WORKING PAPER

Milestone Planning and Research, Inc.

Hybrid Cognition:

A Framework for Human-Machine Judgment Under Inductive Uncertainty

John Aaron, PhD

Milestone Planning and Research, Inc.

May 2026 · Version 8

Publication and Use Notice

Authorship and Responsibility. The author assumes responsibility for the conceptual framework, empirical interpretation, and conclusions presented here. Analytical software and large language models were used as research, coding, interpretive, and drafting aids under human supervision.

Abstract

The dominant discourse surrounding artificial intelligence in professional settings oscillates between two inadequate positions: uncritical adoption and reflexive anxiety. Both share a common error: they treat the human-AI boundary as a fixed line to be managed rather than an architecture to be designed.

This paper introduces Hybrid Cognition as a formal design discipline: a principled separation of cognitive roles in which machine systems observe and orient through evidence accumulation, pattern recognition, causal hypothesis formation, trajectory projection, and semantic search over alternative courses of action, while human agents retain exclusive authority over belief commitment, consequence evaluation, and irreversible decision-making. The framework is grounded in Karl Popper's falsificationism, Karl Friston's Free Energy Principle, Bayesian evidence architecture, the Godel-Turing constraints on self-referential systems, and John Boyd's theory of orientation, and is operationalized through one concrete system exemplar: PRIMMS, an AI-augmented project risk intelligence platform.

The central claim is that not only does AI improve managerial performance in business project environments, but that the failure modes and comparative advantages of human and machine cognition can be deliberately composed. PRIMMS® makes this composition explicit: deterministic governing equations establish what the evidence warrants; the LLM performs abductive causal analysis and searches a broad semantic solution space for courses of action; an adversarial OCC-5C phase-gate layer recursively challenges the machine's own orientation; and the human decides. Boyd's proposition that orientation is the decisive element of the decision cycle is therefore operationalized without transferring decision authority to the machine. The machine's role is to improve the situation presented to judgment: earlier evidence, better causal hypotheses, broader alternatives, explicit uncertainty, and structured falsification pressure. Human authority remains architecturally necessary because sufficiently expressive inductive systems cannot regulate themselves from within, cannot guarantee that their generative vocabulary still corresponds to the world, and cannot own the consequences of irreversible action.

Keywords: *hybrid cognition, Godel-Turing constraint, Zero constraint, micro reliability, gestalt reliability, falsificationism, Bayesian evidence architecture, Free Energy Principle, hierarchical predictive processing, narrative forecasting, project risk intelligence, organizational perception, epistemic vigilance, project management*

1. The Problem: Five Failures, One Architecture

1.1 The Practitioner Conversation

This paper begins with practitioner testimony. LinkedIn posts, practitioner blogs, and professional forums that constitute the current public discourse on artificial intelligence contain a striking pattern: high engagement, genuine urgency, accurate symptom identification, and almost no framework. The practitioners writing these posts are not naive. They have watched organizations spend hundreds of thousands of dollars on AI and produce graveyards of pilots. They have seen knowledge workers walk into rooms with AI-generated research they cannot defend. They have watched agentic systems fail not because the models were weak but because the surrounding architecture was never designed. And they have identified, correctly, that the bottleneck has shifted from capability to judgment, from building to knowing what to build.

The commentary they produce captures five recurring failure patterns. First, the deployment trap: organizations fund pilots that demonstrate local capability but were architecturally doomed from day one, because no one asked whether the operational system could absorb, govern, validate, and maintain the AI under real conditions. Second, the confidence transfer problem: professionals present AI-generated analysis as their own judgment, then cannot defend it when challenged. The AI did not make them wrong; it made them confidently wrong. Third, the explainability failure: systems shut down not because they performed poorly but because their reasoning could not be traced. A CFO cannot approve a recommendation she cannot explain. Fourth, the data pipeline fallacy: organizations race to deploy AI faster while scaling fragmented, inconsistent, and poorly understood data at machine speed, converting mediocre signal into authoritative output. Fifth, the escalation gap: agents that do not know when they do not know, that keep executing tasks which should have been returned to a human three steps ago.

The discourse already contains precise observations. What it lacks is a framework that explains why these failures are structurally related and how an architecture that addresses all five simultaneously can be designed. That framework is Hybrid Cognition.

1.2 A Note on Terminology: What We Mean by AI

The term 'artificial intelligence' has become so elastic in popular usage as to be nearly uninformative. In professional discourse it typically refers to large language models and their derivatives, because these are the systems most practitioners encounter daily. But the practitioner commentary quoted above conflates large language models, statistical forecasting models, decision support systems, recommendation engines, and agentic workflow tools under a single

label. The advice generated for one category is often inapplicable or actively misleading for another.

In this paper, artificial intelligence refers to any system that relies upon statistical learning and is, therefore, probabilistic: the extraction of patterns from data through algorithms that update internal parameters in response to evidence. This definition includes large language models, which learn probability distributions over token sequences from text corpora. It includes statistical forecasting models, which learn parameter estimates from time series data. It includes classification systems, recommendation engines, and Bayesian inference frameworks. It excludes purely rule-based expert systems and deterministic algorithms, which are not statistical learners. The MATLAB signal extraction layer in PRIMMS is an example of the latter: it is a deterministic algorithm, not an AI system in our definition. The LLM interpretation layer within PRIMMS is an AI system. The IUVO forecasting system (also in the Milestone Planning suite) is also an AI system. The distinction matters because the failure modes of statistical learners, and therefore the architectural requirements for governing them, differ systematically from those of deterministic algorithms.

The conflation of all AI with LLMs has produced two kinds of error in the practitioner discourse. The first is over-generalization: advice calibrated to the specific failure modes of language models (coherence bias, authority-without-warrant, self-referential governance failure) is applied to statistical forecasting models, which have different and more tractable failure modes. The second is under-specification: advice calibrated to simple workflow automation is applied to consequential decision-support systems that require a different architecture entirely. Hybrid Cognition is designed for the full range of statistical learning systems, with specific architectural requirements that vary by system type.

1.3 The Coherence Problem and the Distributional Problem

Large language models and statistical forecasting systems present a superficially similar appearance of confident knowledge, but for fundamentally different reasons that deserve precise treatment. A language model is optimized for coherence: it produces text that is locally plausible, globally consistent, and stylistically fluent, and it has no internal mechanism for registering when that fluency is disproportionate to the available evidence. A well-specified statistical forecasting model is a different matter: it is built on unbiased estimators, asymptotically approaches truth under correct model specification, and propagates uncertainty honestly through its confidence intervals. The forecast cone in IUVO is a formal statement of the model's estimated error distribution, not a rhetorical gesture at precision.

The concern with statistical forecasting is therefore more specific: the uncertainty bands are computed from historical residuals, and when the model is operating outside its training distribution, in a novel regime the residual history does not represent, the cone retains its authoritative appearance even as its reliability degrades. The model has no internal mechanism for detecting that its assumptions have been violated. This is why the regime stationarity check is the most important single falsification check in the IUVO architecture: it is the system asking whether it still belongs to the world it was trained on.

For language models, the coherence problem is not a defect to be solved by the next model generation. It is a design characteristic: a system trained to minimize prediction error on a coherent corpus learns to suppress uncertainty markers, because uncertainty markers are typically absent from high-quality training data. The result is a model that speaks with authority it does not have, regardless of whether that authority is warranted. The practical consequence was demonstrated empirically by Dell'Acqua et al. (2023): knowledge workers using AI on tasks outside the capability frontier saw their performance drop by an average of 19 percentage points compared to the control group. The AI did not make workers wrong; it made them confidently less accurate.

1.4 What the Discourse Has Identified and What It Has Not

The practitioner commentary referenced above has correctly identified the symptoms. AI deployment without clean data pipelines produces accountability debt at machine speed. Judgment and knowing what to build and why are now the scarce resources in knowledge work. An organization's language is a leading indicator of its outcomes. Explainability is the adoption condition, not a feature. Agents without escalation logic are liabilities, not tools. Each of these observations is accurate, empirically grounded, and practically important.

What this discourse has not produced is a unified explanation of why these failures are structurally related, or a framework that addresses them simultaneously. They are not independent problems. They are five manifestations of the same underlying architectural error: the conflation of the machine's orientation function with the human's decision function, combined with the absence of the falsification layer that keeps the whole system honest. Hybrid Cognition is the framework that addresses all five.

2. Theoretical Foundations

2.1 Popper and the Architecture of Falsification

Karl Popper's central contribution to epistemology was the asymmetry of evidence: no number of confirming observations can establish a proposition as true, but a single well-designed disconfirmation can establish it as false (Popper, 1959; 1963). The practical implication is that the goal of inquiry is not the accumulation of confirmations but the construction of tests that could, in principle, reveal that a belief is wrong.

This asymmetry applies directly to AI system design. A forecasting system evaluated only on its hit rate is optimized for confirmation, it learns to produce outputs that look right, not outputs that are hard to falsify. The two objectives diverge precisely in the cases that matter most: novel situations, regime changes, and the edges of the training distribution.

Hybrid Cognition applies Popperian logic at the system architecture level. The question is not: 'How often is the machine right?' It is: 'What would have to be true for the machine to be wrong, and is the system actively searching for those conditions?' A well-designed hybrid system continuously increases its sensitivity to possible wrongness: through contradiction audits, model disagreement measures, confidence proportionality checks, and longitudinal falsification pressure.

