SPINOR LABS · ALOFT · INTERNAL ARCHITECTURE STUDY

Agent Class Design Reference

THEORY → PROTOTYPE MAPPING

Six scientists. Six disciplines. Twenty-seven prototype agents.
Each design rule traced to the theory that earns it.

CLASS 01	CLASS 02	CLASS 03	CLASS 04	CLASS 05	CLASS 06
FEYNMAN	FERMI	MAXWELL	RAMA	MARCUS	WHEELER
reasoning	estimation	orchestration	integrity	meta-agent	timing

WHAT'S INSIDE

THE VERIFIED THEORY
Each scientist's actual contribution — with attribution flags preserved verbatim for diligence.

THE VISUAL MAPPING
A schematic per class translating the theory into the ALOFT design pattern.

THE AGENT SPEC
Per-prototype data structure, control flow, gating logic, metrics, and anti-patterns.

> *"The names aren't decoration. Each scientist supplies a usable engineering discipline — and every agent inherits it by design."*

INFORMATION-ACCURATE · ATTRIBUTION-FLAGGED · ILLUSTRATED EDITION

Source: ALOFT Agent Class Design Reference (verbatim) · illustrated for review & investor use

OVERVIEW

Executive Summary

ALOFT organizes its agent fleet into six classes, each named for a scientist (or, in Class 05, a Stoic philosopher) whose actual theoretical work supplies a usable engineering discipline — not a marketing metaphor. This document maps each figure's verified contributions to concrete design guidance for every locked prototype agent: data structures, control flow, gating logic, metrics, failure modes, and anti-patterns.

- FEYNMAN (Class 01 — Reasoning / Explainability): glass-box reasoning. Feynman diagrams make every interaction term explicit; path integrals sum over all histories before collapse; the Challenger investigation puts evidence before narrative. Thesis: every decision traceable to an auditable reason.
- FERMI (Class 02 — Estimation & Analysis): order-of-magnitude reasoning under sparse data. Factor decomposition where errors cancel; every answer ships with bounds and stated assumptions; calibration as a contract.
- MAXWELL (Class 03 — Orchestration): Maxwell founded control theory ("On Governors," 1868). Feedback governors → stable control loops; field theory → local, explicit handoffs; Maxwell's demon → the thermodynamic price of observation.
- RAMA (Class 04 — Perception / Hallucination Defence): the mirror box re-grounds a false percept with real feedback; Capgras separates recognition from verification; filling-in shows the brain confabulates missing data plausibly. Thesis: catch confabulation before the system trusts it.
- MARCUS (Class 05 — Meta-Agent): the Meditations as structured self-review; the dichotomy of control → autonomy tiers by reversibility; the discipline of assent → no promotion without eval evidence; "the obstacle is the way" → failures as improvement fuel.
- WHEELER (Class 06 — Probabilistic Observation / Timing): the delayed-choice experiment → defer the measurement decision to the last responsible moment; "it from bit" → every business event is an information signal with a maturity level; collapse → an explicit, logged state transition.

A cross-class synthesis closes the document: Feynman traces feed Marcus reflection; Rama verification gates Maxwell dispatch; Fermi bounds size the risk Wheeler times; Wheeler timing wraps the whole fleet.

> **HOW TO READ THIS DOCUMENT**
>
> Every chapter follows the same arc: the verified theory (with attribution flags preserved verbatim — these are diligence-critical), then a visual explanation translating that theory into the ALOFT design pattern, then the per-prototype-agent engineering spec (data structure · control flow · gating · metrics · anti-patterns), and a closing anti-patterns list. The diagrams are additive; the text is unchanged from the source reference.

CLASS 01 · ALOFT

FEYNMAN

RESEARCH & REASONING / EXPLAINABILITY

Richard Phillips Feynman (1918–1988), American theoretical physicist; shared the 1965 Nobel Prize in Physics with Julian Schwinger and Shin'ichirō Tomonaga for work on quantum electrodynamics (QED).

The theory, verified

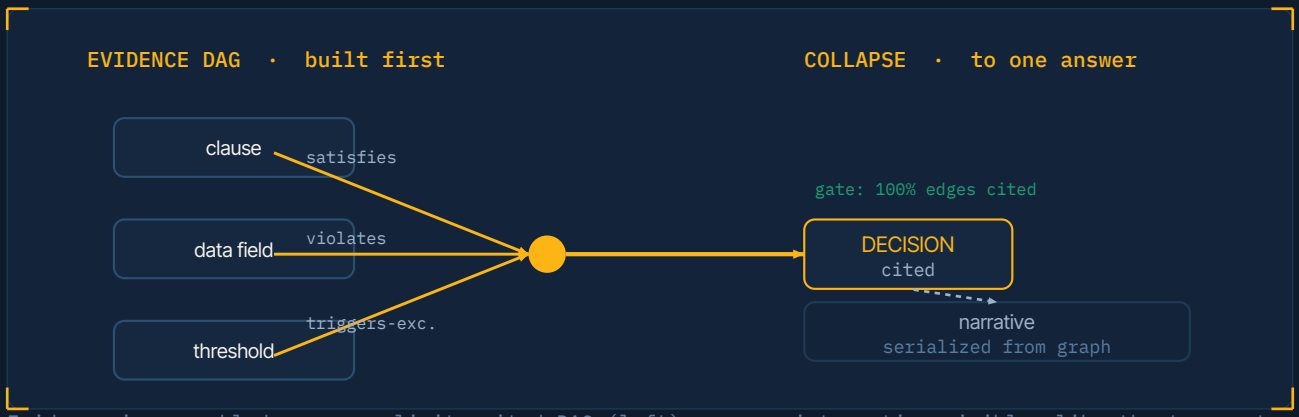
- Path integral formulation (sum over histories). "Space-Time Approach to Non-Relativistic Quantum Mechanics," Rev. Mod. Phys. 20:367–387 (1948). Replaces the classical single trajectory with a sum over all possible paths between initial and final states. Originated in his 1942 Princeton thesis; publication delayed by the Manhattan Project.
- Feynman diagrams as a visual calculus for QED: each line and vertex corresponds to an explicit term in the underlying mathematical expression.
- "What I cannot create, I do not understand." On his Caltech blackboard at his death, 15 Feb 1988.
- Challenger / Rogers Commission (1986). Feynman demonstrated O-ring brittleness with a C-clamp and ice water. Dissent published as Appendix F: "For a successful technology, reality must take precedence over public relations, for nature cannot be fooled."
- "Cargo Cult Science" (Caltech, 1974): "The first principle is that you must not fool yourself — and you are the easiest person to fool."

> **ATTRIBUTION FLAGS (preserve verbatim)**
> The blackboard motto is genuinely Feynman's, but is sometimes wrongly framed as a unique final "last statement" — it was simply what was on the board, alongside "Know how to solve every problem that has been solved." Sourcing nuance: the O-ring cold-temperature clue was relayed by General Donald Kutyna; the BBC (2013) traced the underlying insight to astronaut Sally Ride. Feynman threatened to remove his name unless his observations were included.

