

A T R E A T I S E

NINETY PERCENT

*How a large enterprise runs on artificial intelligence at a tenth of
its headcount, and the four dials that take it there*

INTERIOR, *n.* Everything a firm contains that exists to move context between minds
that cannot see the same state at once. Not the work. The cost of doing the work with
skulls.

EDGE, *n.* Where a promise meets a counterparty, a market, a regulator or physics, and
the world's verdict arrives. The only place the loop closes.

DIAL, *n.* One of four numbers that set how fast the edge answers, and therefore how
far authority reaches.

Bo Chen

September 2026

Written with Claude Fable 5.1 from the programme's own instruments: the synthetic firm ACME-500,
the resident kernel FUSOR, the CUDA org solver family, the flight computer, and the receipts of 13 to
15 September 2026. Every number carries a label: *measured*, *planted*, *budget*, or *derived*.

I · THE CLAIM

A large enterprise whose headquarters is made of knowledge workers, the people who read, decide, coordinate, report, approve and escalate, runs on artificial intelligence that exists today at a tenth of its current headcount, and the path to that number is arithmetic. No new kind of model is required and no research result is waited on. The pieces are a language model anyone can rent, a graphics card anyone can buy, an append-only record, a small amount of deterministic code that decides nothing but permits, a simulator that is the same binary as the product, and one idea the industry has not worked out: the interior of a corporation is not work. It is the cost of doing work with minds that cannot see the same state at the same time, and once one machine holds the whole state at once, the interior is not automated. It is deleted.

This document says how, with the mathematics that governs it, the code that exists, the measurements taken on a synthetic firm of five hundred seats, the exact path by which a headquarters goes from five hundred people to fifty and then past it, and the four dials that set the number. Where a number has been measured it is marked as measured; where it is a budget computed from measured inputs it is marked as a budget; where it is planted in the simulator to model a claim it says so. That discipline is the instrument.

Four sentences carry everything that follows. The interior of an organization is context logistics for skulls, and it is deleted by a fold over a record the machine can read. The edge of an organization, where a promise meets a counterparty, a market, a regulator or physics, is where the world's verdict arrives, so authority is earned there at the rate the world answers, and that rate is a design variable with four dials. The judgment that stands at every decision is bought, not built; what is built is the harness around one call to it, and the harness is language-free, which is why it trains on ten thousand synthetic firms before a real one sees it and why the simulator and the product are one program. And the enterprise learns at the rate its state changes while it is licensed at the rate the world answers, so the machine's share of the work is a curve that falls every quarter as thought compiles into number.

II · WHAT A HEADQUARTERS IS

An organization is a function from arrivals to committed effects. Orders arrive and filled orders leave. Claims arrive and settlements leave. Invoices go out and payments come back. Candidates apply and hires are made. Filings fall due and are made. Each arrival is a promise the firm has made to something outside itself, with a class, a value, a due day, a set of facts that decide its correct disposition, and a delay after which the world says whether the disposition was right. Call that function Φ . Everything the firm contains that is not Φ is what it cost to run Φ on people.

Run it on people and a second machine grows around the first. People need eyes, so applications exist to render database state pretty enough to think over. People cannot see the same state at once, so meetings exist to synchronize them. People cannot carry state between them without projecting it, so reports, decks and status updates exist. People cannot verify each other, so approval chains exist, and the chain is called a hierarchy. People have to be loaded, so onboarding exists. People cluster by the tools they can hold, so departments exist. And people need support, and support needs support, so the α -functions exist: human resources, information technology for people, payroll, facilities, the project office, the quality function that exists to catch human error, each with its own interior, recursively.

Every one of those structures compensated for a limit of skulls, and every one has a name for what it becomes when the state is on a plane the machine can read.

THE STRUCTURE	WHAT IT COMPENSATED FOR	WHAT IT BECOMES ON THE PLANE
the meeting	skulls cannot see the same state at once	a standing query over the field; there is no verb for it
the report, the deck, the status	state had to be carried as a projection	a view rendered from the field, at any date, free
the application and its screens	a skull needs eyes	the template: the screen's query log is the class's context
the hierarchy of approvals	trust among skulls was a chain	the licence table: authority per class and band, a table not a tree
onboarding	a skull has to be loaded	the cognitive index: the cell is kept already read
the department	skulls cluster by the tools they can hold	the wire: an obligation chain from arrival to discharge
the α -functions	skulls need support	nothing

