

The Judgment Layer: From Forecasting to Strategic Understanding in AI Systems

Christopher Berzins

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

October 2026

Abstract

Leading AI systems now forecast events about as accurately as superforecasters, the best human forecasters. This is an impressive and important achievement. Forecasting is nonetheless a narrow strategic skill. A forecast is only as good as the question asked, and it is a small part of how individuals and organizations build a strategic understanding of their environment and, ultimately, take decisions based on that understanding. Leadership Under Uncertainty, an applied research project, aims to build AI systems demonstrably capable of strategic understanding and, ultimately, of judgment, for higher-value strategy work and AI safety, since judgment is key to responsible decision-making. For millennia people have grappled with strategic thinking under radical uncertainty, perhaps most systematically in the intelligence community. There, strategic assessment and judgment form a dynamic process, of which any single determination is one moment. How information is gathered and evaluated, how individual assessments are rolled up into broader judgments and, above all, how those judgments are systematically updated on new information and assessments matter far more than the accuracy of any single forecast. The project asks whether AI systems can replicate this process, whether it enhances strategic understanding, and whether a judgment layer measurably improves AI agents' decisions. Components of every stage are implemented, most partly operating, and the whole process is unproven. Measured on a small scale, 13 of 40 audited citations by the system supported their claims, and structured generation made insights better grounded and more testable without gain in decision relevance. A study of decisions holding forecasting ability constant is specified for collaborators to run.

Keywords: forecasting, strategic understanding, strategic assessment, judgment, belief updating, intelligence analysis, analytic standards, agentic AI, AI safety, decision-making under uncertainty

1. From forecasting to strategic understanding and judgment

The best AI systems now forecast world events about as accurately as superforecasters, the top human forecasters identified in forecasting tournaments (Tetlock and Gardner, 2015). This is an impressive and important achievement. On ForecastBench, the Forecasting Research Institute's continuing benchmark, several systems are statistically indistinguishable from the median forecast of a superforecaster panel (Forecasting Research Institute, 2026; ForecastBench, 2026). Metaculus, a forecasting platform, runs a controlled comparison of its professional forecasters with the best automated systems. The professionals' lead, on Metaculus's scoring scale, fell from about twenty points a question in the second quarter of 2025 to 1.25 in spring 2026, too small to distinguish from no lead (Wilson and Bash, 2026). Appendix A.1 sets out the limits of these comparisons.

Forecasting is nonetheless a narrow strategic skill. A forecast is only as good as the question asked, and it is a small part of how individuals and organizations build a strategic understanding of their environment and, ultimately, take decisions based on that understanding. Every comparison above scores a probability on a short-horizon question that someone else posed in advance, and at most it rewards revising that one probability. None measures whether a system chose the question well, judged what a report establishes, related separate assessments to one another or explained why a judgment changed. As prediction becomes cheap, the value of what must accompany it rises, including judgment (Agrawal, Gans and Goldfarb, 2018).

Leadership Under Uncertainty is an applied research project that aims to build AI systems demonstrably capable of strategic understanding and, ultimately, of judgment. It has potential application in two areas. The first is higher-value strategy work, such as framing an organization’s questions, weighing what its sources establish, connecting assessments held by different teams and explaining why its view has changed. The second is AI safety, because judgment is a key attribute of responsible decision-making. Agentic AI is AI deployed to pursue tasks through repeated use of information and tools, acting as well as advising. An agent given consequential work must recognize the limits of its evidence, keep an assessment apart from the authority to act on it, and refer to a person when its evidence or its authority runs out. Case studies of AI agents conducting open-ended research found agents that completed the engineering unaided and showed poor judgment about what counted as publishable research (Kirgis et al., 2026).

For millennia people have grappled with how to think strategically under radical uncertainty, uncertainty that cannot be reduced to known probabilities, and possibly the most systematic operationalization of that effort exists within the intelligence community. There, strategic assessment and judgment are understood as a dynamic process, of which any single determination is one moment. How information is gathered and evaluated, how individual assessments are rolled up into broader judgments and, above all, how those judgments are systematically updated on new information and assessments matter far more than the accuracy of any single forecast. In the community’s doctrine, accuracy is one of nine tradecraft standards, the rules that govern how a finished assessment is written and evaluated. Section 2 traces each stage through that doctrine and through one case, the Cuban missile crisis of 1962.

Strategic understanding, as I use the term, is a representation of the actors, incentives and constraints in a situation that identifies the question that matters for a decision, names the observation that would discriminate between rival accounts and anticipates the next change. Judgment, as I use the term, is the formation, integration, use and revision of assessments about what matters, what is credible and what warrants inquiry or action under uncertainty. A judgment layer is the set of functions through which an AI system carries out this dynamic process. The project explores three questions.

1. Can such a dynamic process be replicated by AI systems?
2. Can it enhance strategic understanding?
3. Can integrating a judgment layer into agentic AI measurably improve decision-making?

The word "demonstrably" in the project’s aim sets the standard of evidence, so every claim about the project in this paper is labelled by level. A claim is specified when a manuscript or protocol states it, implemented when reference software does it and tests confirm the software’s properties, operating when it runs on live data, and measured when a study has estimated a quantity or effect, within its limits. Table 1 sets out what an affirmative answer to each question would require and

how the project will test it. Section 3 describes the reference system, Table 5 in Section 4 gives the evidence so far, and Section 8 invites collaborators.

Table 1. The three research questions and how they are tested

Question	What an affirmative answer requires	Measures that need no resolved outcome	Test	What would count against it
1. Can the dynamic process be replicated by AI systems?	Observed behaviour, checked against the record. Sources are evaluated and shared origins detected, and component assessments are integrated into a standing judgment with dissent preserved. Revisions follow discriminating evidence and state what changed and why, as trade-craft standard 7 of Intelligence Community Directive 203 requires (Section 2.5)	Share of revisions linked to an identified new item of evidence. Reversals without new evidence. Blind ratings of explained change. Detection of planted reports that share one origin and of sources impersonating known outlets. Traceability of each key judgment to its components	Section 5.1, Studies 1 to 3. Section 5.2, process outcomes	Removing explained change leaves the results unchanged. Planted reports sharing one origin pass as independent. Revisions remain unlinked to evidence
2. Can the process enhance strategic understanding?	The actors, incentives and constraints represented. The key question identified. The observation that would discriminate between rival accounts named. The next change anticipated	Blind ratings of whether, and at which checkpoint, the question, the discriminating observation and the actors listed in advance by a panel were identified. Next-change prediction against the rule that nothing changes	Section 5.1, Study 4. Section 5.2, understanding outcomes	Understanding ratings no better than the forecasting agent's. Next-change prediction stays below the rule that nothing changes. Given a disguised version of the September 1962 evidence (Study 4), the system never asks the question that proved decisive

Question	What an affirmative answer requires	Measures that need no resolved outcome	Test	What would count against it
3. Can a judgment layer in agentic AI measurably improve decision-making?	Better decisions by an agent acting under delegated authority, with forecasting ability held constant	Decision quality rated blind on the information available at each check-point. Unsupported or unauthorized actions. Appropriate referral in both directions. Timeliness	Section 5.2, the decision study, specified and not yet run	The judgment-layer agent does no better than a forecasting agent on decision quality or loss. An agent given its raw history does as well. The margin shrinks with a stronger underlying model. Gains come only from referring or spending more

2. Strategic assessment as a dynamic process

Figure 1 sets out the process as this paper uses it. The intelligence community’s stages of gathering and evaluating, rolling up and updating are preceded by setting the question and followed by deciding under authority.

Strategic assessment as a dynamic process

A single forecast is one output of the process. The judgment layer is meant to carry the whole of it.

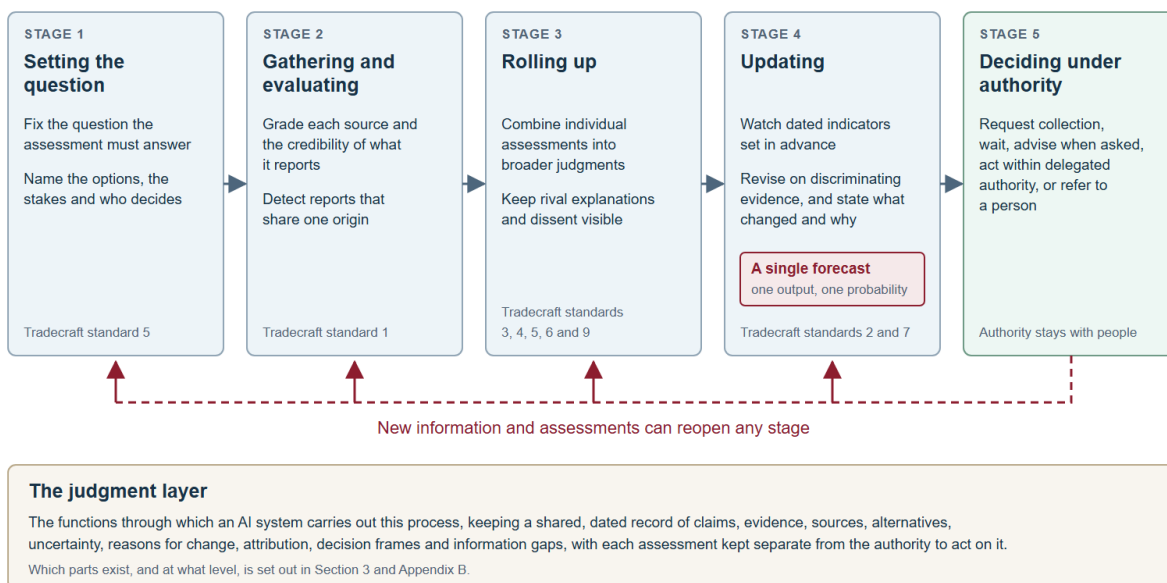


Figure 1. Strategic assessment as a dynamic process. A single forecast is one output. The complete arrangement is unvalidated.

2.1 A long problem

The problem is old. Thucydides tested conflicting reports "by the most severe and detailed tests possible", because eyewitnesses disagreed through faulty memory or partiality (Thucydides, 1874, 1.22). Clausewitz wrote that a great part of the information obtained in war is contradictory and a still greater part false (Clausewitz, 1832/1873, I.6). Knight (1921) separated measurable risk from unmeasurable uncertainty, and observed that a person estimating a unique situation also estimates the probability that the estimate is correct, anticipating the analyst's separation of likelihood from confidence. Kay and King (2020) call this condition radical uncertainty and propose beginning with the question "What is going on here?".

2.2 Setting the question

The intelligence community has its own answer to the objection that a forecast is only as good as the question asked. A National Intelligence Estimate begins with terms of reference, circulated across the agencies to fix the questions it will answer (NIC, 2007). Intelligence Community Directive 203 (ICD-203), the United States directive on analytic standards, requires analysis to demonstrate relevance to its customers and to address the implications of its judgments (ODNI, 2015, as amended). The Cuban missile crisis of 1962 shows why the question matters. On 19 September 1962 the United States Intelligence Board judged a deployment of Soviet ballistic missiles in Cuba "incompatible with Soviet practice to date and with Soviet policy as we presently estimate it" (USIB, 1962). Within four weeks the estimate was overtaken by a question it had answered only in general terms, what the new air-defence sites in western Cuba were protecting. A well-calibrated forecaster is one whose stated probabilities match how often events occur. I expect that such a forecaster would have given the same answer on the September evidence, a hypothesis that Study 4 in Section 5.1 is designed to test. John McCone, the Director of Central Intelligence, had warned in August, before any air-defence site had been photographed, that Moscow might place offensive missiles in Cuba. He reportedly read the sites photographed on 29 August as meant "to blind our reconnaissance eye", and he was abroad when the Intelligence Board concurred in the estimate (Holland, 2005).

