board results with the largest effects have not yet been added to the programme's archived evaluation record and should be re-run or released before anyone relies on them. The element-removal study departed from its registered outcome measure, and the forecasting comparison carries a possible exposure to resolved outcomes. A substantial part of what has been built is switched off or paused.

A judgment record can be complete and wrong, and it can describe a process the model did not follow. Section 5 shows that a record can clear every structural check and support none of its claims, and the upper rungs of the fidelity ladder have no scorer. Measures that reward the fields the representation requires would validate the representation circularly, which is why the studies use outcomes that mean something outside it and why the programme's own evaluators are measured against outcomes. The supply of adjudicated outcomes is the binding constraint on the decisive test.

The safety claims are scoped to mistakes and unwarranted reliance. Competent judgment can serve harmful objectives, human approval does not make every action acceptable, and the architecture supplies no general guarantee against misuse or loss of control. Adoption of the analytic standards is no evidence of their effectiveness, and the tradecraft literature is the first to say so. The hypothesis of Section 6 would be undermined by simpler arrangements matching the linked representation, by a raw-history rival matching the learning policy, by a margin that shrinks as the base model improves, by review that suppresses good ideas, or by learning that fails to transfer beyond familiar cases. Each is a result the programme is designed to be able to produce, and each would be reported.

9. Conclusion

Judgment decides which questions receive attention, which explanations are pursued and which possibilities become grounds for action. For AI undertaking open-ended work, these choices are part of the capability being delegated. The central opportunity is cumulative improvement in judgment: an explicit account of evidence and alternatives can guide inquiry, discovery can change that account, and evaluation can identify lessons for subsequent work. ICD-203 provides the institutional precedent, and the programme's vocabulary, benchmark, insight architecture, sealed record and operating feeds give the research concrete material. Its measured results so far establish that the elements the standards require are the ones widely used models omit unprompted, that structure secures checkability without yet securing discovery, and that a record can pass every structural test and still support nothing. The hypothesis therefore rests on the connections, and Section 6.2 is the test.

In 1962 the estimate asked whether the adversary would do something the estimators found unreasonable, and the correction came from asking what the evidence was protecting. The lesson had been written down and was known inside the agency, and nobody applied it. A useful outcome of this research would be AI that helps people ask the second kind of question sooner, that recognizes

when several reports have a single source, and that carries what it learned to the next inquiry in a form that can be checked. These are substantial research problems with practical consequences, and they are better pursued together.

Terms used in this paper

Agent: a system that pursues a task through repeated use of information and tools. *Brier score*: the mean squared error of a probability forecast, lower being better, with 0.25 the score for always answering one half. *Calibration and discrimination*: whether stated probabilities match observed frequencies, and whether they separate events that happen from events that do not. *Closed pattern vocabulary*: the fixed list of kinds of insight the search may return. *Conviction*: the probability that an insight's claimed mechanism holds, scored like a forecast. *Coverage gate*: a check that a system's confidence is conditioned on how much of the evidence it examined. *Crosswalk*: a mapping from the programme's vocabulary to an external standard. *Element-competence score*: the mean of a system's scores on the nine elements of Table 2. *Fidelity ladder*: the levels at which a record can correspond to the reasoning that produced it, from structural coverage through relational integrity to causal sensitivity. *Handoff format*: the envelope in which one agent passes a judgment artifact to another. *Kappa*: agreement between raters beyond chance, 0 being chance and 1 perfect. *Locator*: a pointer into a source document that a citation must resolve to. *Promotion gate*: the test a candidate must pass to leave quarantine or to replace an incumbent method. *Reliance controller*: the rule that decides whether an evaluator's verdict is accepted, referred to a second evaluator or deferred to a person. *Sealed*: hashed when written, folded into a published daily root, and therefore checkable by a third party as having existed before its outcome. *Shadow score*: a model-judged score reported but not used for decisions. *Test conditions*: the fixed procedure that runs tasks, models and judges.