THE STRUCTURE	WHAT IT COMPENSATED FOR	WHAT IT BECOMES ON THE PLANE
the documented process	the humans' static, shared version of a thing done	latent per instance; compiled per class when stable
the accounting close	a reconciliation between books people kept apart	a fold: every effect carries its inverse, so the ledger is double-entry by construction

The size of the second machine is not a guess. A synthetic firm of five hundred seats, built to model the claim and instrumented to the minute, spends its day as follows: glue (reading, skimming, re-reading yesterday) 32.4 percent; meetings 16.6; fetching context from applications 16.3; framing it 8.4; the decision itself 15.9; committing it 6.9; transport 3.2; rework 0.3 (planted). The decision and the commit, the two acts that are actually Φ , are under a quarter of the day. Half the day is coordination and a quarter is assembling context. A middle manager's product is four artifacts, and three of them are projections of a ledger that could simply be visible.

Flattening has a computable criterion. A write is scaffolding if nothing outside the firm reads it: no foreign key to a counterparty, no filing, no payment rail, no delivery, no still-active human seat. The irreducible output of a case is the minimal set of writes that satisfies every external reader. The flattening dividend of a wire is one minus the ratio of irreducible writes to all historical writes, and it is computable from the schema and the change stream before any model runs. On the synthetic firm's schema it is between 0.61 and 0.77 (measured, planted inputs). That is the fraction of the headquarters that exists to feed skulls.

What one skull carried. A human at a decision held three things at once: world knowledge (how invoices, disputes and contracts work), the firm's own dynamics (who answers what, when, and what follows), and authority (the right to commit). The design splits them. The rented model's pretraining holds the first. The tape holds the second, learned in the gym and owned. The licence table holds the third, deterministic and signed. Three engines, three clocks, three trust regimes for what one skull used to carry, and that is the exact sense in which a language model already contains the compression of everything: it holds the world's knowledge, the firm learns its own dynamics, and authority is a table.

The change of basis. The headquarters has five hundred seats and twenty-four rooms. The context a decision needs is a property of its class, not of the person who held it: which fields decide a credit check are the same for every credit check. So seats collapse to classes, and the context a class needs is one template, compiled once from the fossil the people left behind (the screens they opened, the joins in the query log, the validation rules) and tightened against arrived outcomes. Five hundred vantage points become two dozen templates, and the templates are the whole of what the seats held that was worth keeping. The dependency structure is kept; the workflow, who hands what to whom, is deleted.

The stack ends at the session layer. The seven layers of networking served the transport of context between skulls. In the flattened firm the physical layer is the card, the link is the tape, the network is the lanes, the transport is the fold, and the session is the cognitive index, the cell kept read. Presentation and application are gone, because those two layers were the interior: screens were presentation for eyeballs and applications were interfaces for people. Switchboarding cognition means the stack stops there.

III · THE PREMISE, EXACTLY

At every moment in which any employee makes any decision, a frontier language model given the same context, the same records the person had open, the same document, the same role, decides at least as well. That is the premise, and its exact form is stronger than its slogan.

Two public datasets in which the recorded decision is predicted from the recorded field say that the incumbent's judgment is mostly reproducible: on one, 0.716 of the decision is predictable and the headroom from the organization's own record beyond that is about zero; on the other, the headroom is 0.066, of which 0.061 is a single external score (measured). On the recorded field the incumbent's judgment is reproducible, and the bar the machine must clear is the incumbent's realized rate, which the incumbent's own record shows to be ordinary. The synthetic twin says the same from the other direction: with both arms reading the same physics, the machine is within a few points of the people it replaces on correctness (good rate 38.06 against 42.70; wrong 11.94 against 8.40; late 50.01 against 48.90; measured, planted physics).

The exact claim is therefore this: at the same context the machine is within a few points of the incumbent, and on everything else that decides an outcome it is categorically different. It reads the whole record every time. It never forgets yesterday. It does not cut corners at demand 1.5, when the people are drowning. It is present at every cell at the moment the row lands, and it costs a few dollars an hour. In the twin, that presence drained more than twice the incumbent's settlements, ended the year with a sixth of the incumbent's open backlog, and returned a total cost 583 million dollars lower at the writ's own price for work left undone (measured, planted physics). The money was never in the cleverness. It was in the queue. Presence is the mechanism, and presence is a property of the harness, not of the model.