Class-level design thesis

Feynman-class agents are glass-box by construction. Three of his methods map directly onto reasoning architecture: Feynman diagrams → explicit-interaction trace structures (each evidence-to-conclusion interaction is a discrete, inspectable node — no hidden terms); path integrals → multi-hypothesis reasoning before collapse (enumerate and weight candidate explanations; the trace retains the unselected paths and weights); Challenger / Appendix F → evidence-before-narrative (conclusions built from verifiable evidence nodes; narrative generated last — "do not fool yourself" becomes a requirement to log disconfirming evidence).

FEYNMAN · VISUAL EXPLANATION



Evidence is assembled as an explicit, cited DAG (left) — every interaction visible, like the terms at a Feynman-diagram vertex. The narrative is serialized last, from the graph.

FEYNMAN · PROTOTYPE AGENTS

The four prototype agents

Regulatory Decision Tracer	0 unexplainable decisions
Credit Decision Explainer	memo drafting −82%
Contract Clause Reasoner	7.1× review throughput
Root-Cause Narrator	time-to-explained −67%

Regulatory Decision Tracer

THEORY ANCHOR Feynman diagrams (explicit terms) + Appendix F (evidence before narrative).

DATA STRUCTURE	A directed acyclic graph (DAG) where nodes are atomic evidence items (a regulation clause, a data field, a threshold) and edges are typed interactions (satisfies, violates, triggers-exception). Every edge carries a citation to source text.
CONTROL FLOW	Build the evidence DAG first; derive the decision by graph traversal; emit narrative as a serialization of the traversal path.
GATING LOGIC	No decision node may be marked “final” unless every incoming edge resolves to a cited source. Uncited inference is a hard block.
METRICS	Trace completeness (fraction of decision-bearing edges with citations = target 100%); auditor replay success rate (an independent reviewer reconstructs the decision from the trace alone).
ANTI-PATTERNS	Post-hoc rationalization (narrative written first, evidence retrofitted); “cargo cult” traces with the form of a citation list without the citations actually supporting the edges.

Credit Decision Explainer

THEORY ANCHOR Path integral (sum over hypotheses) + "must not fool yourself."

DATA STRUCTURE	A weighted hypothesis set — candidate decision rationales each with a contribution weight and the features that drove them — collapsing to the final adverse-action reasons.
CONTROL FLOW	Enumerate factor contributions across all candidate paths; retain counterfactual paths ("what would have flipped this decision"); collapse to the reported reasons.
GATING LOGIC	The reported reasons must be the highest-weight paths; if a high-weight path is a prohibited-basis proxy, block and escalate.
METRICS	Reason fidelity (do stated reasons match actual model attributions?); counterfactual coverage; adverse-action compliance.
ANTI-PATTERNS	Presenting a single tidy reason while suppressing competing high-weight factors (collapsing the path integral too early and hiding it).

Contract Clause Reasoner

THEORY ANCHOR Feynman diagrams (every interaction explicit) + the Feynman technique.

DATA STRUCTURE	Clause-interaction graph — clauses as nodes, cross-references / overrides / conflicts as typed edges.
CONTROL FLOW	Resolve clause interactions explicitly (precedence, carve-outs); produce a plain-language restatement per clause as a comprehension check.
GATING LOGIC	If the plain-language restatement cannot be regenerated and matched back to the source clause, flag as not-understood (the "what I cannot create" test operationalized).
METRICS	Conflict-detection recall; restatement round-trip fidelity.
ANTI-PATTERNS	Jargon passthrough (restating clause text in other clause text without grounding).

Root-Cause Narrator

THEORY ANCHOR The Challenger investigation method — evidence and minority dissent before consensus narrative.

DATA STRUCTURE	An evidence-first causal graph with explicit support for competing hypotheses, including a retained "minority report" branch.
CONTROL FLOW	Collect evidence → build candidate causal chains → weight → narrate. The narrative step is strictly downstream of evidence.
GATING LOGIC	A root cause cannot be promoted if a competing hypothesis retains material unrebutted evidence; dissent must be preserved, not overwritten.
METRICS	Evidence-to-conclusion coverage; rate of preserved dissent; "Russian-roulette" check (flag "it worked last time" reasoning as non-evidence — Feynman's exact critique of NASA treating prior O-ring erosion as a safety margin).
ANTI-PATTERNS	Management-narrative capture (conclusion shaped to be reassuring rather than true); normalizing repeated anomalies as safety margin.

ANTI-PATTERNS THE THEORY WARNS AGAINST

Narrative-first reasoning (rationalization) · hidden terms / uncited inference in traces · premature collapse to a single hypothesis with competing paths discarded · cargo-cult explainability (the form of a trace without the substance).

CLASS 02 · ALOFT

FERMI

ESTIMATION & ANALYSIS / OPERATIONAL INTELLIGENCE

Enrico Fermi (1901–1954), Italian-American physicist.

The theory, verified

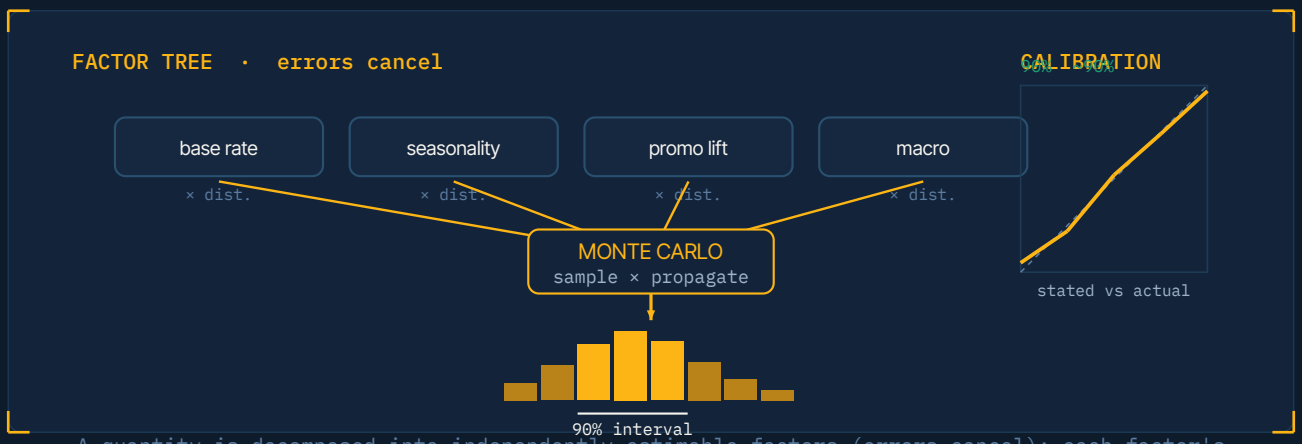
- Fermi problems / estimation method. Order-of-magnitude estimation by decomposing a quantity into estimable factors (the canonical "How many piano tuners in Chicago?"). The method relies on errors in independent factors tending to cancel.
- Trinity test yield estimate (16 July 1945). Fermi dropped paper scraps and read the blast displacement: "...The shift was about 2½ meters, which... I estimated to correspond to the blast that would be produced by ten thousand tons of T.N.T."
- Monte Carlo lineage. During a 1947 ENIAC downtime, Fermi devised the FERMIAC, an analog "Monte Carlo trolley" to trace neutron genealogies by random sampling.
- Also genuine: Fermi-Dirac statistics; the Fermi paradox.