Appendix A. Internal studies: methods and limits

A.1 Structured insight generation

Fifteen shared evidence contexts from an AI-compute feed were supplied to Gemini 2.5 Flash under structured and plain-prompt conditions, with the same context and token budget. Each requested ten items. Claude Sonnet 4 evaluated the outputs with balanced presentation order and concealed condition identities. The reported 0–100 scores were 86.3 versus 58.0 for evidential grounding, a difference of 28.3 with a bootstrap 95% interval of [17, 41], and 53.3 versus 30.7 for falsifiability, a difference of 22.7 with interval [2, 40]. Decision-relevance and non-obviousness differences favoured the plain-prompt condition, with intervals spanning zero. Overall preference favoured the structured set in seven contexts and the plain-prompt set in eight. This small comparison used a single generator–judge pair, and the generator was the model that serves as the production critic, not the production writer. Visible differences in presentation prevent full blinding. The comparison isolated the generation scaffold and supplies no result for the complete operator-and-critic pipeline, human-assessed usefulness or sustained performance.

A.2 Forecasting with architectural scaffolding

The comparison used 344 matched resolved questions. After repair of a failure-handling defect, the structured arm's Brier score was 0.1743, compared with 0.1776 for the retained plain-model baseline, a difference of -0.0033 with a 95% interval of $[-0.0244, 0.0181]$. The result establishes no

accuracy advantage or equivalence. The original structured run had counted 140 failed outputs as forecasts of 0.50. The replacement run produced forecasts for all 348 questions attempted, with 344 entering the matched comparison, the plain-model arm was retained from the earlier run, and the July 2026 rerun used questions already resolved, leaving possible outcome exposure and differences in evaluation timing as limitations.

Related internal comparisons are reported as they stand and have not been independently reproduced. Realistic retrieval added about 2.6 per cent to accuracy, and an earlier figure of 20 per cent was an artifact of a curated question set. A comparison of two models of the same family on post-cutoff questions showed no material difference in Brier score, while the larger model was materially better at prose synthesis. A multi-model deliberation panel costing roughly fifty cents to a dollar per forecast scored a Brier of 0.2919 on 143 resolved questions, worse than the 0.25 obtained by always answering one half, and was retired. A pre-registered test of learning from worked examples on 120 questions returned a null ($p = 0.93$). A post-hoc calibration model fitted to 1,160 crowd forecasts recovered from a public prediction market reduced expected calibration error from 0.0489 to 0.0277 while reducing Brier score by 5.2 per cent. Taking market prices near resolution instead of thirty days out changed the market's Brier score from 0.122 to 0.028 on the same questions, an artifact of timing, and shrank the measured gap between system and market from 0.152 to 0.055.

A.3 Agreement on structural conformance

Two standards-compliant SHACL engines, shacl-engine and pySHACL, evaluated sixty instances against the published ontology rules after conversion through the common JSON-LD context. Both returned the same verdict on every instance: 45 passed and fifteen failed, three of the failures being zero-shot single-prompt instances produced as an ablation arm, each carrying about two violations. The historical TypeScript reference validator had accepted 57, and twelve of those were rejected by both standards-compliant engines, all on one constraint requiring an evidence item to link to a source. The published shapes still lack two minimum-count rules the reference validator enforces, so the two validators do not yet encode the same definition in both directions. An earlier vacuous all-pass result, caused by missing type information, was withdrawn after correction.

A.4 Removal of representational elements

A pre-registration fixed on 2 May 2026 specified the rule under which an element of the vocabulary would count as load-bearing: a confidence interval for the effect of its removal excluding zero, an effect size of at least 0.30, correction for multiple comparisons across the nine elements, and a minimum number of tasks on which the element had been removed cleanly. Runs on forty and then sixty resolved forecasting questions removed one element at a time from the output structure, with Gemini 2.5 Flash generating and a single model judge from a different family, Claude Sonnet 4, scoring specific capabilities. A three-layer detector excluded runs in which the removed element had been rebuilt in prose. Under the locked rule at sixty questions, eight of nine elements were load-bearing, with effect sizes ranging from about 0.4 for the information gap to about 2.8 for the uncertainty estimate, source at about 1.1, and attribution was the exception (effect size about 0.14) on single-judge event forecasts, where the distinction between provenance of evidence and provenance of judgment does little work. Evidence could be removed cleanly on only about 5 per cent of tasks at forty questions because the model supplied it unprompted.

The registered outcome measures were calibration, auditability and adaptability. They were replaced after it emerged that a capable forecaster's point probability did not depend on which fields

accompanied it, and the deviation is recorded in a public log. The findings are therefore confirmatory under the locked rule and exploratory with respect to the measure, they concern one task family and one generator, they rely on model judges, and the claim element was treated as definitional and was not experimentally removed. They do not establish that all nine elements are universally necessary.