Two further facts about the correctness gap put the last of the debate to rest. Buying every missing determinant, moving every fact that reached a human only by telephone into a system of record, closed a quarter of the gap; the rest is the judge's own read noise, a property of the

model that closes with a calibrated one (measured). And the premise needs the model to be there, with everything read, and it needs the people who are slower than it to stop being the bottleneck. Both are already true.

What the premise leaves to the design is the generalization: how to stand the model at every point of the firm's core wire without hand-wiring each one. The answer is that the points are not wired. They are classes, the context of a class is a template, and the template is a fold.

IV · TEXT-TO-VIDEO, AND THE VIDEO THE FIRM ALREADY IS

A video used to be made by a pipeline: a camera, then a compositing tool, then a modelling package, then a renderer, then post. A model trained end to end went from words to output in one step and never learned the compositor. Nobody built an agent that imitates a compositor. The pipeline was a fact about the limits of the tools, not a fact about videos.

The interior of a corporation is that pipeline. The meeting, the deck, the ticket, the escalation, the weekly rollup, the manager whose product is four artifacts: these are the steps between arrival and committed effect that exist because context had to move between skulls. An organization run on artificial intelligence does not learn them. It maps arrivals to committed effects with no interior, and what replaces the interior is a fold. A field over every open promise picks the cell that needs attention; a template compiled once per class renders the context that class needs; attention inside the model picks the tokens that matter; one budgeted fetch covers the residual; and every fetch the plane could not satisfy tells the plane where to grow. The middle management of every organization was that fold done by hand.

The analogy runs deeper than a pipeline. Text-to-video skipped its pipeline because a pixel is a free, dense verdict: a billion videos are a billion supervised targets and a human eye grades a frame in a hundred milliseconds. A corporation's committed effects are graded by the world on the branch it took, at the rate the world answers. The design does three things with that fact, and each is a section of this paper. It manufactures a dense grader for the interior, the synthetic firm with planted truth, so the instrument itself is graded in seconds (§XIV). It turns the world's verdict rate into a design variable with four dials (§IX). And it notices that the firm's state, as distinct from its verdicts, arrives every tick, so that the model learns at video rates even while its authority is earned at verdict rates (§X).

And the firm is already a video. Its state is a lattice of promises against seats and time; one frame per tick; promises as sprites; the field as intensity. Read that way the three engines of §VI are image primitives over one canvas. The solver is a convolution: contention is local, so

every contradiction is a stencil over a cell and its neighbours. The event model is masked diffusion over the frame: decode every open cell in parallel under law masks, remask the uncertain, decode again. The resident's glance is attention over patches. The organization is a video of its own lattice, and the event model is its action-conditioned next-frame predictor.

V · THE TWO PLANES

Everything the enterprise has is on one thin data plane. Pressed against it, as close as the hardware allows, is a second plane: the artificial-intelligence layer that reads, writes and acts on the first. Closeness is the point, because every intermediate process, every human touch point and every round trip lives in the gap between the two planes, and closing the gap deletes them. Here is the geometry that makes as close as possible a measured quantity.

The unit is the promise, not the table row. An obligation the firm owes to something outside itself: an order to fill, a claim to adjudicate, an invoice to chase, a hire to close, a filing to make. A firm the size of the synthetic one has on the order of ten thousand open at once. Everything else on the disk is either an input to a promise or archive.

The commitment plane is a lattice. Every open promise is a fixed-size row, 32 bytes of scalar state and a 128-byte integer embedding with a 4-byte scale, 164 bytes in all (measured layout), on a lattice whose cells are a seat at a slot. A promise occupies a run of cells; deadlines are horizontal lines; the verdict horizon per class is a vertical extent. Ten thousand promises at 164 bytes is under two megabytes of state, and the whole open set lives in the graphics card's memory permanently.

The locality law. Put on the plane what contends, and nothing else. Two promises can conflict only if they contend for something, so contention is local in the lattice, and every contradiction the solver finds (a cell overcommitted, an entity double-booked, a dependency started before its predecessor can finish, a deadline missed) is a stencil over a cell and its neighbours, never an all-pairs scan. Identity is local in the entity rather than the lattice, so the duplicate check (one promise recorded in two systems of record with no shared key) gets its own bucketing by entity, its own stencil over the embeddings, and a confidence, because it is the one contradiction a model judges rather than a query certifies. Two stencils, two keys, both microseconds. This is why the plane is a lattice and not a graph: a lattice makes the solver a convolution and the glance a picture.