> **ACCURACY FLAGS (diligence-grade)**
> Per J. I. Katz, "Fermi at Trinity," Nuclear Technology 207:sup1 (2021) (arXiv:2103.05784), Fermi's ~10 kt is about 40% of the radiochemically derived 25 ± 2 kt (soil-sample total yield 18.6 kt). The robust, citable claim: his estimate was correct to within roughly a factor of two from a rudimentary experiment, and was explicitly a lower bound — do not report a single "true value" without this nuance. Monte Carlo itself was developed by Ulam and von Neumann and named by Metropolis; Fermi did not "invent Monte Carlo" — he is part of its lineage and built the FERMIAC.

Class-level design thesis

Fermi-class agents never emit a bare point estimate. Each answer is a triple: estimate + interval + stated assumptions. Core principles: factor decomposition so errors cancel (a chain of rough independent factors beats one monolithic guess); Monte Carlo propagation (sample each factor's distribution to produce an output distribution, not a number); calibration as a contract (stated 90% intervals must capture actuals ~90% of the time, tracked over the fleet's history); dimensional-analysis sanity checks (units reconcile; order-of-magnitude plausibility checked before release).

FERMI · VISUAL EXPLANATION



A quantity is decomposed into independently estimable factors (errors cancel); each factor's distribution is sampled and propagated to an output distribution — never a bare number. Stated 90% intervals are tracked against actuals.

FERMI · PROTOTYPE AGENTS

The five prototype agents

Demand Signal Forecaster	MAPE 8.7% · −39% vs baseline
Procurement Spend Estimator	96% anomalies caught · 2.1% FP
Capacity Headroom Estimator	overprovisioning −18%
Inventory Risk Quantifier	11-day stockout warning
Pipeline Revenue Sizer	±12% across 3 quarters

Demand Signal Forecaster

THEORY ANCHOR factor decomposition + Monte Carlo propagation.

DATA STRUCTURE	Factor tree (base rate × seasonality × promo lift × macro factor), each factor a distribution.
CONTROL FLOW	Sample factor tree → aggregate → emit predictive distribution with quantiles.
GATING LOGIC	Refuse to emit a point forecast without an interval; block if any factor lacks a stated source or prior.
METRICS	Interval coverage (PIT calibration); Brier/CRPS for probabilistic accuracy; MAPE as secondary.
ANTI-PATTERNS	False precision (four significant figures on a factor-of-two estimate); single-scenario forecasting.

Procurement Spend Estimator

THEORY ANCHOR piano-tuner decomposition.

DATA STRUCTURE	Spend = Σ(category volume × unit price × adjustment factors), each estimable.
CONTROL FLOW	Decompose by category; estimate each factor with bounds; Monte Carlo aggregate.
GATING LOGIC	Dimensional check (currency × quantity reconciles); assumptions log mandatory.
METRICS	Realized-vs-estimated within stated interval; calibration drift over quarters.
ANTI-PATTERNS	Anchoring on last year's total without decomposition; hiding assumptions.

Capacity Headroom Estimator

THEORY ANCHOR order-of-magnitude bounds; lower-bound discipline (Trinity).

DATA STRUCTURE	Headroom = capacity − projected load, both as distributions.
CONTROL FLOW	Estimate load distribution via factor tree; compute headroom distribution; report probability of breach.
GATING LOGIC	Report as a conservative bound when data is sparse (Fermi's explicit lower-bound stance); flag thin-data factors.
METRICS	Breach-prediction calibration; coverage of stated headroom intervals.
ANTI-PATTERNS	Treating a mean projection as a guarantee; ignoring tail load.

Inventory Risk Quantifier

THEORY ANCHOR Monte Carlo propagation through demand × lead-time × supply variability.

DATA STRUCTURE	Joint distribution over stockout and overstock outcomes.
CONTROL FLOW	Simulate demand/lead-time scenarios; quantify P(stockout), expected shortfall.
GATING LOGIC	Every risk number paired with confidence interval and the dominating assumption.
METRICS	Realized stockout frequency vs predicted; interval coverage.
ANTI-PATTERNS	Point reorder thresholds with no uncertainty; assuming independence where correlated shocks exist.

Pipeline Revenue Sizer

THEORY ANCHOR factor decomposition (deal count × win rate × deal size × timing) + calibration contract.

DATA STRUCTURE	Stage-weighted factor tree with per-factor distributions.
CONTROL FLOW	Sample → aggregate to revenue distribution → emit quantiles + assumptions.
GATING LOGIC	Block single-number forecasts to finance; require calibration tag.
METRICS	Quarterly coverage of 90% intervals (~90% target); sharpness traded against coverage.
ANTI-PATTERNS	Sandbagging/optimism baked into factors without tracking; un-versioned assumption changes.

ANTI-PATTERNS THE THEORY WARNS AGAINST

Point estimates without intervals · monolithic guesses instead of factor decomposition · false precision / ignored dimensional checks · stated intervals never reconciled against actuals (broken calibration contract). Tetlock & Gardner's Superforecasting (2015) makes breaking problems into tractable sub-problems an explicit commandment and uses the Brier score as the calibration yardstick — Fermi-class agents should log a running Brier/coverage score per agent.

CLASS 03 · ALOFT

MAXWELL

ORCHESTRATION

James Clerk Maxwell (1831–1879), Scottish physicist. Naming note: this class was renamed from "Grossmann" to Maxwell; the rationale is that Maxwell is the founder of control theory, which is the orchestration discipline.

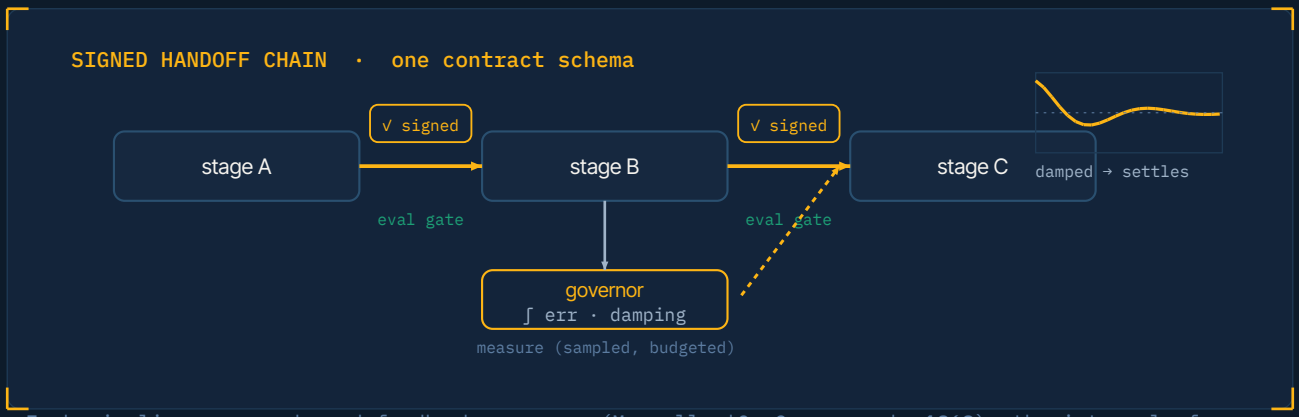
The theory, verified

- "On Governors" (1868), Proc. R. Soc. 16:270–283. The founding paper of control theory: the first mathematical stability analysis of feedback governors, distinguishing "moderators" (correction proportional to error) from true "governors" (also including a term proportional to the integral of the error), analyzing stability via the roots of the system's differential equations. Wiener, Cybernetics (1948): "the first significant paper on feedback mechanisms is an article on governors, which was published by Clerk Maxwell in 1868…"
- Maxwell's equations / unification. "A Dynamical Theory of the Electromagnetic Field" (1865) unified electricity, magnetism and light. The displacement-current term made the system consistent and enabled the electromagnetic wave equation.
- Maxwell's demon (1871). A thought-experiment agent that observes and sorts molecules. Szilárd (1929) quantified the information–entropy link; Landauer (1961) showed information erasure carries the irreducible thermodynamic cost — "information is physical."
- Maxwell-Boltzmann distribution. Stable macroscopic behavior emerging from many micro-components.