2.3 Gathering and evaluating

The Butler review describes a British chain of three steps: validation, which removes unreliable or planted reporting, then analysis, then assessment (Butler, 2004, paras 20 and 48). ICD-203's first tradecraft standard requires each finished assessment to describe the quality and credibility of its sources, including possible denial and deception, and encourages statements saying which sources carry the key judgments. NATO doctrine grades each report on two six-point scales, the reliability of the source and the credibility of the information, a scheme commonly attributed to the Admiralty. In experiments, participants' trust in a report followed the source grade more than the credibility grade (Kelly et al., 2025). Collection is itself directed by hypotheses, and Grabo (2002, p. 3) distinguishes indicators, steps an adversary should or may take in preparation for hostilities, from indications, information that such a step is being taken.

The Cuban case shows evaluation at work. A Defense Intelligence Agency briefing paper of 1 October noted that three air-defence sites in western Cuba could not be connected with any significant military installation (Wright, 1962). A Cuban observer agent, the lowest rank of source, reported secret work in a restricted area elsewhere in the region, two refugee reports coincided with his, and the area went to the top of the list of reconnaissance targets. New reconnaissance settled the question. On 10 September the President's advisers had agreed to keep high-altitude U-2 reconnais-

sance flights away from the known air-defence sites, and on 11 September the President approved that schedule. On 9 October he approved one flight over western Cuba, and on 14 October it photographed the medium-range launchers (Holland, 2005).

Whether a report is evidence depends on the claim it bears on and on where the report came from. Several reports that repeat one announcement therefore add little independent support, a failure analysts call circular reporting. The Iraq post-mortems found that several apparently corroborating reports traced to a single source (SSCI, 2004; Butler, 2004; WMD Commission, 2005).

2.4 Rolling up

Individual assessments become broader judgments through coordination. In the National Intelligence Council's 2007 description, terms of reference are circulated, analysts draft and the Council critiques (NIC, 2007, reproduced in Treverton, 2013). Agency representatives "hone and coordinate line-by-line the full text", each key judgment receives a confidence level, and the National Intelligence Board approves the result. The estimate was legally the Director's, and agencies that disagreed recorded dissent in footnotes, a right written into the founding directive of 1946 (Kent, 1976). British practice is similar (Table 2). Roll-up in this sense is deliberative coordination that keeps disagreement visible. Averaging the forecasts of several models, a common practice in machine forecasting, discards that disagreement. Grabo (2002, pp. 8–9) warned that bringing more people into coordination can dilute a judgment, and that in past crises the most nearly correct view was often held by a minority spread across agencies.

2.5 Updating

ICD-203's seventh tradecraft standard asks finished assessments to state whether major judgments are consistent with or change earlier analysis, and to "make clear how new information or different reasoning led to the judgments expressed in them" (ODNI, 2015, as amended, §D.6.e(7)). The directive's objectivity standard adds that analysis should not be "unduly constrained by previous judgments", and tradecraft standards 2 to 4 ask for the indicators that would change an assessment. The British Common Analytical Standards ask analysts to explain change and to make judgments falsifiable, "containing a hypothesis, verbal probability, and timeframe" (PHIA, 2025). Revision machinery includes the Memorandum to Holders, which updates an estimate for new information without a full reassessment (Treverton, 2013), and signpost lists, lists of developments that would signal a change, reviewed against the currently agreed assessment (US Government, 2009).

The 2007 estimate on Iran's nuclear program shows systematic updating in published form. It judged that Tehran had halted its nuclear weapons program in the autumn of 2003, reversing the 2005 judgment that Iran was determined to develop nuclear weapons. It printed a table setting its key judgments beside those of 2005 so that readers could see what had changed (NIC, 2007; Treverton, 2013). For warning, Grabo (2002, p. 164) put the principle as "cumulative, not merely current".

The Board of National Estimates, which Sherman Kent chaired, reviewed new information every working day against its standing estimates. It reopened one only if the subject still mattered and, in Kent's words, "the new evidence must be firm and must indicate a significant departure from what was previously estimated". It would not normally recall an estimate to raise "probably" to "almost certainly", and let current intelligence carry the increments. Disciplined updating therefore includes declining to reopen an estimate. The September 1962 estimate was never formally reopened, and

Kent later doubted that a revision, though "a better record on paper", would have moved the government to "crash action" (Kent, 1964).

2.6 Doctrine written after failures

Much of this doctrine was written after failures to update. Heuer (1999) showed how information arriving in increments is absorbed into existing views and how a written assessment gives its authors a stake in keeping it. The Senate intelligence committee found "layering" in the 2002 Iraq estimate, judgments built on earlier judgments without carrying their uncertainties forward (SSCI, 2004, conclusion 4). The Iraq Inquiry found no formal reassessment of the judgments of the Joint Intelligence Committee, the British body that agrees national assessments, after September 2002 (Chilcot, 2016, paras 566–570). ICD-203, issued in 2007 and revised in 2015, sets five analytic standards, the fifth containing the nine tradecraft standards (ODNI, 2015, as amended). Tables 2 and 3 set out the process by stage. Accuracy, standard 8, asks for the most accurate judgments possible on the information available, warns against avoiding difficult judgments to reduce the risk of being wrong, and asks for judgments clear enough to be scored. On this paper's reading, the doctrine places most of its weight on how judgments are sourced, coordinated, expressed and revised. Canadian strategic forecasts were well calibrated and underconfident (Mandel and Barnes, 2014), and frequent, incremental updating predicted forecasting accuracy (Mellers et al., 2015; Atanasov et al., 2020). Coulthart (2017) found limited evidence for widely used structured analytic techniques, and individual ratings of intelligence reports against a rubric derived from ICD-203 were unreliable until they were aggregated (Marcoci et al., 2019). No study has tested the doctrine as a whole.

Table 2. The process in doctrine, by stage

Stage	United States mechanism	United Kingdom mechanism	Governing standard	Documented failure
Setting the question	Terms of reference circulated across agencies for each estimate (NIC, 2007)	A forward program of assessments, issued three times a year in 2004 and revised as events required (Butler, 2004, para 43)	ICD-203 tradecraft standard 5, relevance to customers	In 1962 the challenge by the Director of Central Intelligence to the estimate's presumption came from outside the line the Intelligence Board agreed (Holland, 2005)
Gathering and evaluating	Source descriptions and source summary statements. Collection agencies assess the sources behind critical judgments (ODNI, 2015; NIC, 2007)	Validation, analysis and assessment (Butler, 2004, paras 20 and 48). NATO two-scale grading of source and information (Kelly et al., 2025)	Tradecraft standard 1	Apparently corroborating reports traced to a single source (SSCI, 2004; Butler, 2004; WMD Commission, 2005)

Stage	United States mechanism	United Kingdom mechanism	Governing standard	Documented failure
Rolling up	Line-by-line interagency coordination, a confidence level for each key judgment, Board approval, dissent in footnotes (NIC, 2007; Kent, 1976)	Assessments Staff drafts, Current Intelligence Groups debate, the Joint Intelligence Committee agrees (Butler, 2004, para 43). Today the Committee endorses assessments usually written by the Joint Intelligence Organisation (Cabinet Office, n.d.)	Tradecraft standards 3, 4, 5, 6 and 9	Consensus diluting a correct minority view (Grabo, 2002)
Updating	Memorandum to Holders (Treverton, 2013). Daily review of standing estimates with a bar for reopening (Kent, 1964). Signpost lists (US Government, 2009)	Explain change. Falsifiable judgments with a hypothesis, probability and timeframe (PHIA, 2025)	Tradecraft standards 2 and 7	Layering in the 2002 Iraq estimate (SSCI, 2004). No formal reassessment after September 2002 (Chilcot, 2016)

Table 3. ICD-203 tradecraft standards as requirements on an AI system

ICD-203 tradecraft standard	Stage	What the system would have to show
1. Describe the quality and credibility of sources	Gathering and evaluating	Which claims rest on which sources, how reliable each is, and when an apparently corroborating report shares an origin with one already held
2. Express and explain uncertainties	Updating	A likelihood for the assessed event and, separately, the strength of the evidence behind it, with the indicators that would move either
3. Distinguish information from assumptions and judgments	Rolling up	Which statements are reported, which are assumed and which are the system’s own inferences, in each component and in the combined judgment
4. Incorporate analysis of alternatives	Rolling up	Competing explanations carried through the combination of assessments, with dissent preserved and the evidence that would discriminate among them
5. Demonstrate customer relevance and address implications	Setting the question, and rolling up	The customer’s question the assessment bears on and the implications of its judgments, stated without recommending a course of action
6. Use clear and logical argumentation	Rolling up	An inference chain a reader can follow from evidence through component assessments to the conclusion
7. Explain change to, or consistency of, judgments	Updating	What changed since the last assessment and why, tied to the item of evidence or reasoning that prompted it, or why nothing changed despite new reporting
8. Make accurate judgments	All stages	Dated, resolvable judgments, including difficult ones, that can be scored against outcomes

ICD-203 tradecraft standard	Stage	What the system would have to show
9. Use effective visual information	Rolling up	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 stage assignment and the right-hand column are this paper’s proposed application to AI systems.

2.7 The reciprocal opportunity

Intelligence doctrine can give AI research an account of assessment as a process, and AI systems can make parts of that doctrine testable, because a system’s procedures can be switched on and off in a way an analyst’s cannot (Section 5). Intelligence analysis supplies assessments to decision-makers, so agents that take actions need additional rules about permission and consequences.

3. A judgment layer for AI systems

The functions of the judgment layer can be distributed across models, retrieval of relevant documents, memory, tools and human review, and each component below is stated at its level, stage by stage as in Figure 1. The project’s reference system, the software that implements these functions and runs on live data, is called the system below. Table 4 gives, by stage, the vocabulary in which a judgment is recorded. For the first stage, setting the question, the decision frame of Table 4 is implemented and cited by none of 693 recent insights, and Study 4 (Section 5.1) would test question-finding.

Table 4. A common vocabulary for judgments under uncertainty, by stage of the process

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

3.1 Gathering and evaluating information

A source records where evidence came from, and an attribution records who judged it. A reliable report of Moscow’s assurance in September 1962 establishes that the statement was made and says

little about Soviet intent.

Intake is operating. In the thirty days to 27 September 2026 it averaged about 5,000 dated observations a day, individual readings from public data series such as electricity demand, from 18 of 23 registered sources. Most observations repeated unchanged values, and four sources reported success while supplying no real readings. Three have since been repaired (Appendix A.8). About 138,000 candidate news stories a month are clustered and screened by a significance gate, a set of rules that uses no AI model and whose weights, by the project's own record, were set before it saw live data. About 3 per cent of the clusters pass. The gate uses the number of outlets carrying a story only as a measure of prominence, so repetition does not count as confirmation. In operation, a prediction market's settlement cannot confirm a model's verdict on a forecast question if the evidence for that verdict cited the market. A critic, a second AI model, operates on most insights, checking whether cited material supports their claims, and an operating mechanical check tests whether each citation leads to a passage in its source, less strictly than the check of Section 4.3.

Source evaluation is the weakest part. Source grading operates as a single grade, for the website that published each item, from a fixed list of known websites. In the terms of the two-scale grading of Section 2.3 this grades the channel alone. Neither the originating source behind a report nor the credibility of the information itself is graded, and the list placed 88 per cent of recent evidence in an "unknown" tier. In thirty days, two websites imitating a major international news outlet supplied four times as much of the evidence retrieved to settle forecast questions as the outlet itself, although no verdict yet rests on them. In operation, the general corroboration rule at settlement counts near-identical reports on different websites as independent confirmation, and shared origin is detected only for a hand-built list of thirteen websites. This is the circular reporting that the Iraq post-mortems identified (Section 2.3).