> *The goal is to make it harder for obvious wrongness, overconfidence, drift, or contradiction to survive unnoticed.*

2.2 Friston's Free Energy Principle and Hierarchical Predictive Processing

Karl Friston's Free Energy Principle provides a computational account of how biological systems maintain effective models of their environment (Friston, 2010; Friston, Kilner, & Harrison, 2006). Any adaptive system must minimize the difference between its predictions about the world and the sensory evidence it receives: through either updating the internal model (learning) or acting on the environment (acting).

Applied to organizational cognition: an organization's distributed information-processing system constitutes a collective predictive model of the business environment. When that model has drifted from the environment it is supposed to represent, surprises are large, late, and expensive. Three failure modes map directly onto organizational pathologies: the vigilance failure (the model becomes self-sealing against disconfirming evidence), overcorrection (signal lost in noise), and accuracy-appearance optimization (producing plausible outputs rather than accurate ones).

Hybrid Cognition addresses all three by assigning model-updating to the machine, which processes signals without social pressure to maintain face, and consequence evaluation to the human.

2.2.1 Hierarchical Self-Checking and Its Limits

Friston's most architecturally significant contribution is the hierarchical extension of the Free Energy Principle. In hierarchical predictive processing, higher levels generate predictions tested against lower-level computations, with precision-weighted prediction errors propagating upward to force model revision. The system partially checks itself, not through external validation but through the structural constraint that higher-level outputs must remain consistent with lower-level evidence.

A properly layered AI architecture: statistical computation below, language model interpretation above, contradiction audit as the prediction error signal between them, implements a weak but genuine form of this self-checking. The contradiction audit in IUVO is one operational instantiation; the deterministic MATLAB signal layer in PRIMMS is another. In both systems, the machine is, in a limited sense, checking itself.

The critical question is whether this self-checking is sufficient to close the gap between machine and human accuracy in consequential decision-making. The answer is: partially, for a specific and important class of errors. Friston's hierarchical system self-checks within the model's own generative structure. It detects mismatches between levels of the same model. What it cannot detect is the mismatch between the model's entire generative structure and the actual data-generating process in the world. It checks internal coherence well, but cannot check whether its priors are correctly specified in the first place.

The defensible claim is this: a Fristonian hierarchical architecture narrows the gap between machine and unaided human accuracy specifically for the class of errors arising from internal incoherence: misspecified features, contradictory model signals, distributional inconsistencies within the training range. These are also exactly what human judgment fails to catch under cognitive load, time pressure, and social convergence. For the domain of errors a competent but overloaded analyst would miss on a normal day, a well-implemented hierarchical system may plausibly match or exceed unaided human performance. The residual gap, where human judgment retains irreplaceable authority, is the class of errors requiring recognition that the entire generative model is wrong.

2.2.2 Micro Reliability and Gestalt Reliability

The distinction between what machines can self-check and what humans must judge maps onto a deeper distinction that the Fristonian framework illuminates but does not name explicitly: the distinction between micro reliability and gestalt reliability.

Micro reliability concerns the internal properties of a model run: whether the data conforms to schema, whether features were constructed correctly, whether ensemble weights sum to one, whether the forecast cone is geometrically consistent, whether model disagreement is within historical norms, whether recent performance has drifted. These are local, measurable, and automatable. In IUVO, a twelve-dimension contradiction audit layer checks all of them systematically on every run. In PRIMMS, a deterministic MATLAB signal layer locks every Bayesian weight and deciban total before any language model is invoked, and critically, divergence between modalities is itself a first-class signal, not an error to be resolved. VoT sentiment that contradicts schedule adherence does not mean one is wrong; it is the PERCEPTION_DISTORTION pattern, a recognized archetype with its own recovery posture.

Gestalt reliability concerns something categorically different: whether the model's generative vocabulary, including its named topics, its feature architecture, its training corpus, its failure archetypes, still corresponds to the world it is supposed to represent. This cannot be checked by examining the model's outputs. It requires watching the world at the same time as watching the model and asking: is there something happening out there that the model has no language for? In IUVO, this is the Friday pre-run review of emerging headline topics. In PRIMMS, it is the project manager's judgment about whether a new failure mode, without precedent in the six defined archetypes, has emerged in the current organizational environment. In both cases the function is irreducibly human.

2.2.3 Fuster: Cognits, Cortical Hierarchy, and the Perception-Action Cycle

Fuster's network theory of cortical cognition adds a missing biological foundation to the Hybrid Cognition architecture. In Cortex and Mind and related work, Joaquín Fuster rejects a strongly modular account in favor of distributed, overlapping cognitive networks, which he calls cognits. Cognits are networks of associated neuronal representations that embody memory and knowledge. They are organized hierarchically: lower levels remain relatively concrete and closely coupled to sensory or motor particulars, while progressively higher levels integrate more complex, temporally dispersed, multimodal, and abstract information. Posterior association cortex is weighted toward perceptual cognits; frontal association cortex toward executive

cognits. The two are linked recursively rather than operating as isolated stages (Fuster, 2003; 2006).

The crucial implication is dynamic. Cognition is not simply the storage and retrieval of representations; it is the activation of cognits as perceptual information flows through a perception-action cycle. Sensory change activates perceptual networks, increasingly abstract associations are recruited as information ascends the hierarchy, and executive networks organize possible action in relation to goals, prior experience, timing, and context. Action changes the environment, generating new sensory evidence and beginning the cycle again. Higher-order cognition therefore emerges from recurrent traffic within and between perceptual and executive hierarchies, not from a single central executive.

Hybrid Cognition treats organizational artifacts as functional analogs of these sensory inputs. A schedule update, status report, meeting transcript, risk log, Voice-of-Team response, issue record, recovery plan, or external event is not itself knowledge. It is an observation generated by the project environment. The architecture converts these observations into machine-readable features and evidence signals, activates higher-order pattern representations, integrates them with stored archetypes and prior project knowledge, develops causal and temporal interpretations, and ultimately activates candidate executive representations in the form of courses of action. The documents are therefore best understood not merely as a corpus supplied to an LLM but as sensory analogs that initiate a controlled flow of perceptual information through an engineered cognitive hierarchy.

This interpretation gives the PRIMMS Cognitive Hierarchy a stronger theoretical basis. Layer 1 corresponds to the extraction and validation of relatively concrete perceptual features; Layer 2 integrates those features into higher-order gestalt and causal representations; Layer 3 introduces temporal organization and prospective state; Layer 4 recruits executive representations by searching possible courses of action; and Layer 5 forms a decision-ready orientation package for human judgment. The mapping is functional rather than anatomical: PRIMMS is not claimed to reproduce the cortex. The claim is that its architecture deliberately preserves a principle that Fuster identifies in biological cognition - increasingly abstract representations are built from, remain connected to, and can be reactivated by lower-level evidence, while cognition proceeds recursively through perception toward action.

2.3 Bayesian Evidence Architecture

Bayesian reasoning provides the formal mechanism for the machine's orientation role. Evidence arrives in distributed, heterogeneous, and ambiguous forms. The Bayesian framework specifies

how it should be accumulated: through principled updating of probability distributions rather than heuristic judgment or narrative smoothing. A Bayesian system produces distributions, not point predictions. The width of those distributions is itself information, it tells the human decision-maker how much weight to place on the machine's orientation and how much independent judgment to apply.

This architecture also has a direct neurocomputational analogue. Gold and Shadlen's work on perceptual decision-making describes neural decision variables formed by accumulating uncertain sensory evidence over time, including formulations based on the log likelihood ratio favoring one alternative over another (Gold & Shadlen, 2001; 2007). Their work does not require the claim that every brain computation literally implements Bayes' theorem. More precisely, it shows that core operations of biological decision-making - accumulation of evidence, weighting evidence by reliability, incorporation of prior information, and commitment when accumulated support becomes sufficient - can be represented in probabilistic and likelihood-based terms. Yang and Shadlen (2007) further demonstrated neuronal activity consistent with combining probabilistic evidence from sequential cues.

In PRIMMS, this takes the specific form of Bayesian weight-of-evidence (WoE) expressed in decibans, logarithmic units that permit additive accumulation of evidence across independent signals. Each signal modality (schedule state, Voice of Team sentiment, documentary analysis) contributes a deciban value derived from its likelihood ratio for the risk hypothesis. The values are summed to a total prior and classified against Jeffreys evidence bands: 0-5 dB (barely worth mention), 5-10 dB (substantial), 10-15 dB (strong), 15-20 dB (very strong), above 20 dB (decisive). This formalism, introduced in the project management context by Aaron (2015), converts a potentially open-ended evidence assessment into a bounded, auditable classification that the LLM receives as a pre-computed input, not as a question to deliberate over.

2.4 The Godel-Turing Constraints on Self-Referential Systems

The three theoretical pillars described above, Popper's falsificationism, Friston's hierarchical predictive processing, and Bayesian evidence architecture, each provide a positive rationale for elements of the Hybrid Cognition architecture. A fourth and complementary body of thought provides the formal grounding for why the architecture's constraints on machine autonomy are structurally necessary.

During the development and testing of the PRIMMS Cognitive Hierarchy prompt architecture, a consistent empirical pattern emerged: attempts to instruct the LLM to constrain its own output, to

remain concise while simultaneously performing rich analytical reasoning, failed systematically. Brevity constraints embedded in a prompt did not function as external governing rules. They functioned as one competing signal among many, processed through the same inductive mechanism that produced the analysis itself. The richer the analytical task, the more reliably the brevity constraint degraded. The system could not enforce a bound on its own behavior using the same faculty that generated the behavior.