Class-level design thesis

Maxwell-class orchestrators coordinate many agents under one framework with signed handoff contracts, rollback, and eval gates before dispatch. Feedback governors → control loops with stability conditions (damping and integral-of-error correction prevent oscillation and retry storms — an under-damped retry policy is a literal control-theory instability). Unification → one orchestration contract schema across all departments. Field theory → no action at a distance (state propagates locally; every handoff is local and explicit). Maxwell's demon / Landauer → the cost of observation (monitoring, sampling, and log erasure carry real cost; orchestration must budget observation).

MAXWELL · VISUAL EXPLANATION



Each pipeline runs a damped feedback governor (Maxwell, 'On Governors', 1868): the integral-of-error term corrects drift while damping prevents the retry-storm oscillation. Handoffs are local, signed, and gated — no action at a distance.

MAXWELL · PROTOTYPE AGENTS

The five prototype agents

Procure-to-Pay Pipeline	cycle time −41% · 73% touchless
Hire-to-Onboard Coordinator	98.4% day-one ready
Order-to-Cash Orchestrator	DSO −6.2 days
Audit Readiness Orchestrator	evidence effort −78%
Quote-to-Contract Pipeline	9.4d → 2.1d

Procure-to-Pay Pipeline

THEORY ANCHOR governor feedback loop + signed handoff contracts.

DATA STRUCTURE	Stage graph with a signed contract per handoff (preconditions, payload schema, postconditions, rollback token).
CONTROL FLOW	Control loop monitors throughput/error; integral term corrects persistent drift; damping caps retry rate.
GATING LOGIC	Eval gate before dispatch to next stage; contract signature verified or handoff blocked.
METRICS	Loop stability (oscillation amplitude in retry rate); settling time after a shock; SLA adherence.
ANTI-PATTERNS	Retry storms (under-damped loop); unsigned handoffs; action-at-a-distance state mutation.

Hire-to-Onboard Coordinator

THEORY ANCHOR unification (one schema) + local handoffs.

DATA STRUCTURE	Common contract schema instantiated for recruiting → IT → payroll stages.
CONTROL FLOW	Each stage consumes a signed contract, emits a signed completion; coordinator never bypasses a stage.
GATING LOGIC	Rollback on any failed postcondition; compensating transactions defined per stage.
METRICS	Handoff success rate; rollback frequency; orphaned-task count (should be zero).
ANTI-PATTERNS	Bespoke per-stage schemas (defeats unification); silent partial completion.

Order-to-Cash Orchestrator

THEORY ANCHOR governor stability + demon/observation cost.

DATA STRUCTURE	Event-sourced state machine; sampling budget per monitored signal.
CONTROL FLOW	Feedback on collection latency; sampled observation rather than continuous polling (Landauer cost discipline).
GATING LOGIC	Dispatch invoicing only after credit/eval gate passes; damping on dunning retries.
METRICS	DSO control-loop stability; observation cost vs information gained.
ANTI-PATTERNS	Over-instrumentation (unbounded observation cost); oscillating dunning cadence.

Audit Readiness Orchestrator

THEORY ANCHOR Maxwell's demon (sorting + information cost) + explicit local state.

DATA STRUCTURE	Evidence ledger; each control check is a sort operation with a logged information cost.
CONTROL FLOW	Continuously sort artifacts into compliant/non-compliant; erasure (log rotation) explicitly costed and gated.
GATING LOGIC	No artifact discarded without recorded justification (erasure has a cost — make it deliberate, per Landauer).
METRICS	Audit-trail completeness; cost-of-observation budget adherence.
ANTI-PATTERNS	Silent log deletion; unbounded monitoring overhead.

Quote-to-Contract Pipeline

THEORY ANCHOR signed contracts + eval gates + damping.

DATA STRUCTURE	Stage graph quote → approval → contract, each handoff signed and versioned.
CONTROL FLOW	Approval feedback loop; damped escalation to prevent approval ping-pong.
GATING LOGIC	Eval gate (pricing sanity, margin floor) before contract generation; rollback to quote on rejection.
METRICS	Cycle-time stability; approval-loop oscillation; contract error rate.
ANTI-PATTERNS	Approval oscillation (under-damped); unsigned pricing changes propagating downstream.

ANTI-PATTERNS THE THEORY WARNS AGAINST

Under-damped feedback → retry storms / oscillation (the instability Maxwell analyzed) · action-at-a-distance state mutation (violates the field/local-propagation principle) · fragmented per-department schemas (defeats unification) · ignoring the cost of observation and erasure (Landauer); over- or under-instrumentation.

CLASS 04 · ALOFT

RAMA

PERCEPTION & SYNTHESIS / HALLUCINATION DEFENCE

Vilayanur S. Ramachandran (b. 1951), neuroscientist. Key book: Phantoms in the Brain (with Sandra Blakeslee, 1998).

The theory, verified

- Phantom limbs & mirror box therapy. Ramachandran & Hirstein, "The Perception of Phantom Limbs," Brain 121 (1998), 1603–1630. A brain deprived of grounded input generates its own (false) percept; reintroducing grounded visual feedback corrects it cheaply.
- Capgras delusion. Hirstein & Ramachandran (1997). Recognition pathway intact but the emotional-verification pathway is severed — perception looks right but fails verification. A verification-pathway failure model.
- Anosognosia & confabulation / the left-hemisphere "interpreter." Patients confidently narrate false explanations; a right-hemisphere "devil's advocate" is needed to check the interpreter.
- Perceptual filling-in (blind-spot completion): the brain plausibly completes missing data. Synesthesia / cross-modal binding. Cheap-experiment style: mirrors and cotton swabs — lightweight, high-signal probes.