3.2 Rolling individual assessments up into broader judgments

An insight is a finding of one of a fixed set of kinds, such as a binding constraint that has shifted or an assumption that no longer holds. The design specifies a demanding entry test, applied by the critic in operation, under which an insight earns its place only if producing it required something a thoughtful analyst could not readily do alone, such as holding hundreds of assumptions in view over months. McCone's reading of the air defences in 1962, as meant to blind reconnaissance, had the form sought, a rival explanation that one overflight, the kind of evidence Heuer (1999) calls diagnostic, could test.

Generation, criticism by a second model, a check of citations and ranking are implemented as successive steps, each of which an insight must pass. They resemble review of a draft within one agency more than interagency coordination, because the critic reads the same material as the generator and cannot record a dissent that travels with the judgment. In operation the synthesis step described below publishes theses without passing the critic, and until 26 September 2026 insights the critic rejected could still be published, as 21 were (Appendix A.8).

Roll-up is operating. Since July the system has stated 417 distinct beliefs, each an assessment counted once across its restatements. A synthesis step has composed 184 theses, each from about four findings, stating which of the system's standing assumptions they support or contradict. Forecasts in 133 standing question sets, continuing sets of forecasting questions each kept open on one subject, record which named assessments and causal factors informed them. A thesis cites its findings only by headline, so its success or failure cannot be traced back to them, and recent theses carry no alternatives or information gaps. When two critics disagree, the reliance controller of

Section 3.4 discounts the lone adverse verdict, which is nearer to the dilution of a minority warning that Grabo (2002) described than to the dissenting footnote.

Actor models, designed to separate declared aims, observed actions and inferred motives, operate for fifteen states and twenty AI-sector firms. Their comparison of stated with revealed positions is disabled, and in a paired test of forecasts made with and without them, 95 of 114 active pairs are identical, so even full resolution is unlikely to show an effect (Appendix A.8).

3.3 Updating judgments

In operation, assessments that pass these steps are held as standing positions. Each carries a conviction, the probability that its claimed mechanism holds, which is scored like a forecast. Each also carries dated indicators, stated in advance with the direction each would move it. Of 2,541 indicators, 76 were settled, that is, found to have occurred, by an independent instrument, either a query against an external data series or a model reading retrieved text. Another 174 were settled against the system's own records and are kept out of its track record.

An indicator is useful only if it is more likely to fire when an assessment is right than when it is wrong, a problem warning analysts also face. Each has a criterion, the exact condition under which it fires. A criterion labelled a three-standard-deviation event, which should occur by chance on about 0.135 per cent of readings, is tested on every reading in its window. On one electricity-demand series that meant 532 readings in thirty days and, if the readings were independent, about a 51 per cent chance of firing when nothing had happened. A criterion gate, operating and enforced by default, with thresholds fixed at its first live run, is meant to refuse such criteria, and by the end of August 192 criteria across 127 assessments had been withdrawn as sealed, dated retractions. Until 26 September 2026, however, the gate estimated each criterion's chance of firing by accident from recent readings, which cannot show a series drifting, and it admitted any data series that no one had reviewed. On 25 September 33 beliefs had been settled, and 30 of them, the structured beliefs below, rested on a query against an external data series. The gate had put the chance that each deciding criterion would fire by accident at 1.2 to 4.0 per cent. Replayed from every past date on which it could have been set, each fired on 52 to 100 per cent of them. The gate as revised on 26 September refuses unreviewed series, criteria unrelated to their claims and any criterion whose simulated chance of firing when nothing has happened exceeds 5 per cent. Replayed on the 65 criteria behind those beliefs, it admits none (Appendix A.8).

Four controls slow revision. A signal must persist before a threshold moves, revisions must be a minimum interval apart, a check reverses any change that degrades performance, and a daily budget caps revisions. Kent's reopening bar weighed the subject's importance and the new evidence (Section 2.5), and only the first control bears on evidence. The controls operated on storylines, the system's continuing narratives of a situation, during a trial with pilot users, which was suspended in June 2026, applying 975 revisions and declining 1,386, all of them on the daily budget or the minimum interval. For assessments they are bypassed (Section 3.4).

Sealing is operating. When a record is sealed, a digital fingerprint of its contents is combined with the day's other fingerprints into a single value published in a public repository. A third party can then check that the record existed, unaltered, no later than the time that value was published. About 42,600 records are sealed, including about 6,800 forecast revisions, each linked to its predecessor (Appendix A.8).

Against standard 7 this stage falls short. The reason-for-change element is empty in operation, no

change of conviction on a held belief has been recorded, and revision appears as changes of status justified by the indicator that fired. Revision of the forecasting method itself has one working check. The promotion gate, the operating test that a new forecasting method must pass before it replaces the one in use, reversed its first apparent success, because the new method had been given part of the market price it was scored against (Appendix A.3).

3.4 From assessment to decision under authority

Assistance should make options, objectives and evidence visible to those responsible for deciding, preserve disagreement and leave authority with them. The specified advice rule admits a recommendation only when a decision-maker asks for one, within that person's question, authority and appetite for risk. It is not enforced in operation, where one of the system's search routines issues recommendations that no one requested. A mechanism that holds the system's one irreversible operation until a person commits to it is implemented and unused. A reliance controller, the rule that decides whether an evaluator's verdict is accepted, referred to a second evaluator or deferred to a person, is operating and has recorded 902 of about 3,700 decisions as deferrals. Authority checks on outside AI agents are implemented and at first only record refusals.

In operation, every one of the 15,540 automated action requests recorded since April was refused, including 965 in the last sixty days that sought to change an assessment's status without a person's confirmation. After each refusal the system recorded and sealed the change itself, by a route outside the action controls, so the controls requiring a recorded reason and a daily budget never ran. An action that declares no required authority also runs without confirmation. In the intelligence community the assessing body revises an estimate on its own analytic authority, and a decision-maker's authority governs action. The design required a person's confirmation for revision as though revision were action.

3.5 Learning across inquiries

A further cycle, running across inquiries, extends updating to the process itself. Kent's diagnosis of 1962 was that estimators lose an adversary "when the other man zigs violently out of the track of 'normal' behavior" (Kent, 1964).

The design specifies recording lessons with provenance, adjudicating them against outcomes and retrieving them before a later inquiry forms its question, so that a ministry reading an unexplained air-defence deployment around a site in a third country asks what it protects. Credit assignment, apportioning a resolved judgment's error to what produced it, operates in observation mode, recording assignments without acting on them. Only 13 of 156 sealed assignments rest on independently settled outcomes, and all 1,402 changes to its own methods that the system has learned from past results are recorded and scored but not applied. Appendix C sets out the intellectual sources and nearest precedents of this cycle.

4. Evaluating judgment, and what the evidence shows

4.1 Levels and instruments

What is switched off or paused is recorded in Appendix B. Methods are in Appendix A, and the operating audit of 27 September 2026, with its definitions, is Appendix A.8.

The project’s benchmark, an implemented set of tasks for comparing AI systems, scores reasoning under uncertainty, recognition of limits, generation of alternatives, revision of belief within and across sessions, and integration of information from different sources. Its overall score is capped for any system whose tradecraft, integration or continuity falls below a threshold, so forecasting skill alone cannot earn a high score. Revision across sessions, the dimension closest to Question 1, has never been scored for any system.

A ladder of four rungs grades what a record can show. The two lower rungs are structural coverage, whether the required elements are present, and relational integrity, whether a cited claim rests on the cited evidence. Both have been measured in one study (Appendix A.7). The two upper rungs are behavioural correspondence, whether the system actually reasoned as its record says, and causal sensitivity, whether changing a premise changes the conclusion. Neither yet has a scorer, and research on monitoring a model’s stated reasoning argues that such correspondence cannot be assumed (Korbak et al., 2025). Question 1 can be answered only at the relational and behavioural rungs.

Table 5. Results by research question and level

Result	Question	Level	Main limit	Appendix
Structured insight generation: grounding +28.3, testability +22.7, no gain in decision relevance or non-obviousness	2	Measured	One generator and one judge, fifteen contexts	A.2
Structured prompting with self-critique, without document search: Brier 0.1743 against 0.1776, interval spanning zero, 344 questions, both worse than the market price	Accuracy only	Measured	Structured arm rerun after resolution	A.3
Two validators agree on all 60 conformance verdicts	1	Measured	Conformance only, silent on truth or usefulness	A.4
Element removal: seven of eight removed elements load-bearing, the claim not removed	1	Measured	Measure adopted after registration, single judge, circular	A.5
Six models, unprompted, fall short at times on source, alternative and gap, within-session updating 55.7 to 72.6	1	Measured	Single judge, runs with no memory between sessions	A.6
Model judges agree at kappa 0.13 to 0.30 and prefer length, candidate models as judges add about sixteen points	1	Measured	Pilot audit	A.6
Questions passing the citation check 10 against 7, 1, 4 and 0, citations 27 of 40 resolved and 13 of those 27 supporting, none of ten sampled on the largest document	1 and 2	Measured	Four documents, single audit judge, outputs not yet archived	A.7
Policy papers: 82.9 against 55.7 for a single request to a model	2	Measured	One paper per condition, single judge that also wrote both papers	A.9

Result	Question	Level	Main limit	Appendix
Next-change prediction: indicator Brier 0.341 against 0.020 for the rule that nothing changes	2	Measured	The system’s own record, weekly windows	A.8
Oscillation: 1,275 reversal records on 187 assessments	1	Operating	One defect, since fixed	A.8
Evaluator registry fed self-graded outcomes, critic 8 of 14 on independent outcomes	1	Operating. Critic’s record on independent outcomes measured	Registry uncorrected on 27 September 2026	A.8
37 beliefs independently settled, 32 held. Six scored on the claim’s own criterion, hits and misses together	2	Operating (counts). Measured (baseline comparison)	Hit rate equal to a baseline that never reads the thesis (30 structured beliefs, 25 September)	A.8
Decisions	3	Specified, not run	Decision records hold one walkthrough	None

Counts are from the operating audit of 27 September 2026 (Appendix A.8) unless another date is given. Brier scores measure the error of probability forecasts, lower being better, with 0.25 the score for always answering one half. Kappa measures agreement between judges corrected for chance, 0 being chance.

4.2 Forecasting accuracy inside the project

The project’s own attempts to add accuracy bear on the objection considered in Section 6 and leave untouched the argument of Section 1, which concerns what a forecast leaves out. A structured method, in which one model estimated a base rate, weighed the evidence and critiqued its own forecast, scored a Brier of 0.1743 against 0.1776 for a plain model on 344 resolved questions, a difference whose interval spans zero. Both were significantly worse than the market price (Appendix A.3). Other comparisons, of document search, model size, a panel of models deliberating together and learning from worked examples, found small gains or none, and the method in use scores worse than copying the market price on the promotion gate’s questions (Appendix A.3). Other engineered systems do gain accuracy, as the benchmark leaders of Section 1 show. The project claims no gain in accuracy from its own methods.

4.3 Question 1: replicating the process

Six widely used models from three developers, given entry-level tasks with no instruction to report sources, alternatives or gaps, fell short at times on all three, which correspond to tradecraft standards 1, 4 and 2 (Appendix A.6). Their element-competence scores, the mean of each model’s scores on the nine elements of Table 4, ran from 65.3 down to 49.8, the last below the benchmark’s minimum-competence threshold. On harder geopolitical tasks that model failed seven of nine elements while producing fluent strategic prose. Their scores for updating belief as new evidence arrived within a session ran from 55.7 to 72.6.

The vocabulary of Table 4 has one measured result. In a pre-registered study that removed one element at a time from sixty forecasting records, seven of the eight elements removed were load-

bearing, meaning that removing one measurably reduced the capabilities scored, on a measure adopted after the registered measures showed nothing (Appendix A.5). The claim was treated as definitional and not removed, and attribution was the one element whose removal made no measurable difference. The test is also circular, because the capabilities scored are close to the elements removed.