A.5 Six widely used models on entry-level judgment tasks

Six models were evaluated on fifty tasks each (two models completed 48 and 43 of the fifty) drawn from the benchmark's first layer, stratified across its dimensions and difficulty levels: Claude Sonnet 4.5, Claude Haiku 4.5, Gemini 2.5 Pro, Gemini 2.5 Flash, GPT-4o and GPT-4o-mini, sampled at temperature 0.4. Candidates received a system instruction to answer concisely with explicit reasoning and to follow the requested format, and were not told to report sources, alternatives or gaps. The judge, Claude Opus 4.1, was disjoint from the candidate set, and the harness refuses to start if any judge is also a candidate. For each task the judge returned a dimension score and a band for each of the nine elements, mapped to numeric values with a failure threshold of 0.50, and an element-competence score was computed from the element scores with 90 per cent bootstrap intervals over the task sample.

Every model was deficient or absent on some tasks at three elements, source, alternative and information gap, which correspond to tradecraft standards 1, 4 and 2, and the two OpenAI models additionally failed uncertainty estimate and attribution. Element-competence scores ranged from 65.3 (Claude Sonnet 4.5) to 49.8 (GPT-4o), the last falling below the 50 threshold at which the benchmark caps a system's overall classification regardless of its other scores. On a second-layer geopolitical set run on the top and bottom models, the weaker model scored 37.9 and failed seven of nine elements while producing fluent strategic prose. An earlier run with three of the candidates serving as judges had scored GPT-4o's dimension mean at about 74.2 against 58.7 under the independent judge. The run is single-judge, author-run, stateless (so continuity across sessions was structurally untestable), and uses a first-publication rubric. Reproduction takes about thirty minutes per model with the published task pool and harness.

A.6 Architecture against guard-matched prompting, and citation support, on public board documents

Four publicly available governing-board and committee packs were used, ranging from about 4,800 to about 676,000 characters. The comparison condition ran the strongest evidence-first prompt I could write on the same model (Claude Opus 4.7), then passed its output through the architecture's own locator validator and its lists of banned phrases and banned meanings, so that any remaining difference is attributable to the architecture and not to the guards. Raw output was ten questions per pack in both conditions. After the guards, the prompt retained seven, one, four and zero questions in order of pack size, with the locator validator doing all the filtering, and the architecture retained ten on each pack at 100 per cent locator validity, because it drops claims with unresolvable locators at extraction and generates questions only over the surviving claims. The relationship between document length and collapse is monotonic in this sample of four.

Separately, forty claim–locator–evidence triples from the architecture's stored output, ten per pack, were judged by Gemini 3.1 Pro, a different model family from the generator. Under a strict substring check 67.5 per cent of locators resolved, against the in-pipeline validator's more permissive acceptance, and of the resolved triples 56.3 per cent were judged to support their claim, with no

contradictions. On the largest pack 100 per cent of locators resolved and 0 per cent supported the claim.

A related counterfactual showed that disabling the locator validator would have reported evidence grounding at 100 while 22 to 34 per cent of claims on the three larger packs carried locators that did not resolve to the document, and a planted undisclosed provision of £729 million was extracted, though not promoted to a finding, at a position about five thousand characters in and not at 340,000 or 627,000 characters, consistent with a 120,000-character extraction window. At the frontier, variance between runs of one model was comparable to variance across vendors on the largest pack, three of four models converged on the same score on the smallest, and a larger model of the same family did not move the score. All results are author-run, the audit judge is a single model, and the manual-coded anchor for the audit is reserved for peer review. The procedure and documents are archived, and the outputs of these two runs are not yet.

Appendix B. Programme components by level of existence

Specified means a manuscript or protocol states it. Implemented means reference software does it and tests confirm the properties of the software. Operating means it runs on live data. Measured means a study has estimated an effect, with the study's limits. The table records what is switched off or paused as well as what runs.