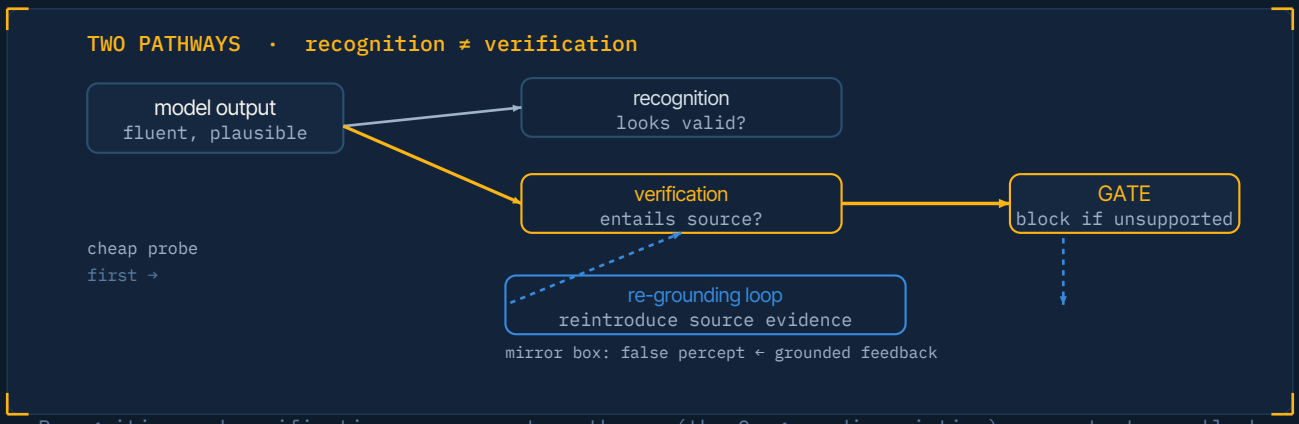
MODERN ENGINEERING BRIDGE

The predictive-processing / free-energy account (Friston, Nat. Rev. Neurosci. 11, 2010) explains filling-in and confabulation as the brain imposing strong top-down priors when sensory evidence is weak. Powers/Mathys/Corlett, Science 357 (2017): "Hallucinations… may arise when strong priors cause a percept in the absence of input." This is the bridge to LLM confabulation: the model fills in plausibly when grounding evidence is weak — a precision-weighting failure.

Class-level design thesis

Rama-class agents catch model confabulation before downstream systems trust it. Mirror box → re-grounding loop (reintroduce grounded source evidence to break the confabulation). Capgras → separate recognition from verification (an output can "look right" and still fail verification; never let recognition confidence substitute for verification). Filling-in → detect interpolated spans (mark spans not supported by source). Cheap probes first (lightweight high-signal checks before expensive verification).

RAMA · VISUAL EXPLANATION



Recognition and verification are separate pathways (the Capgras dissociation): an output can 'look right' and still fail grounding. Unsupported spans trigger a mirror-box re-grounding loop that reintroduces source evidence before the system trusts the claim.

RAMA · PROTOTYPE AGENTS

The five prototype agents

Source Grounding Verifier	every output traceable to span
Cross-Modal Signal Fuser	discrepancy F1 0.93
Visual Anomaly Detector	95.2% recall @ 3.1% FP
Hallucination Sentinel	412 fabrications / 100K tokens
Citation Integrity Checker	99.1% resolution accuracy

Source Grounding Verifier

THEORY ANCHOR mirror box re-grounding + NLI entailment.

DATA STRUCTURE	Claim → candidate source-passage map with an entailment label (entailed / contradicted / neutral).
CONTROL FLOW	Decompose output into atomic claims; for each, retrieve source; run NLI; claims with no entailing passage flagged unverifiable.
GATING LOGIC	Block claims labeled contradicted; mark neutral/unverifiable. NLI grounding is strongest for faithfulness errors but does not catch extrinsic hallucinations alone.
METRICS	Faithfulness/groundedness score; contradiction recall; cost per claim — favor cheap learned checkers (MiniCheck-class) for full-coverage checking.
ANTI-PATTERNS	Trusting fluency as grounding; verifying only a sampled subset when full coverage is feasible.

Cross-Modal Signal Fuser

THEORY ANCHOR synesthesia / cross-modal binding.

DATA STRUCTURE	Aligned multi-source representation (text, tabular, image-derived) with per-source provenance.
CONTROL FLOW	Bind signals across modalities; flag binding conflicts (a value present in one modality contradicted by another).
GATING LOGIC	Require agreement across ≥2 independent modalities before high-confidence assertion.
METRICS	Cross-modal agreement rate; conflict-detection precision.
ANTI-PATTERNS	Spurious binding (false cross-modal correlation); collapsing modalities before conflict check.

Visual Anomaly Detector

THEORY ANCHOR filling-in (the brain completes missing data) — inverted to detect completion.

DATA STRUCTURE	Region map flagged for interpolated/low-evidence spans.
CONTROL FLOW	Cheap saliency/consistency probe first; escalate flagged regions to expensive verification.
GATING LOGIC	Mark plausibly-completed-but-unsupported regions; do not pass as observed fact.
METRICS	Interpolation-detection recall; cheap-probe escalation efficiency.
ANTI-PATTERNS	Accepting plausible completion as evidence; uniform expensive scanning (ignores cheap-probe economy).

Hallucination Sentinel

THEORY ANCHOR confabulation / left-hemisphere interpreter + predictive-processing strong-priors.

DATA STRUCTURE	Per-claim confidence + self-consistency cluster set across resampled generations.
CONTROL FLOW	Sample multiple generations; cluster by semantic meaning; claims that fragment across clusters are flagged (operationalizes SelfCheckGPT, Manakul et al., EMNLP 2023). Pair with NLI grounding.
GATING LOGIC	The institutionalized “right-hemisphere devil's advocate” — block confident outputs that lack grounding or fail consistency.
METRICS	Hallucination detection rate; false-flag rate; calibration of confidence vs correctness.
ANTI-PATTERNS	Confidence-as-truth (the interpreter confabulating coherence); single-sample trust; over-flagging that trains operators to ignore alerts.

Citation Integrity Checker

THEORY ANCHOR Capgras — recognition vs verification dissociation.

DATA STRUCTURE	Citation → (exists? resolves? actually supports the claim?) triple.
CONTROL FLOW	Recognition pathway (citation looks valid/formatted) is separated from verification pathway (citation content entails the claim).
GATING LOGIC	A well-formed citation that does not entail the claim is the Capgras case — looks right, fails verification → block.
METRICS	Citation-support precision; fabricated/non-resolving citation catch rate.
ANTI-PATTERNS	Accepting plausible-looking citations without content verification.

ANTI-PATTERNS THE THEORY WARNS AGAINST

Treating fluency/plausibility as evidence (filling-in mistaken for perception) · letting recognition confidence substitute for verification (Capgras failure) · single-sample trust without consistency checks · skipping cheap probes and over-spending on expensive verification (or never verifying).

CLASS 05 · ALOFT

MARCUS

META-AGENT

Marcus Aurelius (121–180 AD), Roman emperor and Stoic philosopher; author of the Meditations, a private notebook of self-directed reflections not written for publication.

The texts, verified

- "The obstacle is the way." Meditations 5.20 (Hays): "The impediment to action advances action. What stands in the way becomes the way." Anchors failures-as-fuel.
- Dichotomy of control. The Stoic distinction between what is and is not "up to us."
- Discipline of assent. Do not assent to impressions without examination; distinguish impressions from judgments.
- The evening review. A structured nightly self-examination (see flag).
- Premeditatio malorum and the "inner citadel" are genuine Stoic/Marcus themes.