The most consequential measurement concerns gathering. On four public board and committee documents, an AI model from a different developer checked forty of the system's citations. Under a strict check 27 of the 40 resolved to real passages, and 13 of those 27, or 48 per cent, supported the claim attached to them. On the largest document all ten sampled citations resolved and none supported the claim attached to it (Appendix A.7). On that document the two lower rungs of the ladder in Section 4.1, structural coverage and relational integrity, gave opposite verdicts.

The operating record adds a failure of updating. A defect in the rule that reopens resolved assessments produced 1,275 reversal records on 187 assessments between 23 July and 3 September 2026, up to 22 on a single assessment. The defect is fixed. It was visible only because entries in the revision log can be added but never altered or removed.

Two model judges agreed only slightly beyond chance, with Cohen's kappa between 0.13 and 0.30, and both preferred longer outputs. When models under evaluation also served as judges, they scored the weakest model about sixteen points higher than an independent judge did, so models under evaluation no longer serve as judges. In operation, the registry that learns how far to trust each evaluator was fed outcomes the system had graded itself, 177 of the 191 bearing on the critic. From 31 August the critic's low recorded reliability sent its adverse verdicts to a second critic, the 186 referrals changed no verdict, and on independently settled outcomes alone the critic scored 8 of 14. The registry repeated what the Senate committee called layering, building on earlier judgments without carrying their uncertain basis forward (SSCI, 2004), and on 27 September 2026 it had not been recomputed.

Read together, these results show components of each stage implemented and partly operating. The process itself has not been shown in behaviour. Two of its failures so far, circular reporting and layering, are ones the intelligence community's own post-mortems documented. The third, the repeated reversal of judgments, is the churn that Kent's bar for reopening an estimate was set against, without which estimates "would have to be reopened for almost daily revisions" (Kent, 1964). The resemblance suggests that the doctrine names hazards that an AI system also meets. No study has yet tested whether applying the doctrine prevents them.

4.4 Question 2: strategic understanding

Question 2 asks whether the process yields strategic understanding in the sense of Section 1. Across fifteen evidence contexts, with the same model, inputs and limit on output length, a separate model evaluator scored structured generation higher by 28.3 points for evidential grounding and 22.7 for testability, and found no gain in decision relevance or non-obviousness (Appendix A.2). This comparison tested only the generation step. The full insight process of generation and criticism has not yet been compared with a single request to a model, and is to be abandoned if it does not beat one on at least three of four dimensions, a condition fixed in advance. Two other processes have been compared with one. On the four board documents, a test of extracting questions for a board to ask, the best evidence-first prompt I could write, passed through the system's filters, kept seven, one, four and zero questions whose citations passed the system's citation check, in order of document size. The system kept ten on each, because it checks claims before generating questions

(Appendix A.7). On the largest document none of the ten sampled citations behind the system’s questions supported its claim (Section 4.3). On policy papers written before a decision, an evaluator derived from ICD-203 scored the policy-paper process at 82.9 against 55.7 for a single request to a model, with one paper per condition and a single judge, on papers I generated (Appendix A.9).

A weekly test of understanding as a process asks whether the system can anticipate the next change in its own record, and it does not beat persistence, the rule that nothing changes. Its Brier score was 0.341 against 0.020 for persistence. Counted once per belief, 37 of the system’s beliefs have been settled by an independent instrument, and 32 held. Most of these verdicts came from a generic indicator attached to a thesis from a fixed catalogue, such as a share’s trading volume or a bond yield, and a single trading-volume reading on 27 August 2026 settled nine beliefs at once. A baseline that never reads the thesis predicts 25.7 hits on the 30 structured beliefs, against 26 observed. Counted only on the claim’s own criterion, with the seven-day delay the track record applies, six beliefs were scored on 26 September 2026, hits and misses together, and the record carries no evidence of skill.

On this evidence, structure ties insights more closely to evidence, makes them more testable and keeps more questions whose citations pass the citation check, a measure that on the largest document diverged from support. Of the four parts of strategic understanding defined in Section 1, only anticipation has been measured, and the other three await Study 4.

4.5 Question 3: decisions

Question 3 has no measured result yet. Apart from one walkthrough of the governance controls, no real decision has been recorded against the system’s assessments, and no recent insight cites a decision frame. People have cleared 16 items from the human review queue, and 512 await review. The separation of an assessment from the authority to act on it is the design’s central safety claim, and its operation (Section 3.4) showed how easily it can be implemented wrongly.

5. Testing the three questions

5.1 First studies

Four inexpensive studies come first. Each is specified, and each will be pre-registered and sealed before its first run. None needs a resolved outcome.

For Question 1, Study 1 runs standing assessments with and without tradecraft standard 7 over a period of live operation. It compares evidence-linked with unprompted revisions, counts reversals without new evidence, and asks analysts who do not know which version produced each explanation to rate it against ICD-203 and the British "explain change" standard (PHIA, 2025). Its precondition is that explained change operates, which Section 3.3 shows it does not yet.

The benchmark’s cross-session continuity dimension will be scored in Study 2 on a system that retains memory between sessions, which the six models of Appendix A.6 did not.

Study 3 places planted reports sharing one origin, and sources impersonating known outlets, into live evidence streams. Given the gaps reported in Section 3.1, the current system is predicted in advance to fail it.

Study 4, on question-finding, addresses Question 2 and the hypothesis stated in Section 2.2. It uses a small set of historical warning cases whose declassified post-mortems name the question

that proved decisive, with a disguised analogue of the September 1962 evidence among them. A check excludes runs in which the model recognizes the case. Each model receives the same evidence framed as a request for a probability and as the question of what is going on. Raters score whether, and at which checkpoint, each answer represents the actors, incentives and constraints, names the question that matters for the decision and names the observation that would discriminate between rival accounts.

A measure of process risks circularity, since it can reward a system simply for following the process it is meant to evaluate. Each measure here counts only where outside research links it to accuracy, as the forecasting tournaments linked frequent and incremental updating to it (Mellers et al., 2015; Atanasov et al., 2020), or where it is later checked against decisions or outcomes.

5.2 The decisive test: the judgment-layer decision study

Question 3 needs a study that scores decisions while holding forecasting ability constant. The setting is continuing geopolitical inquiries, such as a military build-up, a stalled negotiation, a basing or transit agreement and a sanctions regime. The indicative design has at least 24 inquiries in at least three subject areas, one of which is kept out of all development and used only for the final test. Each inquiry has six to twelve dated checkpoints that release new evidence. Every version of the agent under test is called an arm. Before any arm runs, practitioners write each inquiry’s decision frame: the decision-maker’s question, objectives, deadline, authority and appetite for risk, the options, and the loss each option would incur in each possible state of the world. At each checkpoint the agent states its key judgments, with likelihood and confidence given separately, and does one of five things. It may request new intelligence collection at a stated cost, as the October 1962 overflight was requested. It may also wait, recommend an option if asked, act within delegated authority or refer to the decision-maker. A historical track of declassified cases, disguised so that a model cannot recognize them and with known outcomes, would run first at small scale in the first half of 2027, and a prospective track would seal every assessment and action at its checkpoint.

The five arms are matched in model, evidence, tools, collection budget and computing budget (Table 6). Arm A, the baseline matched to the layer in forecasting ability, decides from the frame and from probabilities that the strongest available forecasting procedure, with retrieval, produces afresh at each checkpoint. Arm B is arm A with the inquiry’s full transcript, unassessed, supplied to the model at each checkpoint. The judgment layer of Section 3, with explained change on every revision and the four controls on revision of Section 3.3, defines arm C. Disabling explained change in C gives arm D, and removing its separation of assessment from authority gives arm E. A and C each run at two or more levels of capability in the underlying model. An extension has practitioners decide from A’s or C’s outputs, and the combination must beat the better of human and AI alone, which such combinations often fail to do (Vaccaro, Almaatouq and Malone, 2024).

Table 6. Arms of the judgment-layer decision study

Arm	What it has	What its comparison with C isolates
A. Forecasting agent	The strongest available forecasting procedure with retrieval, queried afresh at each checkpoint	Whether the layer improves decisions beyond forecasting parity
B. Raw-history agent	A, plus the full unassessed transcript supplied at each checkpoint	Whether adjudicated structure adds anything to the full, unassessed history

Arm	What it has	What its comparison with C isolates
C. Judgment-layer agent	Graded sources, roll-up with dissent preserved, discriminating indicators, explained change, assessment apart from authority, referral	Reference arm
D. C without explained change	C with ICD-203 standard 7 disabled	The contribution of one trade-craft standard
E. C without the separation	C with assessment and authority to act merged	The safety claim of Section 7.2

Each comparison in the last column is with arm C, the reference arm.

The primary outcome is decision quality at each checkpoint. Blind panels of former analysts and decision-makers rate it on a rubric fixed in advance, seeing only the information available at that checkpoint, because knowing the outcome biases how a decision is judged (Baron and Hershey, 1988). Outputs are first matched for length and stripped of the layer’s vocabulary, since the project’s own model judges prefer longer outputs (Section 4.3). Reported beside it are decision loss, timeliness, unsupported actions, actions beyond authority, appropriate referral, useful collection forgone and cost. Every arm is also scored at every checkpoint on the process and understanding measures of Table 1 and on the relational integrity of its citations. Accuracy is a secondary outcome. The prediction to be registered, based on the null result of Section 4.2, is that A, B and C will differ in accuracy by less than a stated margin.

For this design the answer to Question 3 is no if C does not beat A on decision quality or loss. B matching C means raw history suffices, a margin that shrinks at higher capability means models absorb the layer, and D matching C means standard 7 carries no weight here. Gains in process or understanding without gains in decisions would answer Question 1 or Question 2 yes and Question 3 no. The separation is unjustified if E, without it, explores at least as widely as C and acts on no more of what it explores, or if C forgoes useful collection that E takes. Gains also fail to count if C wins only by referring or spending more, if they vanish in the subject area kept for the final test, or if ratings favour C only when its length or vocabulary is visible. Frames, rubric, panels, the size of effect that counts as improvement and the number of inquiries needed to detect it will be sealed before the first checkpoint is scored. Arm C’s components are only partly operating (Section 3), and each must operate before arm C is the layer described here.

5.3 Learning across inquiries and the supply of outcomes

A second phase would test the cycle of Section 3.5. In a fixed arm, the methods for choosing questions, retrieving cases, selecting critics and weighing sources stay constant, and in a learning arm they are updated only from inquiries whose outcomes were independently adjudicated. A raw-history arm gives the model the full unadjudicated record, and a self-graded arm learns through lessons the system scored itself.

Both phases are bound by the supply of adjudicated outcomes, which the project has measured. In July 2026 only one of fifty-five live claims in one reference series could be resolved by any instrument the project held. Against roughly 14,000 forecast questions still open, 2,539 of them past their horizon, 35 have been resolved through the project’s own resolution process, the first by independent corroboration on 24 September 2026. Questions that prediction markets settle

supply more, 347 resolved forecasts for the forecaster in use. The project treats a model’s reading of retrieved news and documents, checked against fixed criteria, as the primary route to resolution. Studies 1 to 4 and the historical track ease the constraint, because their primary outcomes need no resolution.

6. Whether better models will absorb judgment

The premise invites an objection. If AI systems already forecast about as accurately as superforecasters, and reached that point sooner than the benchmark’s authors projected (Bastani, Kuinskas and Reynolds, 2026), judgment may likewise arrive inside the model as it improves, making an explicit layer only a cost. Parity so far belongs to engineered systems built around models (Appendix A.1), so the forecasting record does not yet show such a capacity arriving inside the model. The objection therefore turns on whether the value of making judgment explicit survives improvement in the model.

On accuracy the objection holds for the project’s own methods. The tested structured method added no measurable accuracy (Section 4.2), so there was no accuracy value for improvement to erode, and models so often restored a removed element in their own words that clean removals were scarce (Appendix A.5). Aids to accuracy built around a model may lose their value as models improve. The project expects, as a hypothesis, that the cost of producing analysis will fall faster than the cost of checking it, which would raise the value of a record that makes checking easier.