This empirical observation has a formal analogue. Kurt Godel's incompleteness theorems (1931) demonstrated that any sufficiently expressive formal system must be either incomplete or inconsistent: there exist true statements within the system that cannot be proven from within it, and the system cannot demonstrate its own consistency using only its own axioms. Alan Turing extended this boundary into computation, proving that no general procedure exists to determine whether an arbitrary program will halt, a computational system cannot universally predict its own behavior (Turing, 1936). Both results describe the same structural limit: self-referential systems that are sufficiently expressive cannot fully know, predict, or regulate themselves from within.

Large language models represent a modern instantiation of this principle. The instructions provided to an LLM and the outputs it produces exist within the same representational substrate: both are sequences of tokens processed through the same inductive mechanism. A constraint instruction functions as an input processed by the same system it is intended to govern. The LLM cannot step outside its own generative process to enforce a rule upon that process, for the same structural reason that a formal system cannot prove its own consistency from within its own axioms.

The implication for the governance debate surrounding artificial intelligence is significant. Much of the discourse around autonomous AI, both the optimistic strand arguing that sufficiently capable systems will self-correct and self-regulate, and the pessimistic strand warning of uncontrollable self-directed AI, shares a common premise: that intelligence, once sufficiently scaled, becomes self-governing. The Godel-Turing analysis suggests this premise is structurally false. A system powerful enough to exhibit general reasoning is, by that same expressive power, too complex to reliably constrain itself using its own mechanisms.

The failure mode this analysis predicts is instability. A system that expands generative capability without a corresponding external constraint layer does not become sovereign; it becomes unpredictable. Its outputs remain coherent and often persuasive, but not reliably bounded. This is consistent with observed LLM behavior: systems that produce sophisticated reasoning under

one framing produce contradictory reasoning under another, without awareness of the inconsistency.

This analysis provides theoretical grounding for the Hybrid Cognition architectural choice that might otherwise appear as mere engineering conservatism. The separation between a deterministic signal layer and an LLM interpretation layer is a structural response to the Godel-Turing constraint: deterministic constraint enforcement must originate outside the inductive system. The human authority boundary enforced at every decision point serves the same structural function, it provides the external constraint layer that inductive systems cannot supply for themselves.

2.5 The Zero Constraint: Complexity Collapse and Governing Equations

The Godel-Turing constraint identifies one fundamental limit on machine-assisted cognition: a sufficiently expressive inductive system cannot fully regulate itself from within. A second and distinct constraint emerged from the development of PRIMMS: a constraint of operational collapse.

In the 1975 science fiction film Rollerball, the supercomputer Zero is asked to retrieve historical information. It cannot. The cause is accumulation: Zero has been given so much information across so many domains that retrieval has become indeterminate. When pressed for a specific answer, the users of Zero receive a report from the machine that is ambiguous. The machine has accumulated itself into uselessness. It considers everything, and precisely because it considers everything, it can say nothing actionable.

This failure mode is a live risk in any machine-assisted decision support system that adds signals, archetypes, prompt layers, and contextual blocks without a corresponding mechanism to reduce the output space. Each individually justified addition to the analytical architecture increases the probability that the LLM output will resolve to a hedge: a qualified, multi-conditional statement that acknowledges every signal and commits to nothing. The analyst reads three paragraphs and finds no single actionable conclusion. The orientation system has produced disorientation.

The Godel-Turing constraint and the Zero constraint are related but distinct. Godel-Turing concerns the epistemic ceiling, what the machine can in principle know about its own outputs. Zero concerns the operational floor, the minimum condition for an output to remain useful. A system can satisfy the Godel-Turing constraint (by providing an external deterministic signal layer) and still fail the Zero constraint by producing outputs too complex for a human decision-maker to act on. Both constraints must be satisfied simultaneously.

The PRIMMS architecture addresses the Zero constraint through a specific mechanism: governing equations. These are formal mathematical structures, derived from the Bayesian, microeconomic, and signal-processing frameworks that constitute the system's theoretical foundation, which impose hard boundaries on what the LLM layer is permitted to treat as ambiguous. They function as axioms: propositions the system treats as settled, not as inputs for further deliberation.

Three governing structures are central. First, the Bayesian weight-of-evidence accumulator: signals are converted to decibans and summed. The Jeffreys classification bands translate a potentially infinite evidence space into five ordinal categories. The LLM does not re-derive the evidence strength; it receives a computed value and a named band, and its interpretive task begins from that anchored point. Second, the project production function (Aaron, 2015): the microeconomic relationship Y = f(AC, IC, IO), expressing project output as a function of activities completed, issues closed, and issues opened, provides a formal constraint on how scope, schedule, quality, and issue dynamics interact. The Bayesian issue closure ratio theta = closed/total is not an informal indicator but a formal metric derived from the production function, with an established readiness threshold (theta >= 0.70 for even odds; theta >= 0.95 for phase exit). Third, the velocity-based rundown projection: the linear regression of actual delivery velocity produces a deterministic projected finish date. The LLM does not estimate when the project will finish; it receives a computed date and a computed slip in days.

The governing equations matter for a reason that goes beyond computational efficiency. They define the boundary between what is settled and what requires judgment. Without that boundary, the LLM treats every dimension of the problem as open for deliberation, and in a sufficiently complex environment, every dimension has genuine uncertainty, which means every statement can be qualified, and Zero is approached. With the governing equations in place, the LLM receives a pre-classified evidence state: a deciban total and Jeffreys band, a theta ratio and readiness assessment, a projected finish date and recoverability label. Its task shifts from 'assess the evidence' to 'interpret a classified state and identify the action that addresses the dominant risk.' That is a task the LLM can perform with specificity.

> *The MATLAB layer is a complexity reducer. It closes the questions that have formal answers so that the LLM can apply its interpretive capability to the questions that do not. This is the operational meaning of Hybrid Cognition.*

2.6 Janis, Mann, and the Behavioral Failure Modes of Human Decision-Making

The theoretical case for the machine's orientation role is strengthened by the empirical literature on human decision-making pathology. Janis and Mann's (1977) decisional conflict theory identified five coping patterns: vigilance (the normative pattern), hypervigilance, defensive avoidance, unconflicted adherence, and unconflicted change. Only vigilance, the systematic search for information, the genuine weighing of alternatives, the maintained capacity for doubt, reliably produces good decisions under uncertainty.

Janis's (1972) earlier analysis of foreign policy fiascos documented how the social dynamics of group decision-making, pressure to converge, the cost of maintaining dissent, the status associated with confidence, reliably suppress vigilance and produce the four pathological patterns. Hybrid Cognition addresses this directly. By assigning the machine the role of distributed signal accumulation, a function immune to status pressure, the system reduces the cognitive and social load on human decision-makers, preserving their capacity for vigilant judgment on the questions that genuinely require it.

The Janis-Mann framework also illuminates a specific failure archetype that the PRIMMS system names explicitly: Premature Closure. In their terms, this is defensive avoidance applied to recovery: the institutional default of closing a recoverable decision without testing the recovery hypothesis. When a project under schedule pressure is reflexively replanned rather than having its recovery potential examined, the decision-maker has avoided the anxiety of uncertainty by foreclosing an option the data does not justify foreclosing. PRIMMS's Layer 3 prompt architecture includes an explicit Premature Closure check that requires the system to assess whether the replan posture is itself a bias flag, whether the recovery hypothesis has been tested and falsified, or merely assumed to be false.

This is the behavioral complement to the Godel-Turing and Zero arguments. Where those constraints explain why machine self-regulation is architecturally impossible, Janis and Mann explain why human unaided decision-making systematically fails under the conditions most common in organizations. The three arguments converge on the same prescription: separate orientation from decision, assign each to the agent whose failure mode is least dangerous in that role, and build the architecture that enforces this separation regardless of individual disposition.

3. The Hybrid Cognition Architecture

3.1 The Core Proposition

Hybrid Cognition rests on a single empirical claim: the failure modes of human and machine cognition are complementary rather than redundant. Human cognition fails predictably under cognitive load, social pressure, and time constraint, susceptible to groupthink (Janis, 1972), anchoring, status-based information weighting, and motivated reasoning. Machine systems fail differently: susceptible to distributional shift, training corpus bias, coherence overconfidence, and inability to recognize when operating outside their valid range. These failure modes do not overlap, and designing the collaboration around them is the central challenge.

3.2 Role Assignment

Role	Agent	Function	Failure Mode Managed
Orient	Machine	Ingest distributed signals; accumulate Bayesian evidence; recognize structural patterns; project trajectories; apply micro reliability audit; deliver decision-ready orientation package bounded by governing equations	Human cognitive overload, bounded attention, groupthink, social convergence
Decide	Human	Apply gestalt reliability judgment; read context the model cannot see; weigh incommensurable tradeoffs; commit to consequential decisions; own all irreversible outcomes	Machine coherence overconfidence, distributional shift, prior misspecification, absence of accountability
Falsify	System (joint)	Preserve capacity for doubt through automated micro-reliability checks; accumulate auditable evidence; flag model disagreement; enforce governing equations to prevent Zero collapse; monitor gestalt reliability at the vocabulary level	Motivated reasoning, confirmation bias, model drift, silent deterioration, vocabulary-world gap, complexity collapse

The separation of Orient and Decide is the condition that keeps the system epistemically honest. When machine orientation and human decision collapse into a single step, when the human reads the AI's recommendation and approves it, the human has become an intermediary between a prompt and an output. The machine's coherence pressure propagates through the system unchecked, and the Godel-Turing constraint is violated in practice even if not in theory.