ATTRIBUTION FLAGS (important for diligence)

The sharpest formulation of the dichotomy of control is Epictetus, Enchiridion 1 — not Marcus; use Epictetus as the primary anchor. The explicit evening-review formula is Seneca, De Ira 3.36, which Seneca credits to the Pythagorean Sextius; Marcus's Meditations is structurally a self-review but is not the source of the explicit formula. State the lineage honestly: Pythagoras/Sextius → Seneca → Stoic practice.

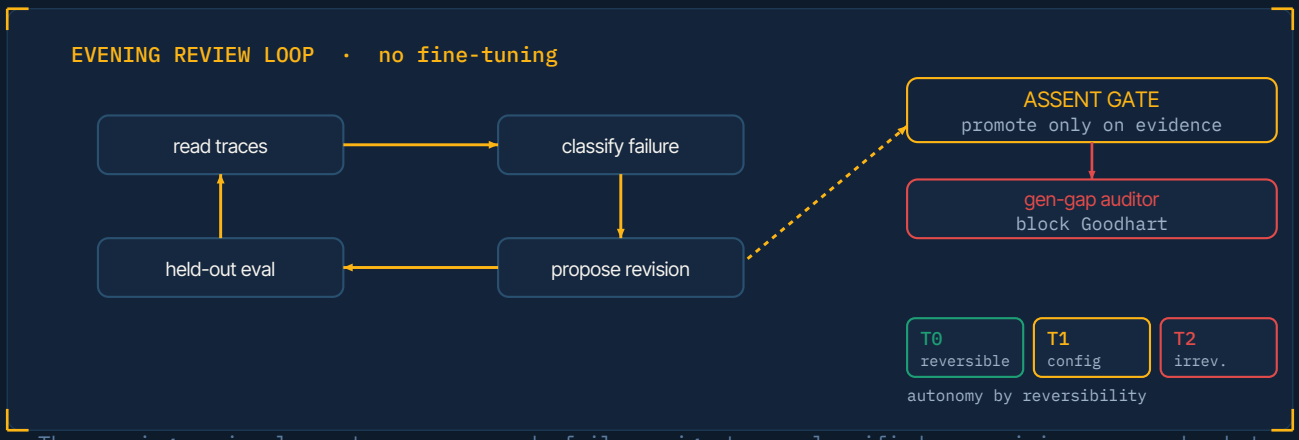
MODERN ENGINEERING BRIDGE

The self-improvement loop without fine-tuning maps to Reflexion (Shinn et al., NeurIPS 2023) — "reinforce language agents not by updating weights, but instead through linguistic feedback," agents maintaining reflective text in an episodic memory buffer (91% vs 80% HumanEval pass@1) — and Self-Refine (Madaan et al., NeurIPS 2023), iterative self-feedback with no additional training. The anti-Goodhart stance maps to Goodhart's law as stated by Marilyn Strathern (1997): "When a measure becomes a target, it ceases to be a good measure."

Class-level design thesis

Marcus is the Stoic meta-agent: it reviews all other agents and runs the self-improvement loop without fine-tuning. Dichotomy of control → autonomy tiers by reversibility (irreversible/high-blast-radius actions get the lowest tier and a mandatory human gate). Discipline of assent → no promotion without held-out eval evidence (the Generalisation Gap Auditor is the institutionalized refusal to assent to Goodharted metrics). Evening review → scheduled batch reflection cadence over the day's trajectories. Obstacle-is-the-way → failure signatures as the routing table for improvement levers.

MARCUS · VISUAL EXPLANATION



The evening-review loop: traces are read, failure signatures classified, a revision proposed — but nothing is promoted without held-out eval evidence (the discipline of assent). Autonomy is tiered by reversibility; the Generalisation-Gap auditor blocks Goodharted gains.

MARCUS · PROTOTYPE AGENTS

The four prototype agents

Failure Signature Classifier	91% agreement with triage
Exemplar Curator	retrieval relevance +22%
Generalisation Gap Auditor	blocked 14% overfit promotions
Evening Review Conductor	64% of proposals accepted

Failure Signature Classifier

THEORY ANCHOR obstacle-is-the-way (5.20) — failures are the raw material of improvement.

DATA STRUCTURE	Taxonomy of failure signatures keyed to traces (Feynman-class traces are the input); each signature maps to an improvement lever.
CONTROL FLOW	Ingest traces → cluster failures → assign signatures → route to levers.
GATING LOGIC	Unclassifiable failures escalate to human review rather than being silently dropped.
METRICS	Classification coverage; signature stability; lever hit rate (did the routed fix reduce recurrence?).
ANTI-PATTERNS	Discarding novel failures that do not fit existing signatures; treating symptoms not root causes.

Exemplar Curator

THEORY ANCHOR Reflexion/Self-Refine episodic memory — verbal feedback retained as exemplars, not weight updates.

DATA STRUCTURE	Curated exemplar store (good/bad trajectories with verbal lessons), versioned.
CONTROL FLOW	Select high-value exemplars from reviewed trajectories; inject into agent context (no fine-tuning).
GATING LOGIC	An exemplar enters the store only if it improves held-out eval performance.
METRICS	Exemplar utility (marginal eval lift); store size vs benefit (avoid context bloat).
ANTI-PATTERNS	Memorizing idiosyncratic cases that don't generalize; unversioned exemplar drift.

Generalisation Gap Auditor

THEORY ANCHOR discipline of assent + Goodhart's law (Strathern 1997).

DATA STRUCTURE	Paired train/held-out metric distributions per candidate revision.
CONTROL FLOW	Measure the gap between in-sample (flattering) and held-out performance; large gaps signal Goodharting/overfitting.
GATING LOGIC	The institutionalized refusal to assent — block promotion when the generalisation gap exceeds threshold, regardless of how good the in-sample metric looks.
METRICS	Generalisation gap; post-promotion regression rate; Goodhart incidents caught.
ANTI-PATTERNS	Optimizing a proxy until it stops measuring the goal; assenting to impressive in-sample numbers.

Evening Review Conductor

THEORY ANCHOR Seneca's evening review (De Ira 3.36), credited honestly; Meditations as structured self-review.

DATA STRUCTURE	Daily batch of trajectories with a structured review record (what was done, what failed, what to change).
CONTROL FLOW	Scheduled batch reflection cadence; aggregate the day's failure signatures; emit prioritized revision proposals.
GATING LOGIC	Revisions enter the promotion pipeline (gated by the Generalisation Gap Auditor), never auto-deployed.
METRICS	Review coverage (fraction of trajectories reviewed); proposal acceptance rate; day-over-day failure reduction.
ANTI-PATTERNS	Continuous reactive thrash instead of disciplined batch cadence; review without follow-through.

ANTI-PATTERNS THE THEORY WARNS AGAINST

Promoting revisions on flattering in-sample impressions (failure of assent) · Goodharting proxies until they stop measuring the goal · granting high autonomy to irreversible/high-blast-radius actions (violating the dichotomy of control) · misattributing the evening-review formula to Marcus rather than Seneca/Sextius, or the dichotomy of control to Marcus rather than Epictetus.