Two consumers, one layout under a permutation. For the solver the plane is seats by slots, because capacity is per seat per slot. For the mind the same lattice is read as class stage along the wire by time, because that is the order a promise moves in and the order a glance should show. The field is the height of every cell: how loaded the class is, how close its deadlines are, the shadow price of its column, whether its placement is decided or contested, its residual against the class baseline, its contradictions.

The field is a standing query. Every tick the solver runs an entropic transport in the log domain over every open promise against every seat and stock, with finite capacities and an explicit column for work left unplaced that has a real capacity and a real price, and it never materializes the plan. What it keeps are the duals: one number per promise and one per column, in memory proportional to their sum rather than their product. The column duals are prices, what one more unit of that seat or that backlog is worth to the firm this period, nonzero exactly when the capacity binds. From the same solve come the support spectrum (the exponential of each promise's row entropy, the effective number of places it could go, which separates the decided placements, the automatable mass, from the contested ones, the residual that earns a frontier call), the breach forecaster (a column's price times the inverse of the promise's slack, a by-product of a pass already run), and the contradictions, ranked by the price of the column they sit in. Storage in eight-bit integers, arithmetic in single precision with a fixed order of summation, no atomics, no fast math: two runs are byte-identical (measured). This is stencil relaxation, the oldest workload in scientific computing, and it runs in microseconds over tens of thousands of cells on the same cores that will later run the model. It decides nothing. It ranks, prices, and tells the loop which few thousand of the open promises deserve a mind this period.

The cognitive index. The hot set, the cells the field ranks highest, is kept already read: the class template renders the cell's canonical context once, and it is held as the model's own attention pages, keyed by tape position, template pin and judge hash, reference-counted, copy-on-write, evicted by the field's duals. The store's bytes are the model's tokens and the store's index is the model's read state, and a warm cell is judged by a zero-copy fork with no re-encoding. This is the literal form of the AI layer pressed against the data layer, and it is why the design runs on hardware the firm owns: an owned card reuses its attention state, and a rented endpoint asked to reuse its own cache re-encodes the entire context (measured, 516 tokens, image included).

Closeness has three floors. Tokens re-encoded per judgment: zero for a warm cell, by the cognitive index. Hops per effect: one, the gate, because there are three verbs against the store (subscribe, query, write through the valve) and no application tier. Representations per fact:

two, the row and its rendered frame, because the alphabet of typed rows is the machine's language. The interior of a skull-run firm scores badly on all three; the design holds all three at their floors.

The residency arithmetic decides the card (measured constants on the resident: 17 KB of attention state per token and 52.7 MB of recurrent state per sequence):

$$\text{warm cells} \approx (\text{VRAM} - \text{weights} - \text{scratch}) / (52.7 \text{ MB} + 17 \text{ KB} \cdot \text{tokens per cell}).$$

At 500 tokens per cell that is about 160 warm cells on a 16 GB card with a 6 GB resident and about 2,200 on a 141 GB card (budget). So the hot set is the field's top-K, and thinness is what makes it big enough: every field ablated out of a template is more cells held read. One resident card per thousand warm cells; the rest of the plane is priced by the field and read on change.

The archive plane. Mail, documents, chat, filings, the firm's past and its counterparties' words, at a hundred million to a billion tokens, sit on a second plane beside the first. They are joined to promises by registration (deterministic keys first, probabilistic merge second, a model adjudicating only flagged merges, every merge journaled and reversible) and reached by exact scan under a pinned index: two million fragments scored in about two milliseconds, resident on the card, bit-identical across runs (measured), so that a question like which of our procedures now conflict with the regulation that changed this morning is answered exhaustively rather than by sampling. The archive is scanned; it is never held. The embedder has four jobs: registration; novelty, a new document's residual against the field, from retelling through routine and related to bridging and novel, so that novel routes to the frontier and routine stays on the card; the bounded fetch, a phrase in and a read plan out, radius-ordered under the token budget the margin sets, with a stop rule, its spans located in the tape by deterministic code; and the unkeyed duplicate stencil. The same 128-d int8 vector serves the seat fit in the transport, the duplicate stencil, the coarse tier of retrieval and the novelty residual. One vector, four consumers.

The tape under both. Every event in the life of the firm, an arrival, a fetch, a proposal, a hold with its margin, an escalation, an effect with its inverse, a receipt, an outcome, a tick, is an append-only, hash-chained forty-byte record, and every table in the system, both planes included, is a deterministic fold over the tape. A hold is a row, because an organization fails by omission and an omission that leaves no trace cannot be graded. The plane is a view of the tape. The tape is the proof. A tape is a scan, not a parse: two million rows are eighty megabytes and a fold over them is one pass.