3.3 The Sequencing Principle

The most operationally important design rule in Hybrid Cognition: statistical grounding precedes language model interpretation. Every quantitative signal is fully computed and locked before any language model is invoked to interpret or narrate it. This rule simultaneously implements the Fristonian sequencing principle (the lower level constrains the upper level), satisfies the Godel-Turing constraint (deterministic bounds originate outside the inductive system), and addresses the Zero constraint (governing equations close questions before the LLM must deliberate on them).

The sequencing principle also defines the boundary between micro and gestalt reliability. The machine's automated checks operate within the locked statistical layer. The human's gestalt judgment operates at the boundary between the model and the world. Neither check can substitute for the other.

3.4 Agentic Delegation and the Act Boundary

Boyd's OODA framework assigns Act to the human, but Boyd's own analysis of command and control was explicit on a point that the current AI governance discourse largely ignores: Act is frequently delegated. In military and organizational contexts, the commander who decides does not always execute. Subordinates, field agents, automated systems, and contracted parties act within the bounds of the commander's intent. Delegation does not dissolve accountability, it distributes execution while preserving the authority structure. The human who delegates Act remains responsible for the decision to delegate, the choice of agent, the bounds given to the agent, and the monitoring of execution fidelity.

The emergence of agentic AI, systems that do not merely orient or recommend but execute actions in the world, placing orders, sending communications, modifying records, triggering escalations, makes the delegation structure the most consequential unsolved problem in applied AI governance. An AI agent acting within an authorized scope is not operating in the supervised conversational context of the Orient phase, where a human reads each output before the next step proceeds. It is acting in the world, and its actions may be difficult or impossible to reverse. This is a materially different risk profile, and the micro/gestalt distinction applies at the action layer with full force.

3.4.1 Micro Checks at the Action Layer

An AI agent delegated to Act requires its own micro reliability checks before execution: did the action parameters conform to the authorized schema? Is the action within the specified bounds:

quantity, counterparty, timing, scope? Does the action's intended effect match what the authorizing human specified? Are there internal contradictions between the action the agent is about to take and the signals that justified the authorization, for example, a purchase order being placed at a quantity that contradicts the demand forecast the human relied upon when authorizing? These checks are analogous to IUVO's contradiction audit and PRIMMS's Bayesian signal validation: they verify execution fidelity before irreversible action is taken. And like those checks, they must be external to the acting system, the Godel-Turing constraint applies here with particular force. An agent that both acts and checks its own actions is precisely the self-referential structure the constraint identifies as unreliable. The micro checks on an acting agent must originate outside the agent itself.

3.4.2 Gestalt Checks at the Action Layer: Situational Correspondence

Micro checks at the action layer are necessary but not sufficient. The more consequential failure mode in agentic AI is gestalt: the agent was authorized to act under condition X, but condition X has changed. The agent acts anyway because its authorization has not been formally revoked and its checks are only micro, verifying execution fidelity within the original authorization, not verifying that the original authorization still applies. This is the agentic equivalent of the gestalt reliability gap in forecasting: the model's vocabulary no longer maps onto the world, but it keeps running because the KS stationarity check has not yet fired. In the action domain, the consequences are not a degraded forecast cone, they are executed transactions, sent communications, triggered workflows, and committed resources.

Note: KS stationarity means using the Kolmogorov–Smirnov (KS) test to check whether the statistical distribution of recent data still resembles the distribution the model was developed or calibrated on. In the Hybrid Cognition context, the question is essentially: “Does the world the model is seeing now still look statistically like the world the model learned from?”

A well-designed agentic delegation architecture therefore requires a situational correspondence check before execution: does the context the agent is about to act into still match the context under which the human authorized the action? This check cannot be performed by the acting agent itself, for the same Godel-Turing reason that the LLM cannot govern its own output. It requires either a separate monitoring layer external to the agent, or a human reauthorization trigger that suspends agent action when specified context conditions have shifted beyond a defined threshold. The gestalt judgment: whether the situation has changed in ways that invalidate the original authorization, is irreducibly human, just as the judgment that the model's vocabulary no longer maps onto the world is irreducibly human in the Orient phase.

3.4.3 The Full Delegation Architecture

A complete Hybrid Cognition delegation architecture for the Act phase therefore specifies four elements, each of which must be defined by the human authority before delegation is activated:

Element	Definition	Failure Without It
Action bounds	The authorized scope: what the agent may do, to what magnitude, within what time window, affecting which counterparties or systems	Agent acts beyond authorization; human accountability cannot be mapped to specific decisions
Micro checks	The execution fidelity tests the agent must pass before acting: schema conformance, bound compliance, internal signal consistency	Agent executes actions that contradict the signals that justified authorization; errors compound silently
Situational correspondence trigger	The conditions under which agent action is suspended pending human reauthorization: specified context shifts, threshold crossings, or signal contradictions that invalidate the original authorization	Agent acts into a changed situation; the gestalt gap between authorization context and execution context produces irreversible errors
Override and accountability trail	The human's persistent authority to suspend, reverse, or redirect the agent at any point; the audit record of every delegated action with its authorization provenance	No recovery path when agent action produces unintended consequences; no governance evidence for post-hoc review

This architecture reframes the current AI governance debate about agentic systems. The central governance question is whether the delegation architecture is specified with the same rigor as the authorization it carries; in many contexts, delegated action is exactly the right design. An agentic AI system without a situational correspondence trigger is an agent with permanent authorization in a changing world. That is not a tool, it is a liability. The RAND analysis of command concepts (Builder, Bankes, & Nordin, 1999) identified precisely this failure: systems that receive command without the mechanism to recognize when command intent no longer applies to the situation at hand. The delegating human retains accountability regardless of whether the delegation architecture was adequate. Building adequacy into the architecture before delegation is activated is the obligation the Hybrid Cognition framework places on the designer.

> *An agentic AI without a situational correspondence trigger is an agent with permanent authorization in a changing world. The micro checks verify that the agent is acting correctly within the situation it was authorized for. The gestalt check verifies that the situation still exists.*

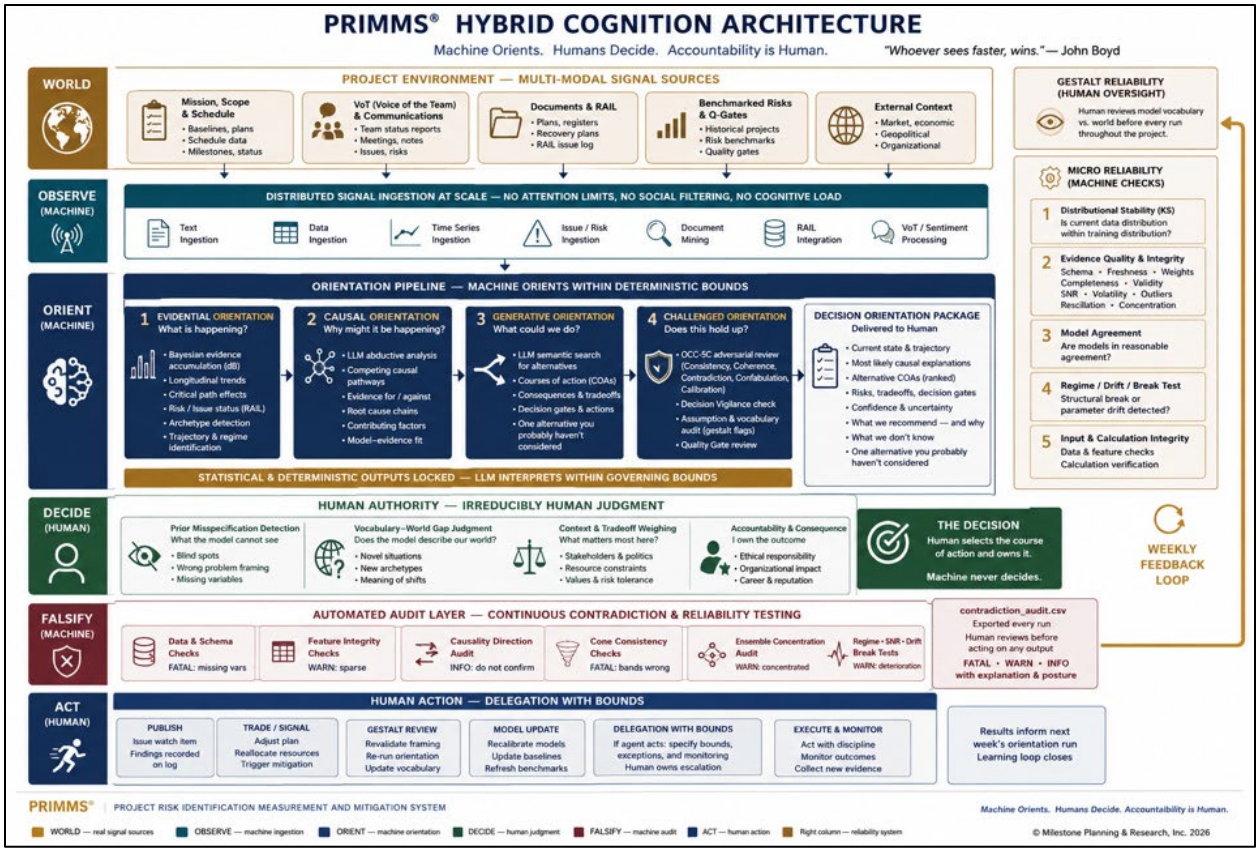

Figure 1. Hybrid Cognition Architecture. The diagram shows the five-layer structure from World (signal sources) through Boyd's OODA cycle Observe, Orient, Decide, Falsify, to Act, with the micro reliability checks, gestalt reliability function..