Table B1. Programme components by level of existence

Component	Level	What it establishes	Not yet established
Philosophical account of judgment: evidence as a relation, commitment as the unit, a threshold for what counts as judging	Specified	Definitions and propositions that can fail	Its necessity claims
Organizational theory: judgment-centric work, live and frozen judgment, four levels of judgment, twenty-one propositions with disconfirmation criteria	Specified	Institutional propositions specific enough to fail	Any of them empirically
Shared vocabulary of nine elements with strategic extensions, a published schema and validator, crosswalks to provenance and doctrine standards, and non-strategic instantiations	Implemented	That a checkable, exchangeable judgment record can be built and passed between agents	That it improves inquiry over simpler records
Insight architecture: closed pattern vocabulary, generation, adversarial critic, mechanical audit, ranker, a standing set of dated positions	Implemented and operating	A testable design for discovery with a track record	Full-pipeline effectiveness against a plain model
Evaluation instruments: two-layer benchmark, capped composite, evaluator-reliability controller, coverage gate, fidelity ladder	Implemented, partly measured	Instruments and a measurement discipline	Validity against professional performance, human validation of model-judged scores

Component	Level	What it establishes	Not yet established
Governed action: quarantine and promotion of insights, authority levels, a staging mechanism for irreversible actions adopted so far by one operation, advice as deferral	Implemented	That possibility, assessment and action can be kept apart at the tool boundary	That the separation changes exploration or reliance
Sealed record: forecasts, insights, adjudication verdicts and oversight-exercising reliance decisions hashed at creation, folded into a daily root and anchored in a public repository, with an offline verifier. A September 2026 audit found 43 per cent of forecast records unlinked to any seal and repaired the linkage and the anchoring path	Operating	That prospective, verifiable judgment records are feasible at low cost, and that their coverage has to be audited	Predictive advantage or decision usefulness
Reference feeds: two feeds of sealed judgments on AI compute and on AI policy and economics, published in May and July 2026 and withdrawn from public view in September 2026, and a third on AI futures accruing privately, fed by roughly 3,800 observations a day from eighteen open sources	Operating	A continuing, contested, consequential setting for the studies	Resolved outcomes in sufficient number
Actor models for fifteen states and one industry's principal firms, with a paired-forecast test with and without the model	Operating (models); paired test implemented, paused with 48 pairs unresolved	Third-person models that can be scored	Their contribution to forecast quality
Learning loops: credit assigned backward from adjudicated outcomes with negative controls, source-quality updating, a promotion gate, rules that slow and bound updating	Implemented; few loops have completed a cycle	Mechanisms with their own audit trail	Improvement on later inquiries, which depends on the supply of adjudicated outcomes
Applied profiles: board and decision-forum oversight, pre-decisional policy papers, governance attestation, an analyst's desk check against ICD-203	Specified and implemented, machine-layer evaluations run	Study designs and instruments	Professional usefulness, which the human studies would establish

Research materials and status

The internal results are summarized with their limitations so that the scope of existing evidence is explicit. Several artifacts are public and verifiable now. The vocabulary is published as a JSON Schema, a JSON-LD context, an OWL ontology, SHACL shapes and a concept vocabulary at `https://leadershipunderuncertainty.org/sco/v1/`, and the reference validator checks a submitted artifact at `https://leadershipunderuncertainty.org/api/v1/sco/validate`. The sufficiency test runs at `https://leadershipunderuncertainty.org/api/v1/public/sufficiency-test`, and the agent handoff format is published as an extension descriptor at `https://leadershipunderuncertainty.org/sco/a2a/extensions/handoff-envelope/`

v1. The sealed record's daily roots, a sample seal and a script that walks the chain are in the public anchors repository at https://github.com/cberzins1973/luu-attestation-anchors, and a single forecast can be checked against its seal at https://leadershipunderuncertainty.org/credibility. The twenty-five-task starter evaluation, the sixty-instance conformance corpus, the handoff test dataset and the governance walkthrough are held in the programme's research collection and are available on request until their public release. The judgment record can be read and written by an external agent through published interfaces, and a decision logged with its evidence, alternatives and authority can be presented to an auditor as a dated, hashed attestation. Detailed specifications, reference implementations and study records for the remaining components are held in the programme's research collection, and a complete replication package for the studies of Appendix A is not supplied with this draft. Independent verification will require release of the relevant protocols, input and output records, scoring procedures, software versions and corrections, subject to applicable data rights.

References

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.

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

Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management. SR 11-7 / OCC Bulletin 2011-12.

Butler, R. (2004). *Review of intelligence on weapons of mass destruction: Report of a Committee of Privy Counsellors*. HC 898. The Stationery Office.