The four globalizations. SABRE, the first airline reservation system, made state global in 1960: one true record per seat, readable everywhere, written through one machine, so that overbooking became a policy someone chose rather than an accident that happened. This design does the same to three more things. The port makes judgment global: one reader that

holds every open promise, every proposal a row with a margin. The gate makes authority global: the interior was policy smeared across people, and the licence table is that policy as rows the writ can set. And the counterparty plane makes the verdict global: the world's answer as a row at commit. A firm is finished when its state, its judgment, its authority and its verdict are all rows on one tape.

The frame the model sees on any promise is assembled by four things of four different kinds, which is why there is no regress of a mind deciding what a mind should read. The field, a deterministic solve, picks the cell. The template, a pure function compiled once per class from the fossilized screens and tightened by ablation against arrived outcomes, renders the neighbourhood. One bounded fetch, a breadth-first walk over the registered graph with a depth cap and a token budget the gate sets from the margin, covers the residual. Attention inside the model picks the tokens, and attention is free. The regress bottoms out in a softmax. The glue is a fold.

. . . .

VI · THE ENGINES AND THE SEAM

Three models behind one port, one deterministic kernel, one hand, and one rule that separates them.

FUSOR: the resident at the edge, which says when. Nearly all software that uses a language model today is a function: called, it answers, it forgets, and between calls nothing is running. A function decides nothing about when, because the timing was decided by whoever called it, and it notices nothing, because noticing means being present when nothing was asked. FUSOR is the opposite, and it exists as one C++ process on one consumer card: a tailer reads every word of the world's stream onto a single shared attention state; the end of a completed thought is read off the model's own arithmetic at no extra cost (0.97 on a finished clause against 0.02 mid-phrase, measured); at each boundary a judgment is probed on a zero-copy fork of the live state, 44 milliseconds on a dedicated card (measured), which is the measurement that makes the card dedicated; the margin between speak and hold is the difference of two logits, no bias; a line is cut mid-word when the world answers first and the counterfactual is written to the record; every silence is a row; the switch that turns its output on is a file the kernel can read and never writes; and there is no act verb anywhere in it. In the organization, FUSOR sits on every lane where the world arrives in words, the mailbox, the call, the counterparty's reply, the contract, and turns them into typed rows on the plane at reply speed. The same architecture pointed at the plane instead of a spool is the warm-cell

judge: a resident model holding class templates already read, judging cells as they change, one fork per cell. Several seats share one trunk: a new seat forks onto it at zero memory cost and three co-decode at about a fifth over the cost of one (measured).

The resident is sighted. The weights are early-fusion sighted, and the projector attached to the disposition-tuned mind changes its text answers by nothing, so the eye and the judgment share one skull (measured). A glance is a priced act: the neighbourhood of the lattice rendered as an image tile, a few hundred tokens for a few thousand cells, filed after the line it illuminates. The building is photographed every tick for the price of a paragraph.

The event model: the firm's dynamics, which says what next. The tape at commit grain is a sequence over a few thousand typed tokens, the schema is its tokenizer, and a small model trained on it learns this firm's dynamics. It is a masked-diffusion model over the alphabet: it decodes every open cell in parallel under law masks, remasks the low-confidence cells and decodes again; capacity and conservation enter as the doubly stochastic projection whose duals are the prices; its output per cell is a distribution over the closed verb set plus a placement path with times. It carries two heads on one trunk: a dense head that predicts the next state of the plane under the firm's action, supervised every tick by the world's rows, and a sparse head that predicts the outcome, supervised once per horizon by verdicts. A probe head reads the deterministic field back off its hidden state, so the field and the model's belief are compared every tick. The hand-built lattice is the event model with fixed weights: it stays as the null the event model beats and as the floor that runs the wire alone when every model is dark.

The shape of the driver. The open driving models settled the architecture the event model borrows: a vision-language trunk kept general, a perception head that produces an explicit inspectable grid, and a planning head that produces a short-horizon trajectory by flow matching over the grid, trained by a staged recipe that keeps the trunk's language. For the firm the perception head is deleted, because the state is already a grid the firm owns. The trunk, autoregressive and sighted, is the resident's shape: it reads mail and contracts and glances at the lattice. The planning head, parallel over a grid and refined by iterated denoising toward a plan, is the event model's shape: it decodes placement paths over the lattice under masks. One architecture, two organs, and the perception the driving models spent most of their engineering on comes free.