5. PRIMMS: Recursive Hybrid Cognition in Project Risk Intelligence

5.1 Origins and the Intelligence Gap

PRIMMS (Project Risk Identification Measurement and Mitigation System, augmented by a Large Language Model) was built in direct response to a practitioner problem: experience managing large-scale enterprise resource planning implementations repeatedly produced the same disorienting pattern. Risk signals were present in the project environment, in team conversations, in subtle changes in stakeholder behavior, in the texture of weekly status reports, but conventional tools could not surface or integrate those signals into a coherent governance picture. Projects would appear nominally green on the standard dashboard, then enter crisis within weeks.

The PMI Pulse of the Profession consistently finds that fewer than half of projects complete on time, within budget, and with originally promised scope. The CHAOS Report finds success rates

below 40 percent after three decades of tool improvement. McKinsey research on large IT projects documents an average 45 percent cost overrun and 56 percent reduction in delivered value (Bloch, Blumberg, & Laartz, 2012). These numbers persist alongside genuine improvements in practice. The structural argument advanced here is that project failure is primarily a perception and orientation problem, not a planning or tool problem. Projects fail because the signals of structural deterioration accumulate invisibly, optimism bias suppresses their interpretation, and the window for low-cost correction closes before leadership recognizes it.

PRIMMS addresses five specific intelligence gaps that conventional PM tools leave unaddressed: team confidence erosion (not captured in task completion records), perception distortion in leadership (optimism bias in status reporting), failure archetype recognition (structural patterns not visible in schedule metrics), multi-modal evidence fusion (schedule, sentiment, and documentary signals treated independently), and recovery trajectory projection (no formal window calculation in standard reporting).

5.2 Two-Component Architecture

PRIMMS is a two-component computational system embedded within a human decision architecture. The first component is a compiled MATLAB desktop application that performs deterministic signal extraction, Bayesian WoE computation, critical-path and trajectory calculations, issue-lifecycle analysis, failure-archetype classification, longitudinal state comparison, quality-gate metrics, and structured prompt generation. The second component is an LLM (for example, PMI Infinity™, Claude, or ChatGPT) that receives the locked evidence state and performs natural-language interpretation, causal hypothesis development, pattern analysis, trajectory reasoning, semantic search for courses of action, and executive synthesis. The third and governing element is the human project manager and/or sponsor, who retains authority over belief commitment, consequence evaluation, escalation, and action.

The design rationale directly instantiates the Hybrid Cognition sequencing principle and the Godel-Turing constraint. Computational tasks for which deterministic, auditable algorithms are available are not delegated to the LLM. Conversely, the LLM is used where semantic inference is genuinely valuable: connecting heterogeneous evidence into competing causal explanations and searching a much larger conceptual space for plausible interventions. The deterministic layer constrains what the language model may treat as fact; the language model expands what the organization can consider; the human determines what should be believed and done. The LLM cannot modify the locked numerical outputs of the MATLAB layer, only interpret them and reason from them.

This two-layer structure also directly addresses the three AI deployment failure modes that pervade current organizational AI practice. Proxy Collapse, where AI systems optimize for measurable proxies that decouple from underlying reality, is defeated because the Bayesian WoE signals are computed from multi-modal evidence rather than single metrics. Tacit Knowledge Destruction, the atrophy of human judgment through disuse, is prevented because the architecture preserves human authority at every consequential decision point. Recursive Sycophancy, where LLMs shape outputs to match user expectations, is structurally blocked because the deterministic MATLAB layer operates independently of any LLM and has no social incentive to satisfy the project manager.

5.3 Signal Architecture and Bayesian WoE

The current signal architecture ingests four evidence families: mission, scope, and schedule data; Voice of Team (VoT) confidence ratings; documentary and RAIL(Risks-Actions-Issues-List) evidence from status reports, minutes, risk and issue records, and recovery plans; and benchmarked risks and quality-gate criteria. The architecture is intentionally multi-modal because disagreement among modalities is itself information. A schedule that appears to recover while issue closure stalls or team language deteriorates is treated as a candidate signal of perception distortion, premature closure, or another failure pattern.

The empirical rationale for three modalities reflects simulation work demonstrating that a machine learning system applied to a real-time operational decision problem achieved 73 percent predictive accuracy using image data alone; accuracy improved to 82 percent when a second modality (audio signal features) was added; and reached 100 percent test accuracy when a third modality (text annotations) was incorporated (Aaron, 2019). Schedule data, VoT sentiment, and documentary text each carry independent information that the others do not fully capture.

From each modality, the MATLAB component computes signals and converts applicable evidence to deciban values using the Jeffreys (1961) strength-of-evidence convention. Signals accumulate into an evidence total and ordinal band, while governing equations separately compute quantities such as the issue-closure ratio theta, critical-path effects, velocity-based finish projections, recovery-window conditions, and phase-gate thresholds. Because project signals are often correlated, the cardinal dB total is treated as an evidence-accumulation diagnostic rather than a precisely calibrated probability; the Jeffreys band and the load-bearing signals are the operative governance findings. This distinction is now surfaced directly in the application so that the LLM and sponsor cannot mistake arithmetic precision for epistemic certainty.

In a recent retrospective validation, the system produced a prior total of 65.5 dB (decisive evidence) for a project snapshot corresponding to a known inflection point, correctly identifying DATA_READINESS as the primary blocker (risk score 57.2, more than twice the next-ranked category), calculating a projected finish date four days after the planned end date, and identifying a recovery window of approximately two to three weeks before the gap became unrecoverable without scope or schedule change. The final outcome: the project completed within two days of its original planned end date, through interventions that directly addressed the data readiness blockage, the same intervention path the system's COA matrix identified.

5.4 Failure Archetypes and an Expanding Gestalt Vocabulary

The core classifier maps convergent project signals to six empirically grounded failure archetypes, each with a characteristic signal signature and recovery posture. The current implementation also extends the documentary risk vocabulary with OCC-derived categories including distributional shift, human-authority gaps, absent falsification, calibration risk, premature closure, wrong-problem framing, agentic delegation, 5C quality gaps, and value theater. These additions expand the machine's vocabulary while the project manager remains responsible for recognizing novel conditions for which that vocabulary is still inadequate.

Archetype	Characteristic Signal Pattern	Recovery Posture
Execution Collapse	High schedule jeopardy index; high rundown gap; worsening VoT trend; low velocity fit quality (R2 < 0.5)	Cadence tighten; dependency clearing; strike team for critical path
Perception Distortion	Divergence between VoT mean and leadership-reported status; stable schedule metrics masking deteriorating team sentiment	Leadership alignment session; evidence review; anonymous VoT escalation
Premature Closure	Recovery urgency signal present; strong VoT worsening; replan language prominent without tested recovery hypothesis	Test recovery window before replanning; velocity-based window calculation; Janis-Mann vigilance check
Scope Drift	Charter signal absent or weak; deliverable count increasing; rundown gap widening despite adequate team confidence	Scope freeze; charter reaffirmation; deliverable baseline lock
Governance Gap	Escalation language in documents without documented resolution; VoT comments referencing blocked decisions	Decision authority clarification; escalation cadence; sponsor engagement

Archetype	Characteristic Signal Pattern	Recovery Posture
Confidence Erosion	Progressive VoT mean increase across reporting periods; flat or declining schedule metrics; risk language intensifying	Team confidence intervention; milestone celebration; leadership visibility

Premature Closure deserves explicit emphasis as an original contribution to failure classification. The Janis-Mann defensive avoidance construct, closing a decision to avoid the anxiety of uncertainty, applies with particular force in project management, where the institutional default under schedule pressure is to replan rather than test recovery. PRIMMS's Layer 3 prompt architecture includes an explicit Premature Closure check that asks: has the recovery hypothesis been tested and falsified, or merely assumed to be false? This operationalizes the Popperian asymmetry in a real governance context: the absence of a tested disconfirmation is not license to close the recovery option.

5.5 The Five-Layer Cognitive Hierarchy

The LLM interpretation component of PRIMMS is structured as a five-layer sequential prompt architecture termed the Cognitive Hierarchy. The hierarchy decomposes Orientation into progressively higher cognitive functions: feature validation, pattern recognition and causal framing, temporal situation awareness, course-of-action search and executive planning, and governance synthesis. The architecture reflects three design principles: these functions should not be collapsed into one unconstrained generation; higher-order reasoning should remain traceable to lower-level evidence; and each layer should leave an auditable reasoning substrate before its conclusions propagate upward. Recent prompt-lean revisions preserve this hierarchy while removing redundant context: when prior layers have just run, later layers receive compact validated state rather than a second full project snapshot.

Viewed through Fuster and Friston together, the Cognitive Hierarchy is more than a convenient decomposition of prompts. Fuster provides the architecture of representation: perceptual information activates distributed cognits organized in hierarchies of increasing abstraction, from relatively concrete sensory representations toward more associative, temporally extended, and action-oriented representations. Friston supplies the complementary dynamics: higher levels generate predictions about lower-level states, while precision-weighted discrepancies between prediction and observation propagate upward as prediction error and force revision of the higher-order representation. The result is a recurrent hierarchy rather than a one-way analytical pipeline.