Canadian Centre for Cyber Security. (n.d.). *National Cyber Threat Assessment*, annex on estimative language. Communications Security Establishment Canada.

Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction (WMD Commission). (2005). *Report to the President of the United States*, 31 March 2005.

Coulthart, S. (2017). An evidence-based evaluation of 12 core structured analytic techniques. *International Journal of Intelligence and CounterIntelligence*, 30(2), 368–391.

Director of Central Intelligence. (1984). Special National Intelligence Estimate 11-10-84/JX: Implications of recent Soviet military-political activities, 18 May 1984. In *Foreign Relations of the United States, 1981–1988*, vol. IV, document 221.

Doyle, J. (1979). A truth maintenance system90008-0). *Artificial Intelligence*, 12(3), 231–272.

Friedman, S., Rye, J., LaVergne, D., Thomsen, D., Allen, M., and Tunis, K. (2020). Provenance-based interpretation of multi-agent information analysis. *Proceedings of TaPP 2020*. arXiv:2011.04016.

Gottweis, J., Weng, W.-H., Daryin, A., et al. (2026). Accelerating scientific discovery with Co-Scientist. *Nature*, 655, 487–496.

Greenblatt, R., Shlegeris, B., Sachan, K., and Roger, F. (2023). AI control: Improving safety despite intentional subversion. arXiv:2312.06942.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., and Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19–30.

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence, Central Intelligence Agency.

Holland, M. (2005). The "photo gap" that delayed discovery of missiles in Cuba. *Studies in Intelligence*, 49(4).

Kent, S. (1964). A crucial estimate relived. *Studies in Intelligence*, 8(2), 1–18. Originally classified. Reprinted in D. P. Steury (ed.), *Sherman Kent and the Board of National Estimates: Collected essays* (Center for the Study of Intelligence, 1994), 173–187.

Kirgis, P., Kapoor, S., Schwartz, A., et al. (2026). Can AI agents conduct open-ended AI research? Early evidence from two case studies. arXiv:2607.27191, version 2.

Korbak, T., et al. (2025). Chain of thought monitorability: A new and fragile opportunity for AI safety. arXiv:2507.11473.

Mandel, D. R., and Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences*, 111(30), 10984–10989.

Marcoci, A., Burgman, M., Kruger, A., Silver, E., McBride, M., Thorn, F. S., Fraser, H., Wintle, B. C., Fidler, F., and Vercammen, A. (2019). Better together: Reliable application of the post-9/11 and post-Iraq US intelligence tradecraft standards requires collective analysis. *Frontiers in Psychology*, 9, 2634.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.

Mellers, B., Ungar, L., Baron, J., et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5), 1106–1115.

Office of the Director of National Intelligence (ODNI). (2015, as amended). Intelligence Community Directive 203: Analytic standards. Signed 2 January 2015; public consolidated text accessed 19 September 2026.

President's Foreign Intelligence Advisory Board (PFIAB). (1990). The Soviet "war scare", 15 February 1990. Declassified 2015; National Security Archive.

Rao, A. S., and Georgeff, M. P. (1995). BDI agents: From theory to practice. *Proceedings of the First International Conference on Multiagent Systems*, 312–319.

Russell, S., and Wefald, E. (1991). Principles of metareasoning90015-C). *Artificial Intelligence*, 49(1–3), 361–395.

Senate Select Committee on Intelligence. (2004). *Report on the U.S. Intelligence Community's prewar intelligence assessments on Iraq*. S. Rep. 108-301. United States Senate.

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., and Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. arXiv:2303.11366, version 4.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118.

United States Intelligence Board (USIB). (1962). Special National Intelligence Estimate 85-3-62: The military buildup in Cuba, 19 September 1962. In *Foreign Relations of the United States, 1961–1963*, vol. X, document 433.

Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8, 2293–2303.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. arXiv:2305.16291, version 2.

Weick, K. E., Sutcliffe, K. M., and Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, 16(4), 409–421.

Wilson, B., and Bash, J. (2026). Spring 2026 FutureEval analysis: Pros beat bots, but the gap is nearly gone. Metaculus, 2 September. Author cross-post, 9 September.

Wright, J. R. (1962). Analysis of SAM sites in Cuba, briefing memorandum, 1 October 1962. Avalon Project, Yale Law School.