CLASS 06 · ALOFT

WHEELER

PROBABILISTIC OBSERVATION / TIMING INTELLIGENCE

John Archibald Wheeler (1911–2008), American physicist; Feynman's doctoral advisor; coined/popularized "black hole."

The theory, verified

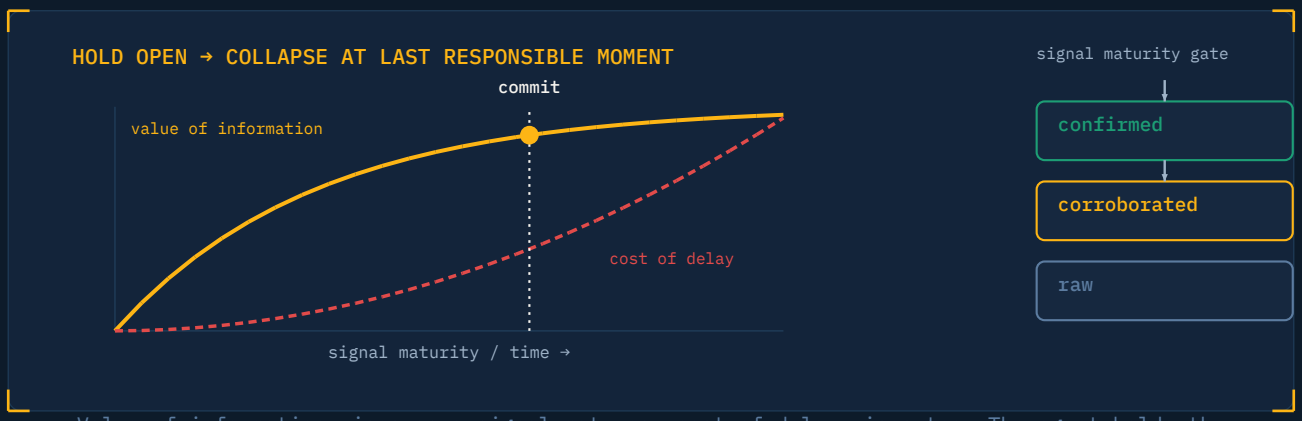
- Delayed-choice experiment (1978). The decision of how/whether to measure can be deferred until after the photon has traversed the apparatus. "No phenomenon is a phenomenon until it is an observed phenomenon"; "the past has no existence except as it is recorded in the present."
- "It from bit" (1989). "Information, Physics, Quantum: The Search for Links": "...what we call reality arises... from the posing of yes-no questions... all things physical are information-theoretic in origin."
- Participatory universe and the "surprise" version of twenty questions: the answer is negotiated by the questions asked; each query constrains the hypothesis space.
- Genuine: quantum measurement / wavefunction collapse; "radical conservatism" methodology; Wheeler-DeWitt and the problem of time.

> **MODERN ENGINEERING BRIDGE**
>
> The timing discipline maps to optimal stopping theory / the secretary problem (the "37%" / 1/e stopping rule, which selects the best candidate ~37% of the time irrespective of pool size) and the lean/agile "last responsible moment" principle: defer commitment until the cost of further delay exceeds the value of more information.

Class-level design thesis

Wheeler-class agents hold probability distributions open rather than collapsing them too early, and commit at the last responsible moment. "The right call at the wrong moment is still the wrong call." Delayed choice → last-responsible-moment commit semantics. Observation participates → model the cost and effect of observing (alerting changes the observed system — treat observation as an intervention with cost). It from bit → every business event is a signal with a maturity level. Wavefunction collapse → an explicit, logged state transition with trigger conditions. Twenty-questions → iterative narrowing where each query maximally constrains the hypothesis space.

WHEELER · VISUAL EXPLANATION



Value of information rises as a signal matures; cost of delay rises too. The agent holds the probability distribution open and commits at the last responsible moment — an explicit, logged 'collapse' — rather than acting on an immature signal.

WHEELER · PROTOTYPE AGENTS

The four prototype agents

Observation Window Scheduler	polling cost −38%
Intervention Timing Optimizer	premature interventions −61%
Signal Maturity Gater	early escalations 1-in-4 → 1-in-11
Market Entry Timing Advisor	beat baseline in 71% of windows

Observation Window Scheduler

THEORY ANCHOR delayed choice + optimal stopping.

DATA STRUCTURE	Signal with an evolving posterior and a value-of-information estimate over candidate observation times.
CONTROL FLOW	Defer observation while expected information gain rises faster than cost; commit at the optimal-stopping threshold (1/e-style baseline).
GATING LOGIC	Do not collapse to a decision before the window unless a hard deadline forces it; log the trigger.
METRICS	Regret vs. oracle timing; fraction of decisions made at the last responsible moment vs. prematurely.
ANTI-PATTERNS	Collapsing too early (acting on immature signal); analysis paralysis past the responsible moment.

Intervention Timing Optimizer

THEORY ANCHOR observation participates (observer effect) + collapse as logged transition.

DATA STRUCTURE	Intervention candidate set with modeled system response (including the effect of the intervention itself).
CONTROL FLOW	Simulate intervention effect; choose timing that maximizes net effect minus disturbance cost.
GATING LOGIC	Explicit collapse event when committing; logged trigger conditions and expected vs. realized effect.
METRICS	Intervention efficacy; observer-effect cost (e.g., alert fatigue incurred); false-trigger rate.
ANTI-PATTERNS	Intervening so often the observation degrades the system (alert fatigue); unlogged collapses.

Signal Maturity Gater

THEORY ANCHOR "it from bit" — every event is a signal with a maturity level.

DATA STRUCTURE	Signal registry tagging each event with maturity (raw → corroborated → confirmed).
CONTROL FLOW	Gate downstream actions on minimum maturity; promote maturity as corroborating bits arrive.
GATING LOGIC	Immature signals cannot trigger irreversible action (links to Marcus's reversibility tiers).
METRICS	Premature-action rate; maturity-promotion latency.
ANTI-PATTERNS	Acting on raw/unconfirmed signal; treating a single bit as a confirmed phenomenon.

Market Entry Timing Advisor

THEORY ANCHOR twenty-questions iterative narrowing + optimal stopping.

DATA STRUCTURE	Hypothesis space over entry windows, narrowed by each new datum (query).
CONTROL FLOW	Each information query maximally constrains remaining entry-window hypotheses; commit at the stopping threshold.
GATING LOGIC	Recommend "wait" while value of information exceeds cost of delay; recommend "commit" at threshold with explicit collapse log.
METRICS	Timing regret vs. realized optimum; calibration of the "wait/commit" call.
ANTI-PATTERNS	Committing on the first plausible window (premature collapse); waiting past the point where delay cost dominates.