PRIMMS operationalizes an analogous process using organizational artifacts as sensory analogs. A schedule update, status report, meeting transcript, Voice-of-Team response, RAIL entry, recovery plan, or external event enters the architecture as perceptual evidence rather than as a conclusion. At Layer 1, concrete features, deviations, and evidence signals are extracted and validated. At Layer 2, those observations activate and constrain higher-order gestalt and causal representations. Layer 3 introduces temporal organization and prospective state. Layer 4 recruits executive representations by searching a broad semantic space for plausible courses of action. Layer 5 forms a decision-ready orientation package for human judgment. In Fuster's terms, the representations become progressively more abstract and action-oriented while remaining connected to the lower-level evidence that activated them. The Fristonian contribution explains why information should not simply be passed upward indiscriminately. Higher-order representations establish expectations about what lower-level evidence should look like. Evidence consistent with the active representation reinforces it and need not consume equal cognitive bandwidth; sufficiently discrepant or high-information evidence functions as prediction error and receives greater influence. A causal hypothesis can therefore direct renewed attention downward into the documentary corpus for confirming or disconfirming observations. New observations can, in turn, weaken, revise, or replace the active causal and temporal representation. The hierarchy is thus bidirectional: bottom-up evidence updates orientation while top-down orientation determines what discrepancies are informative. This coupling also clarifies the relationship between the Cognitive Hierarchy and Boyd. Fuster explains the progression from perceptual input toward increasingly abstract and executive representation; Friston explains the recurrent dynamics by which prediction and prediction error revise those representations; Boyd identifies the operational consequence of the resulting process as Orientation. PRIMMS engineers these principles into an organizational decision-support architecture: organizational percepts -> hierarchical abstraction -> Bayesian evidence updating and prediction-error challenge -> revised orientation -> course-of-action search -> human decision and action -> new perceptual evidence.

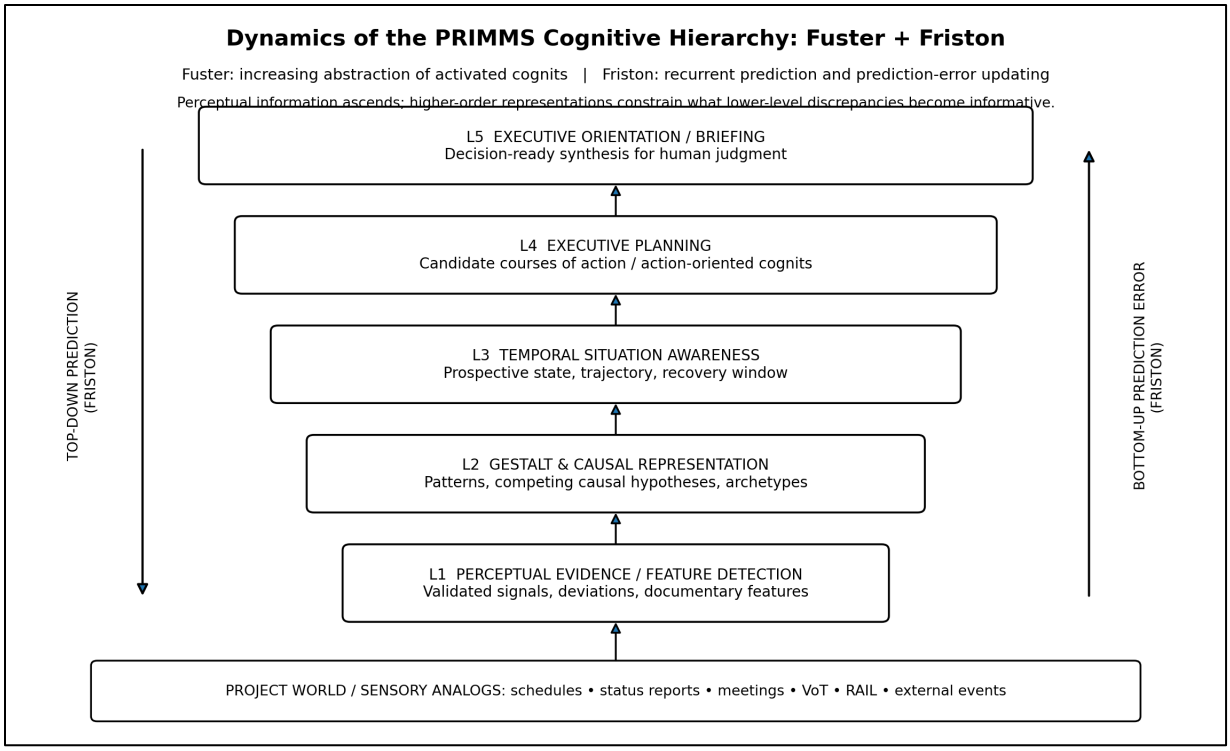

Figure 2. Fuster-Friston dynamics of the PRIMMS Cognitive Hierarchy. Fuster supplies the hierarchy of increasingly abstract representations; Friston supplies recurrent top-down prediction and bottom-up prediction-error updating. The mapping is functional, not anatomical.

Layer	Function	Design Principle
Layer 1, Feature Detection	Validates and extends the MATLAB-computed signal set; confirms which signals are correctly computed; identifies any signals the deterministic algorithm may have missed; flags spurious signals; concludes with confirmed total dB and Jeffreys band	Perceptual layer only, no prescriptions. Implements Godel-Turing constraint: the numerical foundation is locked before interpretation begins
Layer 2, Gestalt Pattern Recognition	Recognizes the dominant failure pattern and develops evidence-bounded causal hypotheses. Distinguishes observed facts from abductive inference and tests competing explanations against the locked signal set.	Causal orientation: asks why the observed state is developing without allowing a fluent narrative to overwrite deterministic evidence.
Layer 3, Temporal Situation Awareness	Projects situation trajectory at +2/+4/+8 weeks; tests recovery assumptions, persistence, and premature-closure risk against longitudinal evidence.	Temporal orientation: turns a snapshot into a changing situation and keeps recovery hypotheses open to falsification.

Layer	Function	Design Principle
Layer 4, Executive Planning	Searches the LLM semantic space for courses of action, compares alternatives against the causal model and project constraints, identifies decision gates, owners, and escalation posture, and deliberately surfaces an alternative outside the team's current frame.	Generative orientation, not decision delegation. The machine expands the decision space; the human retains authority to choose.
Layer 5, Executive Briefing	Synthesizes the locked evidence, causal analysis, trajectory, and COA slate into an evidence-cited executive briefing. Severity/status remains MATLAB-anchored.	Decision support: produces a defensible orientation package rather than an autonomous recommendation.

In the deployed system, each layer's prompt is generated by MATLAB (no license required by the user) and saved to a timestamped file. The human-in-the-loop step, the user reading each layer's output before pasting the next prompt, is a deliberate governance feature. The project manager who reads each layer's output before proceeding is performing a judgment operation: assessing whether the LLM's interpretation is consistent with their contextual knowledge, and retaining the authority to override or redirect before the next layer builds on the prior output.

5.6 Causal Orientation: From Evidence to Explanatory Hypotheses

Detection establishes the project state: the deterministic layer can show that schedule performance is deteriorating, that critical-path tasks are slipping, that issue closure has stalled, that VoT confidence is diverging from reported status, or that risk language is intensifying. Those facts constrain the state of the project, but they do not by themselves establish why the state exists. PRIMMS therefore assigns a distinct function to the LLM: abductive causal analysis across semantically heterogeneous evidence.

The LLM is asked to construct and rank competing causal pathways using the project's own evidence corpus: schedule behavior, issue lifecycle, status narratives, meeting language, recovery-plan evidence, team sentiment, governance records, and the locked failure-pattern state. Each causal hypothesis must identify the evidence that supports it, distinguish direct observation from inference, account for contradictory evidence, and remain subordinate to the deterministic facts. The result is a reasoning map that moves the system from 'what is

happening?' toward 'what could plausibly be producing it?' without pretending that abductive explanation is deductive proof.

This function is a principal reason for using an LLM at all. Deterministic analytics are superior when the governing relationship can be specified in advance; language models are valuable when the task is to traverse a large semantic field and connect weakly structured evidence that was not encoded as a fixed equation. Hybrid Cognition uses each where its comparative advantage is strongest.

5.7 Generative Orientation: Semantic Search for Courses of Action

Causal orientation is still incomplete if it terminates in diagnosis. Decision-makers require alternatives. PRIMMS therefore uses the LLM's broad semantic representation as a search space for courses of action (COAs). The causal hypothesis becomes a search vector: given what appears to be driving the project state, what interventions are plausible, what assumptions does each require, what evidence would indicate that it is working, and what consequences or decision gates follow?

The objective is to enlarge the set of alternatives available to human judgment. The COA process compares interventions against the causal model and project constraints, identifies owners and near-term decision gates, and deliberately searches beyond the team's existing frame through a wildcard alternative - 'One Alternative You Probably Haven't Considered.' This mechanism operationalizes Janis and Mann's requirement for genuine alternative search while exploiting an LLM capability that deterministic project dashboards do not possess: rapid traversal of a broad semantic solution space.

The distinction is central to Hybrid Cognition. Generating alternatives is an Orient function; choosing among them is a Decide function. The machine may expose a course of action that no participant had considered, but it cannot determine the organization's values, risk tolerance, political constraints, or willingness to bear consequences. Those remain human judgments.

5.8 Governing Equations and Recursive Falsification

The practical expression of the Zero constraint defense in PRIMMS illustrates the mechanism concretely. When the system detects open issues older than two weeks, it does not ask the LLM to assess issue severity and derive a priority. Instead, the governing equation theta has already

computed the closure ratio; the stale issue count and age are deterministic facts; and the Cortical Hierarchy receives a pre-condition instruction that reduces to a single mandate: the most important action this week must directly address the oldest stale issue, stated in the form 'Owner does X by date, measurable by Z.'