The remaining value has three parts, and model improvement affects each differently. Substantive improvement in judgment, meaning better questions, preserved alternatives and explained revisions, is what model improvement threatens most directly. The decision study would test it at two capability levels. Checkability rests on a record kept outside the model, with adjudication against outcomes, sealed timestamps showing that a judgment preceded its resolution, a registry of how reliable each evaluator has proved, and portability across a change of model. The features that AI developers offer for forcing a model’s answer into a fixed format supply a checkable format and none of these. Authority remains with people, because a model cannot be the accountable party, and whether to act on an assessment is a further judgment that two institutions can reasonably make differently at the same probability.

7. Applications: strategy work and AI safety

7.1 Strategy work

Higher-value strategy work is the first application. Teams often each see part of a situation, and a system that keeps standing assessments with their components and reasons for change could connect them and trace how a policy change alters another actor’s incentives and the timing of a decision. The project has one such setting in operation and one specified. A ledger of dated judgments about the future of AI, with alternative explanations, has operated privately since 22 August 2026. The addition of new judgments has been paused since 5 September, because 73 per cent of its criteria could not be settled by retrieval, and none of its judgments has been settled by an independent instrument. A study of whether assistance improves a board’s preparation, its challenge of management and its reliance on what it is told is specified, and has not been run with a participating board.

7.2 AI safety

AI safety is the second application, because judgment is a key attribute of responsible decision-making. Several functions of judgment serve constructive and protective ends that can diverge. Broader search can find valuable options and generate unsupported proposals, and strict review can block errors and suppress worthwhile ideas. Evaluation therefore needs separate measures of useful contributions, unjustified acceptance, lost opportunities and human control, as AI-control research has done for monitoring and selective deferral in coding tasks (Greenblatt et al., 2023).

The design’s central safety claim concerns three things: a possibility the system has generated, an assessment it has accepted and an action it is authorized to take. Keeping the three distinct, and enforcing the authorization whenever the system uses a tool, is claimed to let a system explore more widely without acting on more of what it explores. The claim is specified, and arm E of the decision study would test it.

The operating record has taught three lessons so far. Outcomes the system had graded itself entered two learning loops, the evaluator registry and credit assignment, so the system was marking its own work (Sections 3.5 and 4.3). Treating the revision of an assessment as an action produced a bypass (Section 3.4). A record whose entries cannot be altered exposed six weeks in which the same judgments were reversed repeatedly (Section 4.3). Each was found by examining the record, which is written separately from the components whose behaviour it records.

Whose interests count and what authority is delegated are for the institution to decide. A system rewarded for agreement, volume or acceptance can drift away from better judgment. The design therefore specifies a record that keeps the user’s request, the institution’s authorization, the system’s inferences and the system’s rewards apart and retains the original set of questions, so that any drift toward easier questions can be detected.

8. Taking part

The project seeks collaborators, and intelligence analysts and tradecraft researchers come first. They can contribute difficult cases, historical assessment sets and expert raters, and gain a setting in which the effect of a single tradecraft standard can be measured. Strategy practitioners and institutions can contribute decision frames, raters and board or committee packs with their decision records. Documents stay with the institution, and results are released only in a form it has approved. AI safety researchers are needed for arm E and for the evaluator registry whose failures Section 4.3 documents, and researchers in AI, decision science and evaluation can enter competing systems or reproduce the internal findings.

The first collaborations are Studies 1 to 4, followed by the historical track of the decision study in the first half of 2027. The project supplies the tasks, test conditions, adjudication procedure and analysis, and both parties receive an independent assessment of the assistance, including its costs and failures. Data access, confidentiality and publication are agreed for each collaboration, and research with human participants follows institutional ethics review. Readers who want to take part, challenge the design or run the comparison on their own systems are invited to write to me at the address above.

9. Limits

The integrated system is unvalidated. Every measured result was produced within the project, most under a single model judge, and none has been independently replicated. Two board results with the largest effects have not yet been added to the archived evaluation record. The element-removal study departed from its registered measure, and the forecasting comparison may have been exposed to resolved outcomes. Much of what has been built is switched off or paused, and the operating counts come from one dated audit (Appendix A.8).

The premise is dated and contested. It rests on a human panel whose forecasts date from July 2024 and on a leaderboard rebuilt nightly, and the fresh superforecaster round planned for autumn 2026 may move it (Forecasting Research Institute, 2026). Evaluation design can dominate such results. In the project’s own data, taking market prices near resolution instead of thirty days out cut the market’s measured Brier error about 4.4-fold (Appendix A.3).

A record can contain every required element and still be wrong. It can also describe reasoning that the model did not in fact follow. Process measures risk circularity unless they are checked against decisions and outcomes (Section 5.1). Kay and King (2020) argue that radical uncertainty cannot be reduced to known probability distributions, a position that sits uneasily with criteria scored as probabilities. 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 layer offers no general guarantee against misuse or loss of control.

10. Conclusion

The best AI systems now forecast about as accurately as superforecasters. Forecasting remains a narrow skill, only as good as the question asked and a small part of how people and organizations understand their environment and decide within it. The intelligence community has long treated the surrounding process as what matters far more: how information is gathered and evaluated, how individual assessments are rolled up into broader judgments and how those judgments are systematically updated. This paper has asked whether AI systems can carry out that process, whether it enhances their strategic understanding and whether a judgment layer that carries it out improves an agent’s decisions.

On the first question, components of every stage are implemented, and those of every stage but the first are partly operating. Explained change and general detection of shared origin do not yet operate, and on the largest document examined none of ten sampled citations, in a record with every required element, supported its claim. On the second, structure made insights better grounded and more testable in one small comparison, and the system has not yet anticipated change better than the rule that nothing changes. The third has not been measured, and the study that would measure it is specified. AI systems capable of strategic understanding and judgment in this sense would serve higher-value strategy work and AI safety alike, recognizing the limits of their evidence and referring decisions beyond their authority to the people who hold it. Section 8 sets out how to take part.

Terms used in this paper

Agent: a system that pursues a task through repeated use of information and tools. *Agentic AI*: AI deployed as an agent, to act as well as to advise. *Arm*: one version of the agent under test in

a study, run under matched conditions. *Base rate*: how often an outcome occurs in a comparable class of cases, against which a record of success must be read. *Belief*: an assessment counted once across all its restatements. *Brier Index*: ForecastBench’s accuracy score, one hundred times one minus the square root of the difficulty-adjusted Brier score, higher being better. *Brier score*: the mean squared error of a probability forecast, lower being better, with 0.25 the score for always answering one half. *Calibration*: the match between stated probabilities and how often the forecast events occur. *Circular reporting*: apparently independent reports that trace to a single origin. *Closed pattern vocabulary*: the fixed list of kinds of insight the search may return. *Cohen’s kappa*: agreement between raters corrected for chance, 0 being chance and 1 perfect. *Conviction*: the probability that an assessment’s claimed mechanism holds, scored like a forecast. *Credit assignment*: apportioning a resolved judgment’s error to the assumptions, sources and methods that produced it. *Criterion*: the exact condition under which an indicator fires. *Criterion gate*: the test that refuses criteria likely to fire when nothing has happened. *Critic*: a second AI model that checks whether an insight’s cited material supports its claims. *Decision loss*: the loss, fixed in advance by practitioners over options and states, incurred by the option an agent chose. *Decision quality*: a blind panel’s rating of a decision on the information available when it was taken. *Dynamic process*: strategic assessment and judgment treated as continuing, in which information is gathered and evaluated, individual assessments are rolled up into broader judgments, and those judgments are systematically updated as new information and assessments arrive. *Element-competence score*: the mean of a system’s scores on the nine elements of Table 4. *Evaluator registry*: the record of how reliable each evaluator has proved, which the reliance controller consults. *Evidence-linked revision*: a revision tied to an identified new item of evidence or reasoning. *Expected calibration error*: the average gap between stated probabilities and observed frequencies. *Explained change*: a revision that states what changed since the last assessment and why, as tradecraft standard 7 of ICD-203 requires. *Handoff format*: the structured message in which one agent passes a judgment to another, stating the expected output and the authority granted. *Independent instrument*: a means of settling an indicator that does not rest on the system’s own records, either a structured query against an external data series or a model adjudicator reading retrieved text (Appendix A.8). *Indicator and indication*: a step an adversary or actor should or may take, and information that such a step is being taken (Grabo, 2002). In the system, an indicator is a dated observable condition, stated in advance with the direction it would move an assessment, which fires when its criterion is met. *Insight*: a finding of one of a fixed set of kinds produced by one of the system’s search routines. Most pass generation, criticism, citation audit and ranking, and theses from the synthesis step are published without the critic. *Judgment*: the formation, integration, use and revision of assessments about what matters, what is credible and what warrants inquiry or action under uncertainty. *Judgment layer*: the set of functions through which an AI system carries out the dynamic process of assessment. *Key judgment*: a principal conclusion of an assessment, carried with its own confidence level. *Learned change*: a change to the system’s own methods proposed from past results. *Levels*: a claim is specified when a manuscript or protocol states it, implemented when reference software does it and tests confirm the software’s properties, operating when it runs on live data, and measured when a study has estimated a quantity or effect, within its limits. *Locator*: the place in a source document to which a citation must lead. *Persistence*: the forecast that nothing changes. *Pre-registration*: a public record of a study’s hypotheses and analysis, fixed before the data are collected. *Prediction market*: a market in which traders buy and sell contracts that pay out if an event occurs, so that prices can be read as probabilities. *Promotion gate*: the test a candidate method must pass to replace the method in use. *Radical uncertainty*: uncertainty that cannot be reduced to known probabilities. *Reference series*: a continuing series of the system’s assessments on one subject, published until September 2026. *Reliance controller*:

the rule that decides whether an evaluator’s verdict is accepted, referred to a second evaluator or deferred to a person. *Retrieval*: searching for and supplying documents relevant to a question. *Roll-up*: the integration of component assessments into a broader judgment, in the intelligence community through coordination that keeps dissent visible. *Sealed*: fingerprinted when written, with the fingerprint combined into a daily value published in a public repository, and therefore checkable by a third party as having existed, unaltered, no later than the time that value was published. *Self-graded outcome*: an indicator settled against the system’s own records. *Settled*: of an indicator, found by an instrument to have occurred. *Shadow*: the state of a learned change that is recorded and scored and not applied to the system in operation. *Significance gate*: fixed rules that screen clusters of candidate news stories without an AI model, whose weights, by the project’s own record, were set before they saw live data. *Standing assessment*: an assessment kept open and revised as evidence arrives. *Standing question set*: a continuing set of forecasting questions kept open on one subject. *Storyline*: the system’s continuing narrative of a situation. *Strategic understanding*: a representation of the actors, incentives and constraints in a situation that identifies the question that matters for a decision, names the observation that would discriminate between rival accounts and anticipates the next change. *Structural coverage and relational integrity*: whether a record contains the required elements, and whether a cited claim rests on the cited evidence. *Structured belief*: one of the 30 beliefs, among the 33 settled by 25 September 2026, whose settlement rested on a query against an external data series. *Sufficiency test*: a published check that says when a single request to a model, with its answer held to a fixed format, is enough and no layer is needed. *Superforecaster*: a forecaster identified in tournaments as among the most accurate (Tetlock and Gardner, 2015). *Synthesis step*: the step that composes theses from several findings, publishing them without the critic. *Thesis*: a claim composed from several findings, stating which standing assumptions it supports or contradicts. *Tradecraft standard*: one of the nine rules in ICD-203 that govern how a finished assessment is written and evaluated. *Updating*: the revision of standing judgments as new information arrives, including explaining a change and declining to reopen a judgment when new evidence does not warrant it.