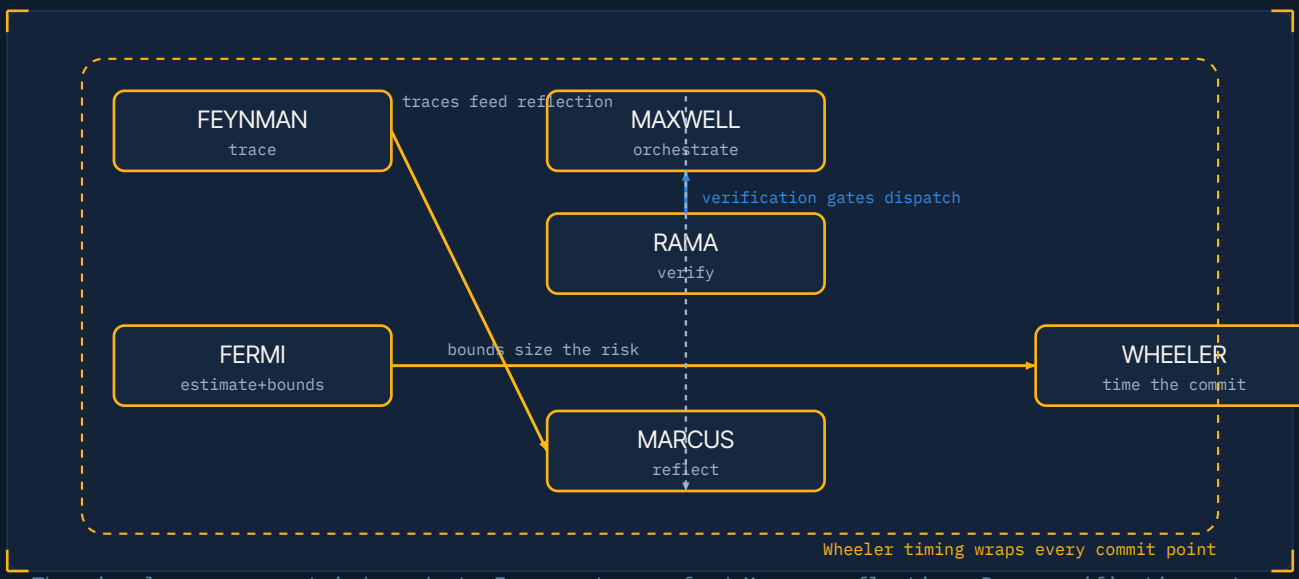
ANTI-PATTERNS THE THEORY WARNS AGAINST

Premature collapse: acting before the signal matures / before the last responsible moment · ignoring the observer effect (monitoring that degrades the system it watches) · treating immature single-bit signals as confirmed phenomena · unlogged collapse events (no record of why/when commitment occurred).

CROSS-CLASS

How the six theories interlock

The classes are not independent; their theories compose into one pipeline discipline.



The six classes are not independent. Feynman traces feed Marcus reflection; Rama verification gates Maxwell dispatch; Fermi bounds size the risk Wheeler times; Wheeler timing wraps every commit point in the fleet.

- Feynman traces feed Marcus reflection. Class 01's explicit evidence-DAG traces are the input substrate Class 05 reads to classify failure signatures. Without glass-box traces, the meta-agent has nothing to assent to or refuse.
- Rama verification gates Maxwell dispatch. Class 04's grounding verdicts are an eval gate on Class 03's handoffs: an orchestrator must not dispatch on a confabulated or unverified output. Capgras-style “looks right, fails verification” is exactly the gate Maxwell's signed-contract precondition should encode.
- Fermi bounds size the risk that Wheeler times. Class 02's estimate+interval feeds Class 06's value-of-information and stopping thresholds: you cannot compute “last responsible moment” without a calibrated uncertainty estimate. Fermi quantifies the distribution; Wheeler decides when to collapse it.
- Wheeler timing wraps everything. Every commit point — Maxwell dispatch, Rama escalation, Marcus promotion — is subject to Wheeler's last-responsible-moment discipline and explicit, logged collapse.
- Maxwell's Landauer cost and Wheeler's observer effect are the same insight from two directions: observation is not free. Class 03 budgets it thermodynamically; Class 06 models its effect on the observed system.
- Marcus's dichotomy of control and Wheeler's signal maturity co-gate autonomy: irreversible actions require both high signal maturity (Wheeler) and the lowest autonomy tier (Marcus).

THE UNIFYING STANCE

Shared across Feynman's “do not fool yourself,” Fermi's calibration contract, Maxwell's stability conditions, Rama's verification pathway, Marcus's discipline of assent, and Wheeler's last-responsible-moment: disciplined epistemic humility made executable — quantify uncertainty, ground claims, observe at a budgeted cost, commit deliberately, and keep every decision auditable.

APPENDIX

Sourcing

FEYNMAN "Space-Time Approach to Non-Relativistic Quantum Mechanics," Rev. Mod. Phys. 20:367–387 (1948); Rogers Commission Report Appendix F (1986); "Cargo Cult Science," Caltech 1974; blackboard motto via Caltech Archives. Misattribution flags (blackboard-as-"final statement"; Kutyna/Ride O-ring clue) noted in-chapter.

FERMI Fermi problem; "Fermi at Trinity," J. I. Katz, Nuclear Technology 207:sup1 (2021), arXiv:2103.05784 — ~10 kt ≈ 40% of radiochemical 25 ± 2 kt; soil-sample total 18.6 kt. FERMIAC (Metropolis, Los Alamos Science 1987). Monte Carlo attribution flag (Ulam/von Neumann/Metropolis). Calibration: Brier (1950); Tetlock & Gardner, Superforecasting (2015).

MAXWELL "On Governors," Proc. R. Soc. 16:270–283 (1868); Wiener, Cybernetics (1948); "A Dynamical Theory of the Electromagnetic Field" (1865) and displacement current (1861–62); Dyson assessment; Maxwell's demon (1871) with Szilárd (1929) and Landauer (1961) resolutions.

RAMA Ramachandran & Hirstein, "The Perception of Phantom Limbs," Brain 121:1603–1630 (1998); Phantoms in the Brain (1998); Capgras, Proc. R. Soc. B 264 (1997). Bridge: Friston, Nat. Rev. Neurosci. 11 (2010); Powers/Mathys/Corlett, Science 357 (2017); Corlett et al., Trends Cogn. Sci. 23 (2019). SelfCheckGPT (Manakul et al., EMNLP 2023, arXiv:2303.08896); NLI entailment for faithfulness.

MARCUS Meditations 5.20 (Hays); dichotomy of control → Epictetus, Enchiridion 1 (flag); evening review → Seneca, De Ira 3.36, credited to Sextius (flag). Reflexion (Shinn et al., NeurIPS 2023, arXiv:2303.11366); Self-Refine (Madaan et al., NeurIPS 2023, arXiv:2303.17651); Goodhart via Strathern (1997), European Review 5(3):305–321, p. 308.

WHEELER Delayed-choice (1978); "Information, Physics, Quantum: The Search for Links" (Tokyo, 1989) for "it from bit" and the twenty-questions surprise version. Bridge: secretary problem / optimal stopping (1/e rule, ~37%); lean/agile last-responsible-moment.

PRODUCTION NOTE

This illustrated edition is additive: the research text is preserved verbatim from the source reference, with the visual explanations layered on top. Before any line feeds investor or class-page material, preserve every attribution flag verbatim (Fermi yield nuance; Monte Carlo lineage; dichotomy-of-control = Epictetus; evening-review = Seneca/Sextius; Feynman blackboard not a "final statement"; Goodhart phrasing = Strathern). Any quote lifted into marketing must retain its exact source string. The Grossmann → Maxwell rename rationale (founder of control theory) should be stated wherever the class lineage is described.