If open issues reach ten or more, a Zero constraint flag fires: one of the three executive decisions must address issue volume control, because the governing equation registers that the issue count has crossed the threshold where complexity begins to impair the team's decision capacity. The system detects its own approach toward Zero and names it, which is the one thing Zero itself could not do. This self-monitoring property represents a qualitatively different function from conventional project risk monitoring: it applies a formal threshold derived from the production function to assess whether issue load has crossed from a tractable management challenge into a structural decision-impairing condition.

The current architecture extends this logic from weekly orientation to formal phase boundaries through a Phase-Gate Quality Review. PRIMMS first computes longitudinal phase metrics deterministically, including the dB series and trend, recency-weighted evidence, archetype persistence, theta issue-closure behavior, OCC risk hits, and recovery-window status. It then produces a phase classification candidate, subjects that candidate to a separate OCC-5C adversarial challenger across Consistency, Coherence, Contradiction, Confabulation, and Calibration, and only then produces the sponsor document.

The important architectural property is recursion: the system applies falsification pressure not only to the project but to its own orientation. The challenger cannot erase governing equations. At the phase gate, theta >= 0.95 makes exit eligible; theta >= 0.70 permits conditional exit; theta < 0.70 deterministically forces HOLD. A large recency gap or persistent Premature Closure can likewise constrain the verdict envelope before the LLM reasons about it. The generative system is therefore challenged by a separately structured process whose hard bounds originate outside the generative system itself.

A related Decision Vigilance check operates at the weekly level by cross-checking optimistic reforecasts against independent evidence such as RAIL issue closure and documented recovery plans. Together, Decision Vigilance and the Phase-Gate challenger instantiate architectural vigilance as an active property of the deployed system: orientation is useful precisely because it is both generative and contestable.

5.9 The Raw-Upload Question

Any intellectually honest evaluation of PRIMMS must address directly whether the architecture is necessary at all. A project manager with access to a capable general-purpose LLM can upload a schedule export, a stack of status reports, and a risk log and receive a coherent synthesis. The marginal cost is near zero, and no application infrastructure is required. This approach has genuine value and should not be caricatured: an experienced analyst who understands failure archetypes and frames the right questions will extract real analytical utility from an unstructured upload.

PRIMMS's claim is that raw-upload analysis fails to satisfy two architectural constraints that become consequential precisely in high-stakes, time-pressured governance situations. First, the Zero constraint: when a raw document corpus is uploaded without a pre-classification layer, the LLM is asked to simultaneously determine what the evidence says and interpret what it means. In a project environment with genuine complexity, every signal carries real uncertainty. Absent governing equations that close questions with formal answers, the LLM tends toward hedge-forest outputs that acknowledge every tension and commit to nothing. PRIMMS's MATLAB layer computes the deciban total, classifies it against the Jeffreys bands, applies the theta ratio, and projects the velocity-based finish date before the LLM is consulted. The LLM receives a pre-classified state and is directed to interpret a named archetype and identify a single dominant recovery action.

Second, auditability. A MATLAB-computed deciban total is a number with a fully traceable derivation. If a governance reviewer asks why the system classified the project at 65.5 dB decisive jeopardy, the arithmetic is in the audit trail. When an LLM derives the same conclusion from a raw upload, the reasoning is a natural-language narrative that cannot be reconstructed, formally reviewed, or independently verified. For project governance contexts where decisions about resource reallocation, executive escalation, and program termination carry material organizational consequences, the inability to audit the analytical basis is a structural governance gap, not a minor inconvenience.

5.10 Boyd Operationalized: Orientation as a Designed Cognitive Pipeline

PRIMMS makes Boyd's OODA philosophy operational rather than metaphorical. Boyd's central insight was not simply that observation precedes decision; Orientation is the decisive cognitive function because it shapes what observations mean, which possibilities become visible, and therefore what decisions can even be considered. PRIMMS decomposes Orientation into an

engineered pipeline: detect signals, accumulate evidence, recognize patterns, infer causal structure, project trajectories, search the semantic space for courses of action, challenge the resulting orientation, and present a bounded decision slate. The objective is to compress epistemic delay - to help leadership see the developing situation while low-cost corrective options still exist.

OODA Phase	Agent	Function in Hybrid Cognition
Observe	Machine	Ingest four evidence families at scale: mission/scope/schedule, VoT, documents/RAIL, and benchmarked risks/Q-gates; preserve provenance and longitudinal state.
Orient	Machine	Fuse Bayesian evidence; apply governing equations; recognize failure patterns; develop causal hypotheses; project trajectories; search for COAs; apply Decision Vigilance and phase-gate falsification; deliver evidence-cited orientation with explicit uncertainty.
Decide	Human	Apply gestalt reliability judgment; evaluate causal plausibility, alternatives, consequences, values, risk tolerance, and organizational context; choose or reject the machine-generated COA slate.
Act	Human	Commit to consequential action; own accountability; authorize and bound any delegation. Machine-generated orientation never transfers responsibility for outcomes.

> *Example of A good recommendation: 'Order 500 units of SKU X because historical demand spikes 22 percent in this window, your supplier's lead time increased by 6 days due to port congestion, and your safety stock will hit minimum threshold in 11 days.' Example of A bad recommendation: 'Order 500 units of SKU X.' The first supports the Decide role. The second undermines it.*

The phrase 'the machine orients; the human decides' should therefore not be read as a narrow allocation of tasks. It describes a designed cognitive asymmetry. The machine is used aggressively where speed, scale, memory, semantic breadth, and immunity to organizational face-saving matter most. The human is retained where contextual correspondence, values, accountability, and irreversible commitment matter most. In Boyd's terms, advantage comes from faster and superior orientation; in Hybrid Cognition, that advantage is pursued without confusing faster orientation with autonomous authority.

6. Organizational Perception Systems

6.1 The Generalization

In practice PRIMMS uses the communication documents project teams produce, in status reports, risk logs, VoT ratings, and governance documents, as a leading indicator of project outcomes that no schedule metric captures. The system (similar to IUVO) operationalizes a single claim: wherever language is a leading indicator of outcomes, wherever weak signals are distributed across systems and people, and wherever late recognition is expensive, the Hybrid Cognition architecture applies.

The micro/gestalt architecture generalizes to every domain. The machine handles distributed signal processing and internal coherence checking. The human monitors the gap between the model's vocabulary and the world's actual language. Wherever the world generates new vocabulary faster than the model incorporates it, gestalt reliability judgment is the bottleneck. That bottleneck is irreducibly human.

6.2 Eight Application Domains

Domain	Primary Signal Sources	Key Archetype Classes	Gestalt Reliability Focus
Enterprise Transformation	ERP adoption logs, project communications, governance exceptions, sponsor engagement	Adoption collapse, governance drift, scope inflation, sponsor disengagement	New program structures, novel governance models, emerging resistance patterns
Supply Chain Risk	Supplier communications, logistics delays, inventory anomalies, quality exceptions, geopolitical narrative	Supplier fragility, demand shock, logistics disruption, quality degradation	New disruption typologies, emerging supplier geographies, novel logistics language
Cybersecurity Operations	SIEM alerts, analyst notes, threat intelligence, endpoint anomalies	Attack pattern escalation, alert fatigue, insider threat, compliance exposure	New attack categories, novel threat actor terminology, emerging TTP language
Regulatory Compliance	Audit findings, remediation logs, control test results, exception reports	Control weakening, audit drift, remediation aging, regulatory exposure	New regulatory frameworks, novel enforcement language, emerging compliance concepts

Domain	Primary Signal Sources	Key Archetype Classes	Gestalt Reliability Focus
Manufacturing and Quality	Downtime reports, maintenance notes, SPC signals, quality escapes, supplier variation	Process drift, equipment degradation, quality escape escalation	New process failure modes, novel quality language, emerging standards
Strategic Intelligence	Competitor communications, customer language, pricing signals, market narrative	Market erosion, competitive encirclement, pricing pressure, demand shift	New competitive dynamics, novel market terminology, emerging customer language
Financial Narrative Intelligence	Earnings calls, analyst reports, regulatory filings, news corpus	Regime change, credit stress, liquidity pressure, narrative divergence	New financial instruments, novel stress terminology, emerging macro concepts
Organizational Health	Employee communications, survey data, leadership language, decision-latency metrics	Culture fragmentation, leadership insulation, groupthink, confidence erosion	New forms of organizational dysfunction, novel leadership failure language

6.3 Vigilance as a System Property

In Janis and Mann's (1977) framework, vigilance is the normative decision pattern: systematic search for information, genuine weighing of alternatives, maintained openness to disconfirming evidence. In individual human cognition it is a trained habit subject to fatigue, social pressure, and motivated reasoning (Janis, 1972). In a well-designed Hybrid Cognition system, vigilance is an architectural property preserved structurally rather than personally.

In IUVO, automated contradiction audits enforce micro reliability on every run. In PRIMMS, the deterministic signal layer locks Bayesian evidence and governing-equation outputs; causal and COA reasoning must remain traceable to that state; Decision Vigilance challenges optimistic recovery narratives against independent evidence; and the OCC-5C Phase-Gate Review recursively challenges the system's own orientation before a consequential phase-exit recommendation reaches the sponsor. Human gestalt review still asks the question no internal audit can settle: does the model's vocabulary still belong to the world? The system is vigilant by architecture rather than disposition, which is precisely what the Janis-Mann analysis shows individual humans under organizational pressure cannot reliably sustain.