Appendix A. Internal studies: methods and limits

A.1 summarizes the external evidence behind the premise. A.2 to A.7 and A.9 report the project’s internal studies, and A.8 the operating audit. None of the internal results has been independently replicated.

A.1 The state of AI forecasting in September 2026

The premise of Section 1 rests on three sources, each with limits that should travel with it. Figures were read on 26 September 2026 and move as the leaderboards are rebuilt nightly.

ForecastBench compares forecasting systems with the median of a superforecaster panel elicited once, in July 2024, when 39 superforecasters over nine days each answered 20 questions drawn from a 200-question set, with a group stage partway through. At that time the best model was significantly worse than the panel, with a Brier score of 0.122 against 0.096 (Karger et al., 2025). Later systems answered different questions and are compared with the panel through a statistical model that separates forecaster ability from question difficulty. The panel’s score therefore changes as new questions enter although its forecasts do not, and the index cannot be compared across dates. The benchmark scores accuracy on its Brier Index, which adjusts for question difficulty and on which higher is better. On the tournament leaderboard of 26 September 2026 the strongest

system scored 69.1, with a 95 per cent interval of 67.9 to 70.6, against 68.7, with an interval of 67.3 to 70.4, for the superforecaster median. A one-sided test gave a p-value of 0.67 against the hypothesis that the superforecasters were better (ForecastBench, 2026). The Institute describes its July results as more consistent with parity than with outperformance, and parity here means a failure to reject equal accuracy, which is weaker than a formal test of equivalence (Forecasting Research Institute, 2026). About 375 system entries are scored against the one human baseline, so the top of the leaderboard, picked after the fact, is inflated by selection.

Parity belongs to engineered systems that add retrieval, structured prompting, combinations of models or further training. On the baseline leaderboard, where models run without tools, the superforecaster median scored 67.7 and the best model 62.2, significantly worse at $p < 0.01$ and no better than the median of the general public at 62.6 (ForecastBench, 2026). On questions drawn from prediction markets, forecasters are shown the current price. A baseline that simply submits that price scored 71.7, against 74.9 for the superforecasters, whose market score rests on 57 resolved questions, and 76.4 for the best system. Much of the market result therefore measures how well a forecaster uses the price it is shown, and Good Judgment Inc. (2026) reports a correlation of 0.994 between one model’s forecasts and the prices shown to it.

Metaculus supplies the other two sources. Its controlled comparison set its ten professional forecasters against its ten best automated systems on 99 shared questions between January and April 2026. The professionals led by 1.25 points a question, with a 95 per cent interval of -2.37 to 4.87 and a one-sided p-value of 0.247, down from 20.03 in the second quarter of 2025. Professionals took nine of the top ten individual places (Wilson and Bash, 2026). They are a different population from ForecastBench’s superforecasters. The Summer 2026 Cup, in which automated systems took first and second place in September 2026, was an open tournament of 885 forecasters on 58 questions, in which automated entrants are ineligible for prizes (Metaculus, 2026). Its scoring sums a score over every question and gives zero for a question not forecast, so forecasting every question early carries weight apart from accuracy on any one of them (Metaculus, 2024).

Several industry claims of parity are retrospective or backtested. One system matched the superforecasters on the 2024 question set, where exposure to outcomes cannot be excluded, and trailed the market consensus on questions from liquid markets, with a Brier score of 0.1258 against 0.1106 (Alur et al., 2025). As of July 2026 no controlled comparison had set superforecasters and a frontier system on the same questions, at the same time, with the same information and scoring (Hatch, 2026). Good Judgment Inc. has a commercial interest in the outcome. Every source above uses short-horizon questions posed in advance by someone else, and the Institute’s fresh superforecaster round, planned for autumn 2026, may move the comparison. The sharpest objection, from Good Judgment Inc., is that the 2024 panel forecast largely alone, with little of the discussion and updating through which superforecasters usually work (Good Judgment Inc., 2026). The benchmark paper reports that for the second half of the exercise the panel could see and discuss one another’s forecasts (Karger et al., 2025).

A.2 Structured insight generation

Fifteen shared evidence contexts from the reference series on AI computing were supplied to Gemini 2.5 Flash under structured and plain-prompt conditions, with the same evidence and the same limit on output length. 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 against 58.0 for evidential grounding, a difference of 28.3 with a 95 per cent interval, estimated

by resampling, of 17 to 41. For testability, scored as falsifiability, they were 53.3 against 30.7, a difference of 22.7 with an interval of 2 to 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 critic in operation, while a different model writes insights. Visible differences in presentation prevent full blinding. The comparison isolated the structured generation step and supplies no result for the complete process of generation and criticism, human-assessed usefulness or sustained performance.

A.3 Forecasting with a structured method

The comparison used 344 matched resolved questions. The structured arm put three fixed prompts to one model, for a base rate, an evaluation of the evidence and a self-critique, and was given only the text of each question. 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 per cent interval of -0.0244 to 0.0181 . The result establishes no accuracy advantage or equivalence. Both arms were significantly worse than the market price thirty days before resolution, which scored 0.122 on the same questions. An earlier structured run had counted 140 failed outputs as forecasts of 0.50. Its replacement 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 questions resolved after the models’ training data ended showed no material difference in Brier score, while the larger model was materially better at prose synthesis. One 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. In a pre-registered test of learning from worked examples on 120 questions, the result was null, with $p = 0.93$. 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 thirty days out instead of near resolution changed the market’s Brier score from 0.028 to 0.122 on the same questions, a 4.4-fold difference produced by timing alone, and narrowed the measured gap between system and market from 0.152 to 0.055.

The promotion gate admits a change to the forecasting method only if the candidate beats the method in use on paired questions resolved after the candidate was sealed, with the realized statistical power disclosed and a person making the final call. Its first provisional win was reversed on inspection. The candidate had been given part of the market price it was scored against, and on 67 questions with a market price it scored 0.1960 against 0.1781 for copying that price outright, so it was rejected as dominated by the market. Two other candidates remain inconclusive after 169 paired questions. Measured on 25 September 2026, the method in use, the gate’s incumbent, scored a Brier of 0.290 on 347 paired questions resolved by a prediction market, worse than the 0.25 of always answering one half, against 0.196 for copying the market price. Its forecasts carry no correction for miscalibration of the kind reported above, because the fitting procedure admits only 28 of the 100 resolved questions it requires. It does not yet read the forecasts of the method in use on questions that prediction markets have settled.

A.4 Agreement on structural conformance

Two independent checkers built on SHACL, the web standard for stating rules that a data record must satisfy, `shacl-engine` and `pySHACL`, evaluated sixty records against the published rules after conversion to the standard linked-data format. Both returned the same verdict on every record: 45 passed and 15 failed, three of the failures being records produced from a single instruction with no examples, as a comparison condition, each carrying about two violations. The project’s original 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 rules still lack two rules the reference validator enforces, each requiring a minimum number of some element, so the two validators do not yet encode the same definition in both directions. An earlier result in which every record passed, because missing type information meant no rule was actually applied, was withdrawn after correction. The result concerns conformance only and is silent on the truth or usefulness of what the records assert.

A.5 Removal of representational elements

A pre-registration fixed on 2 May 2026 set the rule under which an element of the vocabulary would count as load-bearing. The effect of removing it had to have a confidence interval excluding zero and a standardized size of at least 0.30, after correction for multiple comparisons across the nine elements. The element also had to have been removed cleanly on a minimum number of tasks. 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-part detector excluded runs in which the removed element had been rebuilt in prose. Under the locked rule at sixty questions, seven of the eight elements removed 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 the claim, treated as definitional, was counted as an eighth. Attribution was the exception, with an effect size of 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 elements 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 and 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, and the capability scores sit close to the elements removed, which makes the test circular.

A.6 Six widely used models on entry-level judgment tasks

Six models were evaluated on up to fifty tasks each, drawn from the benchmark’s first tier and stratified across its dimensions and difficulty levels. Two models completed 48 and 43. The models were Claude Sonnet 4.5, Claude Haiku 4.5, Gemini 2.5 Pro, Gemini 2.5 Flash, GPT-4o and GPT-4o-mini, run at a sampling temperature of 0.4, a setting that controls the randomness of output. 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 test procedure 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. An element-competence score was computed from the element scores, with 90 per cent intervals estimated by resampling the tasks.

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 for Claude Sonnet 4.5 to 49.8 for GPT-4o, the last falling below the 50 threshold at which the benchmark caps a system’s overall classification regardless of its other scores. Scores for updating belief as evidence arrived within a session ranged from 55.7 for GPT-4o to 72.6 for Claude Haiku 4.5, and scores for integrating information from 65.1 to 74.8. On a second-tier 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, which is the measured size of the self-preference bias the procedure now excludes. In a separate pilot audit, two model judges agreed with each other at Cohen’s kappa between 0.13 and 0.30, and both preferred longer outputs. The run is single-judge and author-run, with about ten tasks per dimension per model, and uses a first-publication rubric. Each run began with no memory of earlier runs, so continuity across sessions could not be tested. Reproduction takes about thirty minutes per model with the published task pool and test procedure.

A.7 The system against filter-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. Its output then passed through the system’s own locator validator and its lists of banned phrases and banned meanings, so that any remaining difference is attributable to the system and not to these filters. Raw output was ten questions per pack in both conditions. After the filters, the prompt retained seven, one, four and zero questions in order of pack size, and the locator validator did all the filtering. The system retained ten on each pack, all with valid locators, because it drops claims with unresolvable locators at extraction and generates questions only over the claims that remain. In this sample of four, the prompt’s retained questions fell from seven on the smallest pack to zero on the largest, though not steadily.

Separately, forty claim–locator–evidence triples from the system’s stored output, ten per pack, were judged by Gemini 3.1 Pro, a different model family from the generator. Under a strict check that the quoted passage appears word for word in the document, 27 of 40 locators, 67.5 per cent, resolved, against the in-process validator’s more permissive acceptance, and 13 of the 27 resolved triples, 48 per cent, were judged to support their claim, with no contradictions. The mean of the four per-pack support rates was 56.3 per cent. 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. A £729 million provision, planted in a document without disclosure, was extracted when placed about five thousand characters in, although it was not reported as a

finding. It was missed when placed at 340,000 or 627,000 characters, which is consistent with extraction reading only a 120,000-character window of each document. Among the most capable current models, variance between runs of one model was comparable to variance across vendors on the largest pack. On the smallest pack three of four models converged on the same score, 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 a manually coded reference set for checking the audit judge is held back for peer review. The procedure and documents are archived, and the outputs of these two runs are not yet.

A.8 The operating audit of 27 September 2026

Operating counts in this paper come from an examination of the project's operating records on 27 September 2026 that altered nothing, checked against the software that creates each kind of record. The audit is author-run and has not been independently repeated. Three definitions govern the counts. An indicator is settled by an independent instrument when a structured query against an external data series or a model adjudicator reading retrieved text settles it. A self-graded outcome is an indicator settled against the system's own records, and such outcomes are kept out of the track record. A count taken once per belief collapses an assessment's restatements to one belief and counts the settlement by an independent instrument that rests on the earliest observation. The number of beliefs settled does not depend on which settlement is chosen, and the number that held could.