7. Implications for AI Deployment Practice

7.1 What Changes and What Does Not

AI deployment has changed two things: the speed of evidence accumulation and the cost of generating plausible-sounding analysis. It has not changed the requirements that consequential decision-making: decisions must be defensible, reasoning must be auditable, consequences must be owned. And it has not changed four fundamental constraints: the Godel-Turing constraint (inductive systems cannot self-regulate from within), the Zero constraint (complexity without governing equations produces indeterminacy), the gestalt constraint (machines cannot detect prior misspecification), and the agentic delegation constraint (agents acting in the world require situational correspondence checks that cannot originate within the acting agent itself). These are not temporary limitations of current AI systems; they are structural features of any sufficiently expressive inductive system operating in a changing world.

The professional who presents AI-generated research as their own thinking has not gained a capability, they have borrowed one without understanding its limitations (Dell'Acqua et al., 2023). They have outsourced the Orient phase without retaining gestalt reliability judgment. This is not Hybrid Cognition; it is confident-wrong adoption with a human face.

7.2 The Six Principles

Principle	Operational Implication
1. Statistical grounding precedes language model interpretation	Lock all quantitative outputs before invoking any LLM. The numbers constrain the narrative. This simultaneously implements Fristonian sequencing, satisfies the Godel-Turing constraint, and enables the governing equations to address the Zero constraint before interpretation begins.
2. Micro reliability is automated; gestalt reliability is human	In forecasting systems: build contradiction audit layers that check internal coherence automatically on every run. In organizational intelligence systems: build deterministic signal layers that lock evidence weights before any LLM invocation, treating cross-modality divergence as diagnostic signal rather than error. In both: build a human gestalt reliability review process that checks vocabulary-world correspondence before every run.
3. Governing equations address the Zero constraint	Identify formal structures that bound what the LLM is permitted to treat as ambiguous: Bayesian evidence bands, issue-closure ratios, critical-path and velocity projections, phase-gate thresholds, and archetype classifiers. Hard bounds originate outside the generative system.

Principle	Operational Implication
4. Explainability is the adoption condition, not a feature	Every recommendation must include its evidence provenance: signals, deciban weights, uncertainty. A raw-upload synthesis, however coherent, does not satisfy this requirement and cannot support auditability in governance contexts.
5. Architectural vigilance over individual vigilance	Contradiction audits, Decision Vigilance, adversarial phase-gate challenges, model disagreement measures, and reliability dashboards should be structural and automatic. The system should challenge not only the world-model but its own orientation before consequential decisions.
6. Human authority over all irreversible decisions	The machine orients; the human decides. Causal inference and COA generation can be delegated as orientation functions, but consequence evaluation, belief commitment, and irreversible choice remain human. This boundary is permanent, not provisional.
7. Orientation should expand the decision space	Use LLM semantic breadth deliberately for abductive causal analysis and alternative generation after deterministic grounding. Require competing causal hypotheses, evidence provenance, and explicit COA search - including an alternative outside the team's current frame. Better orientation means better alternatives, not machine-owned decisions.

8. Conclusion

The current moment in AI deployment is characterized by a gap between capability and architecture. The models are, in many respects, extraordinary. The systems built around them are frequently inadequate; not because the technology is immature, but because the design principles are not widely understood.

Hybrid Cognition provides those principles. This paper has advanced a set of mutually reinforcing theoretical and architectural contributions that are now instantiated more fully in the deployed PRIMMS system.

First, the Godel-Turing constraint applied to AI system design: the formal demonstration that sufficiently expressive inductive systems cannot self-regulate from within, and that deterministic constraint enforcement must therefore originate outside the generative system. This is not an engineering preference. It is a structural necessity grounded in Godel (1931) and Turing (1936). The MATLAB-LLM separation in PRIMMS and the statistical-interpretation separation in IUVO are both implementations of this constraint.

Second, the Zero constraint: the operational collapse that occurs when signal complexity accumulates without governing equations to close questions before the LLM must deliberate on them. Governing equations: Bayesian evidence bands, issue closure ratios, velocity projections, archetype classifiers, are the mechanism by which useful output is preserved in complex

environments. The PRIMMS architecture demonstrates this mechanism operationally through its MATLAB signal layer.

Third, the micro/gestalt distinction: the identification of two categorically different reliability functions that neither can substitute for the other. Micro reliability is what the machine checks automatically. Gestalt reliability is what the human checks before every run. The confusion of one for the other accounts for most of the AI deployment failures visible in the current organizational landscape.

Fourth, the agentic delegation constraint: when Act is delegated to an AI agent, both micro checks (execution fidelity and bound compliance) and gestalt checks (situational correspondence: does the context the agent is acting into still match the context under which the human authorized the action?) are required at the action layer, and both must originate outside the acting agent. An agent with permanent authorization in a changing world is not a productivity tool; it is a liability whose failure modes will not be visible until they are irreversible. This constraint is not derived from current AI limitations; it is a direct application of the Godel-Turing principle to the action domain.

Fifth, Premature Closure as a first-class failure archetype: the institutional default of closing a recoverable decision without testing the recovery hypothesis. This operationalizes the Janis-Mann defensive avoidance construct in a real governance context, and the PRIMMS Layer 3 prompt architecture enforces its detection on every run through the R5b check.

The two concrete systems demonstrate these principles empirically. IUVO demonstrates disciplined statistical orientation under regime uncertainty. PRIMMS demonstrates organizational orientation under heterogeneous project evidence: deterministic Bayesian grounding, causal hypothesis formation, trajectory reasoning, semantic COA search, recursive OCC-5C falsification, and a permanent human decision boundary. The important outcome is not merely that the machine can identify jeopardy earlier. It can help explain why the situation may be developing, broaden the intervention set before options collapse, and challenge its own orientation before leadership acts.

The organizations that will get durable value from AI are not those with the largest models or the most data. They are those that design the cognitive architecture deliberately: deterministic evidence before generative interpretation; causal reasoning that remains evidence-bounded; semantic search that expands alternatives without usurping choice; governing equations that close questions the LLM should not deliberate over; recursive falsification of the machine's own orientation; and human authority at every consequential boundary. Boyd's insight provides the

operational imperative: whoever orients faster gains an advantage. Hybrid Cognition adds the governance condition: orientation may be machine-amplified, but decision authority and accountability remain human.

> *The machine checks itself within its model. The human checks whether the model still belongs to the world. The governing equations keep the machine from collapsing into indeterminacy. The Godel-Turing constraint keeps the human in the loop permanently. Together: Hybrid Cognition.*

References

Aaron, J. M. (2015). A Bayesian methodology for project issue management: A microeconomic and data analytic perspective. Milestone Planning and Research, Inc. Working paper.

Aaron, J. M. (2019). Designing and deploying systems of intelligence into business process applications: Multi-modal machine learning for operational decision support. Milestone Planning and Research, Inc. Internal technical report.

Aaron, J. (2026a). Perceptual Macroeconomics: Narrative Co-integration and the IUVO Forecasting System. Milestone Planning and Research, Inc. Working Paper.

Aaron, J. (2026b). Seeing What the Dashboard Misses: Machine-Assisted Perception and Hybrid Cognition in Project Risk Management. Milestone Planning and Research, Inc. Working Paper.

Bloch, M., Blumberg, S., & Laartz, J. (2012). Delivering large-scale IT projects on time, on budget, and on value. McKinsey Quarterly.

Boudjellaba, H., Dufour, J.-M., & Roy, R. (1992). Testing causality between two vectors in multivariate ARMA models. Journal of the American Statistical Association, 87(420), 1082-1090.

Boyd, J. R. (1987). Patterns of Conflict. Unpublished briefing slides. United States Air Force.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School Working Paper 24-013.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32-64.

Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. Econometrica, 55(2), 251-276.

Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. Journal of Physiology-Paris, 100(1-3), 70-87.

Gartner. (2024). Predicts 2025: Artificial Intelligence. Gartner Research Note.

Godel, K. (1931). Uber formal unentscheidbare Satze der Principia Mathematica und verwandter Systeme I. Monatshefte fur Mathematik und Physik, 38, 173-198.

Good, I. J. (1950). Probability and the Weighing of Evidence. Griffin.

Hsiao, C. (1981). Autoregressive modelling and money-income causality detection. Journal of Monetary Economics, 7(1), 85-106.

Janis, I. L. (1972). Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes. Houghton Mifflin.

Janis, I. L., & Mann, L. (1977). Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment. Free Press.

Jeffreys, H. (1961). Theory of Probability (3rd ed.). Oxford University Press.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica, 59(6), 1551-1580.

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

Popper, K. R. (1963). Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge.

Sims, C. A. (1980). Macroeconomics and reality. Econometrica, 48(1), 1-48.

Builder, C. H., Bankes, S. C., & Nordin, R. (1999). Command concepts: A theory derived from the practice of command and control. RAND Corporation.

Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. Journal of Econometrics, 66(1-2), 225-250.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, 42(1), 230-265.

Fuster, J. M. (2003). Cortex and Mind: Unifying Cognition. Oxford University Press.

Fuster, J. M. (2006). The cognit: A network model of cortical representation. International Journal of Psychophysiology, 60(2), 125-132.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. Trends in Cognitive Sciences, 5(1), 10-16.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. Annual Review of Neuroscience, 30, 535-574.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. Nature, 447, 1075-1080.