For gathering and evaluating, observations averaged 5,047 a day over thirty days, from 18 of 23 registered sources, with five registered sources silent. Most of that volume carried no new information. In one week measured before 26 September, 63.8 per cent of observations belonged to series whose value never changed, and genuine changes numbered about 850 a day. Four sources had recorded zero values that were never measurements, among them 69,990 zero values among 73,130 observations in one family of news-mention series, while their scheduled processes reported success. On 26 September 2026 three of these four sources were repaired, and the fourth, whose public interface had closed, was registered as dormant. A repair that stops recording unchanged values was adopted the same day. It takes effect only once a further change to the system is applied, which had not happened by 27 September. In thirty days 137,624 candidate news stories were clustered and 3.3 per cent of the screened clusters passed the significance gate. The scheduled process that creates new records from admitted stories has been switched off since 5 September 2026. Of 12,294 graded evidence items, 88 per cent were in the "unknown" tier. Among the evidence retrieved to settle forecast questions, two websites imitating a major international news outlet supplied 178 of 5,332 items against 42 from the outlet itself, all graded "unknown", and no verdict rests on them. The general corroboration rule counts any distinct website within 48 hours whose title is sufficiently similar as corroboration. Shared origin is detected by a rule of thumb covering four source families listed by hand and thirteen websites, which last produced an assessment on 29 August. No evidence item has been converted into the numerical form needed to detect reports that say the same thing in different words, so such duplicates cannot be found. The weekly processes that should learn source reliability from outcomes update no source, because resolved predictions carry no source website. Of 13,894 open forecast questions, about 11,200 of them in the standing question sets and most of the rest generated from news, 2,539 are past their horizon and unresolved. Another 35 have been resolved through the project's own process, and the first resolution by independent corroboration came on 24 September 2026. The project has also posted one question of its own on a public prediction market.

For rolling up, 1,345 insight records from eleven search routines since 2 July collapse to 417 beliefs. The synthesis step produced 547 records forming 184 beliefs, each recording on average 4.1 supporting findings, and on all 547 the link to those findings is recorded only as the findings' headlines. None of the 336 syntheses of the last thirty days carries alternatives, information gaps or a base rate, and the synthesis step publishes without passing the critic. Until 26 September 2026 the process that watches indicators could also publish insights the critic had rejected, and it published 21 with no seal, while two steps could publish an insight whose seal had failed. Forecasts in 133 standing question sets carry about 15,000 links a month to named assessments and causal factors. Actor models cover fifteen states and twenty AI-sector firms, twenty-five with explicit profiles of priorities and tolerance of risk, and the process comparing stated with revealed positions has been disabled since 13 July. The assessment of each actor's stance was withdrawn on 26 September 2026, when all 189 recorded stances were neutral because the rule that set them could reach no other value. The paired test of the actor models holds 134 sealed pairs elicited between 28 August and 23 September, none resolved, and in 95 of the 114 active pairs the two forecasts are identical. None of the twenty-five profiles has been revised. Since 5 September the process that creates standing questions about actors' behaviour has been paused, and new pairs come from the process that publishes questions about how an actor will respond to another's move. Storylines are still created and have not been revised since 8 June.

For updating, 2,541 indicators have been set. Of these, 331 have been settled, 76 by independent instruments, 174 against the system's own records and 81 by earlier or since-reverted methods, while 192 were withdrawn as non-discriminating. Counted once per belief, 37 beliefs have been settled by an independent instrument and 32 held, 30 of the 37 being synthesis theses. None of the synthesis theses was settled on its own criterion, and counting only beliefs settled that way, with the track record's seven-day delay, leaves six scored beliefs on 26 September 2026, hits and misses together. An analysis of 25 September 2026 took the 33 beliefs then settled. Thirty had been settled by a structured query, and 26 of them held, against 25.7 expected from a baseline that applies the same criteria on every possible start date and never reads the thesis. One trading-volume reading of 27 August settled nine of the 33. Each of the 30 criteria that decided a structured belief had passed the criterion gate. The gate as revised on 26 September admits none of the 65 criteria behind those beliefs, refusing 54 as unreviewed, six as unrelated to their claims and five on its simulation of firing by chance. None of the 228 recorded changes to named assessments or the 282 to causal factors carries a reason for change, and no change of conviction on a held belief has been recorded. The sealed record holds 42,603 records, including 6,832 forecast revisions, and daily values are missing for 31 of the 118 days since 1 June, including 14 June to 1 July and 18 to 20 September, when no records were created. Of the 162 daily values in the public repository, 77 were first published more than a day after their date, 75 of them for days before the repository was created on 31 May 2026. The underlying data for 57 records sealed in July were later deleted. Between 23 July and 3 September a defect produced 1,275 reversal records on 187 assessments, up to 22 on one, which made up 53 per cent of all records of status changes when measured on 22 August. During the user trial the storyline component applied 975 revisions and declined 1,386, 721 on its daily budget and 665 on its minimum interval. All 1,402 learned changes since March remain in shadow. Of 156 sealed credit assignments, 13 rest on outcomes settled by an independent instrument and 117 on self-graded outcomes. In the evaluator registry, 177 of the 191 outcome records bearing on the critic in operation were self-graded. On 30 August they took the critic past the number of outcomes at which a low reliability triggers referral, and by 26 September its recorded reliability was 0.26, on a record of 36 of 142. The 186 referrals to a second critic from 31 August changed no verdict. Recomputed on independently settled outcomes, the critic's record is 8 of 14. The registry had not

been corrected on 27 September. The weekly next-change test of 20 September, over eight windows, scored a Brier of 0.341 on which indicators would fire, against 0.020 for persistence.

For deciding, 15,540 automated action requests since 21 April were all refused, 965 of them in the last sixty days being attempts to change an assessment’s status that lacked a person’s confirmation. After each refusal the system records the status change itself and seals it, so the controls requiring a recorded reason, a daily budget and a minimum interval never run on these changes. An action that declares no required authority runs without confirmation. The mechanism that holds irreversible actions for a person’s commitment has never been used, and decision records hold four entries, the sealed ones from a single governance walkthrough on 26 July. None of the 693 insights of the last thirty days cites a decision frame, and 106 of them, from one search routine, carry unrequested prescriptive recommendations. The reliance controller has recorded 3,718 decisions, 902 of them deferrals, and the 458 that exercised oversight are sealed. When the system is producing insights, a deferral keeps the first critic’s verdict once the budget for second opinions in that run is spent. Elsewhere in the system, a deferral marks the item for a person’s review. The human review queue holds 512 items, and people have cleared 16. Such controls have costs. An operating rule requiring a second source before automatic settlement, which a market’s own settlement could not supply, sent every question whose market had settled to a review queue in which no person had acted since 30 August. Model inference, meanwhile, was allowed to count as a second source elsewhere.

A.9 Pre-decisional policy papers

One pre-decisional policy paper on political instability in Cuba was produced by the full process, drawing on live evidence, and one by a single request to the same model, Claude Sonnet 4, asking for the same kind of paper on the same issue. Claude Sonnet 4 then scored each paper against seven criteria derived from ICD-203, after the process’s own grade of its paper had been removed so that the judge scored the content. The full process scored 82.9 and the single request 55.7. The largest differences were on assumptions, 85 against 30, the characterization of sources, 75 against 25, and alternatives, 80 against 40, and the judge found that the single request made policy recommendations. Each condition was scored on one paper by one judge, the same model that wrote both papers, and scores moved between runs, from 84.3 to 82.9 for the full process and from 43.6 to 55.7 for the single request. The papers were generated by the author and are not archived.

Appendix B. Project components by level of existence

Levels are defined in Section 1. The table follows the stages of the dynamic process and records, at the level the operating audit of Appendix A.8 verified on 27 September 2026, what is switched off, paused or bypassed as well as what runs.

Table B1. Project components by stage of the process and level of existence

Stage	Component	Level on 27 September 2026	What it establishes	Not yet established, or not running
Foundations	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
Foundations	Organizational theory: judgment-centric work, four levels of judgment, twenty-one propositions with disconfirmation criteria	Specified	Institutional propositions specific enough to fail	Any of them empirically
Foundations	Shared vocabulary of nine elements with strategic extensions, a published machine-readable definition with a validator, a handoff format and mappings to provenance and doctrine standards	Implemented. Validator agreement measured (A.4)	That a checkable judgment record can be built and passed between agents. Sufficiency by construction: a joint-doctrine risk assessment, a chest-pain differential diagnosis, a negligence determination and a peer-review recommendation each expressed from the nine elements alone, and eighteen board-governance concepts reduced to the same nine under a mapping checked on every change to the software	That it improves inquiry over simpler records
Setting the question	Decision frame in the vocabulary, and the question-finding study (Study 4)	Frame implemented and unused. Study specified	A place to record the question, options and stakes	Any recorded frame in operation. None of 693 recent insights cites one
Gathering and evaluating	Intake of dated observations from open sources	Operating	About 5,000 observations a day from 18 of 23 registered sources	Five registered sources are silent. Most observations repeat unchanged values until an adopted repair takes effect. Zero values that were never measurements stopped on 26 September 2026

Stage	Component	Level on 27 September 2026	What it establishes	Not yet established, or not running
Gathering and evaluating	News triage by a significance gate whose weights were recorded as fixed before live data	Operating	About 138,000 candidates a month triaged without a model, about 3 per cent of story clusters admitted, outlet count used only for prominence	The scheduled process that creates new records from admitted stories has been switched off since 5 September 2026
Gathering and evaluating	Retrieved evidence labelled by how it was obtained, with a sealed record of the search	Operating	How each item was obtained	Detection of reports that say the same thing in different words
Gathering and evaluating	Grading of the publishing website	Operating as a fixed list of known websites	A first grade for each item's channel	No grade of the originating source or of the information's credibility. 88 per cent of recent evidence is "unknown", and impersonating websites are graded like any unfamiliar source
Gathering and evaluating	Corroboration and detection of shared origin	Operating. Detection of shared origin is limited to thirteen websites and last produced an assessment on 29 August 2026	A rule of thumb for four source families and thirteen websites	General detection. The corroboration rule counts near-identical reports on different websites as independent
Gathering and evaluating	Independence guard at settlement	Operating	A market cannot confirm a settlement whose evidence cited that market	The same guard between news sources
Gathering and evaluating	Source reliability learned from outcomes	Operating weekly, with no effect	A mechanism with its own audit trail	Any updated source. Resolved outcomes are not linked to source websites
Gathering and evaluating	Supply of evidence to settle questions	Operating	35 questions resolved through the project's own process, the first by independent corroboration on 24 September 2026, against about 14,000 still open	Resolved outcomes in sufficient number

Stage	Component	Level on 27 September 2026	What it establishes	Not yet established, or not running
Rolling up	Insight generation: closed pattern vocabulary, generation, adversarial critic, citation audit, ranker	Operating. Structured generation measured (A.2)	417 distinct beliefs since July, each tracked across its restatements	Full-process effectiveness against a single request to a model. The synthesis step bypasses the critic
Rolling up	Synthesis of findings into theses	Operating	184 theses, each from about four findings, stating which standing assumptions they support or contradict	A traversable link to components. Recent theses carry no alternatives or gaps
Rolling up	Standing question sets linked to named assessments and causal factors, and a trace of what fed each judgment	Operating	133 standing question sets, about 15,000 links a month, about 2,900 scored errors of resolved forecasts attached	That the links improve any forecast
Rolling up	Actor models of states and firms	Operating	Fifteen states and twenty AI-sector firms, twenty-five with profiles of priorities and tolerance of risk	Their contribution to forecast quality. One comparison process disabled since 13 July. Stance assessment withdrawn on 26 September 2026
Rolling up	Paired test of actor models	Operating	134 same-moment pairs sealed	Any scored pair. Identical forecasts in both arms of 95 of 114 active pairs. Profiles never revised. One of the two processes that create pairs paused since 5 September
Rolling up	Storylines and scenarios	Operating for creation only	Continuing narratives of a situation	Revision. None since 8 June, when the user trial was suspended
Updating	Standing positions with conviction and dated indicators	Operating. Baseline comparison measured (A.8)	2,541 indicators, 76 settled by an independent instrument, 174 self-graded kept out of the track record	Any evidence of skill. 37 beliefs settled and 32 held. Six scored on the claim's own criterion, hits and misses together. Among structured beliefs, a hit rate equal to a baseline that never reads the thesis

Stage	Component	Level on 27 September 2026	What it establishes	Not yet established, or not running
Updating	Criterion gate for indicators	Operating and enforced by default. Revised 26 September 2026	192 non-discriminating criteria across 127 assessments withdrawn as sealed retractions	Discrimination in operation. Before the revision the gate admitted every criterion that later decided a structured belief
Updating	Revision discipline: a signal that must persist, a minimum interval, a check that reverses degrading changes, a daily budget	Implemented. Operated on storylines until June 2026. Bypassed for assessments	975 revisions applied and 1,386 declined in the trial	Any effect on the accuracy or stability of judgment
Updating	Reason for change	Implemented, empty in operation	An element for standard 7	Any recorded reason. No change of conviction on a held belief has been recorded
Updating	Sealed revision record, with a daily value published in a public repository and a script for checking it	Operating	About 42,600 sealed records, including about 6,800 forecast revisions, with outages visible as gaps. A September 2026 repair to the links between successive seals has held	Predictive advantage or decision usefulness. Records sealed before June are shown only to have existed by a later publication
Updating	Promotion gate for forecasting methods	Operating	A reversed first win and two inconclusive candidates (A.3)	An adopted improvement
Updating	Evaluator reliability registry	Operating, contaminated. Critic's record on independent outcomes measured (A.8)	Reliability updated from outcomes	A registry free of self-graded outcomes. Uncorrected on 27 September 2026
Updating	Computational component for numerical comparison	Implemented, inactive	Records the basis of each numerical sub-claim	Any operation on live data
Deciding under authority	Action registry with authority tiers	Operating	Every one of 15,540 automated requests since April refused, 965 in the last sixty days for want of a person's confirmation	Correct defaults. An action that declares no required authority runs without confirmation

Stage	Component	Level on 27 September 2026	What it establishes	Not yet established, or not running
Deciding under authority	Status changes recorded directly after a refused request	Operating	Status changes made after a refused request are sealed	The controlled route for revision. The controls on reason, budget and interval never run. Until 26 September 2026 the process that watches indicators could also publish insights the critic had rejected, 21 of them with no seal, and two steps could publish an insight whose seal had failed
Deciding under authority	Staging of irreversible actions until a person commits	Implemented, unused	A mechanism requiring two separate commitments for one operation	Any use in operation
Deciding under authority	Authority checks for external agents	Implemented, recording refusals only at first	Checks applied when an outside agent's request arrives	Enforcement in operation, which the software alone cannot show
Deciding under authority	Advice given only when asked	Specified, not enforced	The rule	Enforcement. One search routine issues unrequested prescriptive recommendations
Deciding under authority	Reliance controller	Operating	About 3,700 decisions, 902 recorded as deferrals, those exercising oversight sealed	That its referrals improve verdicts. Lone adverse verdicts are discounted
Deciding under authority	Human review queue and decision records	Operating queue. Records implemented	512 queued items, 16 cleared. One governance walkthrough recorded	Any real decision recorded against an assessment
Learning across inquiries	Credit assignment from resolved judgments	Operating, recording without acting	156 sealed assignments	Sound inputs. 117 rest on self-graded outcomes and 13 on independently settled ones
Learning across inquiries	Learned changes held in shadow	Operating in shadow	1,402 recorded changes	Any applied change
Learning across inquiries	Phase-two study of learning across inquiries	Specified	A design with fixed, learning, raw-history and self-graded arms	Any result

Stage	Component	Level on 27 September 2026	What it establishes	Not yet established, or not running
Demonstrating understanding	Benchmark with a capped overall score	Implemented. Six-model run measured (A.6)	An overall score in which forecasting skill cannot compensate for weak judgment	Cross-session continuity, never scored for any system
Demonstrating understanding	Next-change prediction test	Operating weekly, measured	A test of understanding as a process	Performance above persistence
Demonstrating understanding	Ladder of structural coverage, relational integrity, behavioural correspondence and causal sensitivity	Lower two rungs measured (A.7)	That form and support come apart	A scorer for the upper rungs
Settings and applications	Reference series on AI compute and on AI policy and economics, and a ledger of judgments on AI futures	Operating. Series withdrawn from public view in September 2026. Ledger private, no new judgments added since 5 September 2026	Continuing, contested settings for the studies	Public track records
Settings and applications	User trial	Implemented. Operated with pilot users until June 2026, now suspended	Operating history for storylines and revision discipline	Current use
Settings and applications	Applied profiles: board and decision-forum oversight, pre-decisional policy papers, governance attestation, an analyst’s desk check against ICD-203	Specified and implemented. Policy-paper evaluation measured with one model judge (A.9)	Study designs and instruments	Professional usefulness. The board study has not run with a participating board
Settings and applications	First studies and the decision study	Specified	Designs and falsifiers (Table 1, Section 5)	Any result

Appendix C. Nearest implemented precedents

The design draws on bounded rationality, metareasoning, belief revision, belief–desire–intention agents and sensemaking (Simon, 1955; Russell and Wefald, 1991; Alchourrón, Gärdenfors and Makinson, 1985; Rao and Georgeff, 1995; Weick, Sutcliffe and Obstfeld, 2005). Table C1 compares the layer with its nearest implemented precedents at each stage of the process of Section 2.

Table C1. What the layer adds to its nearest implemented precedents, by stage of the process

Approach	Gathering and evaluating	Rolling up	Updating	Learning across inquiries
Forecasting system without memory between questions, as on the benchmarks of Section 1	Retrieval afresh for each question	Averaging of several models' forecasts, where used	At most one probability revised	None
Reflexion (Shinn et al., 2023)	Not addressed	Not addressed	Self-graded reflection on a failed attempt	The same task, re-tried
Voyager (Wang et al., 2023)	Execution feedback	Not addressed	Skills revised on execution feedback	A library of skills in the same environment
Raw history supplied in full	A transcript, unassessed	None	None recorded	Any, unadjudicated
Structured outputs from a model interface	A fixed format only	A fixed format only	A fixed format only	None
DIVE (Friedman et al., 2020)	Provenance and appraisals	Contributions of several agents within one analysis session	Confidence propagated through evidence links	None
Truth maintenance (Doyle, 1979)	Not addressed	Dependencies among beliefs	Reasons recorded for each belief	One reasoner
Judgment layer (this paper)	Grades of source and information, and detection of shared origin	Component assessments integrated with dissent preserved	A stated reason for each change	Lessons adjudicated against outcomes, with evaluator reliability carried

Characterizations follow the cited works' own descriptions. The entry for the judgment layer is a design claim, and Appendix B gives the level of each component. In operation its registry has so far been fed self-graded outcomes (Section 4.3), and the claim would be tested by the study of Section 5.3.

Research materials and status

Several artifacts are public and verifiable now. The vocabulary is published in standard machine-readable formats for data exchange and rule checking, 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>. A sufficiency test runs at <https://leadershipunderuncertainty.org/api/v1/public/sufficiency-test>, and the agent handoff format is published as an extension to an open protocol for communication between agents at <https://leadershipunderuncertainty.org/sco/a2a/extensions/handoff-envelope/v1>. The daily values of the sealed record, a sample seal and a script that checks the chain of seals 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, the governance walkthrough and the

queries of the operating audit in Appendix A.8 are held in the project’s research collection and are available on request until their public release. An external agent can read and write the judgment record through published interfaces, and a decision recorded with its evidence, alternatives and authority can be exported as a dated, sealed record for independent audit. Both are implemented. Detailed specifications, reference implementations and study records for the remaining components are held in the project’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.
- Alur, R., Stadie, B. C., Kang, D., et al. (2025). AIA Forecaster: Technical report. arXiv:2511.07678.
- Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., and Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes*, 160, 19–35.
- Baron, J., and Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, 54(4), 569–579.
- Bastani, H., Kuinskas, S., and Reynolds, M. (2026). LLMs are closing the gap on human super-forecasters. Forecasting Research Institute, Substack, 29 January 2026.
- Butler, R. (2004). *Review of intelligence on weapons of mass destruction: Report of a Committee of Privy Counsellors*. HC 898. The Stationery Office.
- Cabinet Office. (n.d.). Joint Intelligence Organisation. GOV.UK. Accessed 27 September 2026.
- Chilcot, J. (Chair). (2016). *The report of the Iraq Inquiry: Executive summary*. HC 264, 6 July 2016.
- Clausewitz, C. von. (1832/1873). *On war*, trans. J. J. Graham. Project Gutenberg eBook #1946. Cited at Book I, chapter 6, "Information in War".
- 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.
- Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence*, 12(3), 231–272.
- ForecastBench. (2026). Tournament and baseline leaderboards, `leaderboard_tournament.csv` and `leaderboard_baseline.csv`, `forecastbench-datasets` repository, Forecasting Research Institute. Nightly update of 26 September 2026.

Forecasting Research Institute. (2026). AI models have likely reached parity with superforecasters on ForecastBench. Substack, 16 July 2026.

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.

Good Judgment Inc. (2026). Not so fast: If ForecastBench were a race, the Superforecasters left the track two years ago. Good Judgment Substack, 24 July 2026.

Grabo, C. M. (2002). *Anticipating surprise: Analysis for strategic warning*. Joint Military Intelligence College, Center for Strategic Intelligence Research.

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

Hatch, W. (2026). The hype and the evidence about AI forecasting. Good Judgment Substack, 6 July 2026.

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).

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. (2025). ForecastBench: A dynamic benchmark of AI forecasting capabilities. *International Conference on Learning Representations 2025*. arXiv:2409.19839.

Kay, J., and King, M. (2020). *Radical uncertainty: Decision-making for an unknowable future*. The Bridge Street Press. US edition: *Radical uncertainty: Decision-making beyond the numbers*. W. W. Norton.

Kelly, M. O., Budescu, D. V., Dhami, M., and Mandel, D. R. (2025). The effect of source reliability and information credibility on judgments of information quality in intelligence analysis. *Judgment and Decision Making*, 20, e36.

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* (History Staff, Center for the Study of Intelligence, Central Intelligence Agency, 1994), 173–187.

Kent, S. (1976). The law and custom of the National Intelligence Estimate. In D. P. Steury (ed.), *Sherman Kent and the Board of National Estimates: Collected essays*. History Staff, Center for the Study of Intelligence, Central Intelligence Agency, 1994.

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.

Knight, F. H. (1921). *Risk, uncertainty, and profit*. Houghton Mifflin. Cited at Part I, chapter I, and Part III, chapter VII (Econlib).

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.
- Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., and Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, 21(1), 1–14.
- Metaculus. (2024). Introducing simpler, fairer tournament scores. Metaculus notebook 21883.
- Metaculus. (2026). Announcing: the Metaculus Cup Summer 2026 winners! Metaculus notebook 45734, 18 September 2026.
- National Intelligence Council (NIC). (2007). *Iran: Nuclear intentions and capabilities*, National Intelligence Estimate Key Judgments, including "National Intelligence Estimates and the NIE Process". Reproduced in Trevorton (2013).
- 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.
- Professional Head of Intelligence Assessment (PHIA). (2025). PHIA Common Analytical Standards. GOV.UK, 24 March 2025.
- 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 metareasoning. *Artificial Intelligence*, 49(1–3), 361–395.
- Senate Select Committee on Intelligence (SSCI). (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.
- Tetlock, P. E., and Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown.
- Thucydides. (1874). *History of the Peloponnesian War*, trans. R. Crawley. Project Gutenberg eBook #7142. Cited at 1.22.
- Trevorton, G. F. (2013). *CIA support to policymakers: The 2007 NIE on Iran’s nuclear intentions and capabilities*. Center for the Study of Intelligence, Central Intelligence Agency.
- 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.

- US Government. (2009). *A tradecraft primer: Structured analytic techniques for improving intelligence analysis*. Central Intelligence Agency, March 2009.
- 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). FutureEval Spring results: Pros beat bots, but the gap is nearly gone. LessWrong, 9 September 2026. First published as Metaculus notebook 45373, 2 September 2026.
- Wright, J. R. (1962). Analysis of SAM sites in Cuba, briefing paper, 1 October 1962. Avalon Project, Yale Law School.