The frontier model: the judgment on the residual, which says what. Every period, for every promise the field selects and the resident marks thin or the memory marks novel, the loop assembles a frame and hands it to a rented frontier model behind the provider seam: one call site, a signed price book, escrow before dispatch, a hard budget stop, a fingerprint with an expiry on every output, quarantine and never a silent drop. It returns a proposal: the decision, a direction, an estimate of how complete the record is, and a sensitivity per shared fact. The

frontier appears in four roles and no others: the judge on the residual (thin margins and novel cases), the mouth's counsel at the edge, the compiler at setup (which reads each class's screens, query logs and policies once and writes the template), and the juror, an independent weights family voting on irreversibles. The counsel memo keyed on the promise and the frame hash means the same aged case is never re-rented for the same frame; on the synthetic firm that memo is the difference between 986 rentals a day and the residual's true count (measured). The copilot shape, a model on every event, is unrepresentable in this design, because there is no port for it.

The kernel: the gate and the governor, which say whether and how much. The gate is small, deterministic, and compiled as its own library with its hash on every decision it authors, so that nothing learned authored this is a fact a build verifies. It reads the proposal (its direction, its sharpness against the class's usual spread, and a conformal novelty score computed from the pass and the field rather than from the model's self-report), the licence for the class, the reversibility of the effect (derived from the schema, never declared by whatever wants to act), the stratum the governor drew, the budget, the outstanding exposure against the writ's cap, and the switch, and it emits exactly one verb, hold, act, draft, frontier or warrant, with a reason from a closed vocabulary of fourteen, in a published order of refusal in which the budget is checked last and exposure after it, so that running out of supervision produces a hold and never an act. Commit reads live state, never the frame snapshot. The governor holds the salt for every canary and audit draw in a file the machine cannot read, keeps the licence table, keeps a retained fraction of every class routed to a person, and keeps the one meter that decides a class the same day it moves: κ , supervision minutes created over supervision minutes removed.

The hand: a single writer. Every effect leaves through an adapter that records the inverse first and mints the receipt itself, with the idempotency key derivable by any fold and receipt statuses executed, refused, failed, unknown and reconciled. An act passes through seven states, each a row: proposed, authorized, reserved, submitted, confirmed, outcome pending, outcome observed. The firm requested, the external system performed, and the obligation was discharged are three rows. Outstanding exposure, the value of unattended effects whose verdict has not landed, is a fold with the writ's cap, checked after the budget, so that risk in flight is bounded and the licence follows the cap.

The seam rule. The kernel owns when, where, whether and how much. The model owns the map from the open alphabet of language to the closed alphabet of typed rows and verbs. The seam between them is a typed row carrying a scalar margin, a provenance, and three hashes: the event schema's, so that the synthetic world and the real mouth emit identical rows and the kernel cannot tell which world it is in; the model's, so that a rented judge is a fingerprint with an expiry rather than a name; and the template's, because the model's

calibration is a property of the context it was given. The kernel never trusts a margin raw. It measures, on a shadow tape, how often this model's confidence at this level on this class turns out right, and the gate consumes that curve. That is what lets a model be swapped without losing the evidence, a better model next year re-earning its rungs on the same tape, and it is what lets the kernel be trained without a model at all.

Unary from the pass, relational from the lattice. What comes free from a forward pass is unary, a property of one token's distribution: where a case ends, whether to speak, what surprised it, how flat the decision is, and which spans of the frame attention rested on, which go to the tape as the reason. What is relational, which mail belongs to which promise, which rows contradict, which seat binds, which promise depends on which, comes from the solver's stencils and the memory's scans, deterministically, and is handed to the mind inside the frame. The mind is never asked to compute a relation it cannot see. Attention is the last fetcher; the field and the template are the first two.

The five kinds of component, and the heart that consumes a curve.

KIND	COMPONENTS	HOW IT IS MADE	WHERE
computed, never trained	the fold; the transport and its duals; the two stencils; the gate; the ladder	written once, deterministic, byte-identical replay	one binary, gym and product
searched	the writ's rates; the read and fetch budgets; the templates' field sets; the structure; the automation order; the tick period	policy search across ten thousand synthetic firms per card with pessimism across worlds; a value head trained on the rollouts the search discards	the gym
trained as a sequence model	the event model: dense on next state, sparse on outcome	pretrained on synthetic tapes where the counterfactual exists; calibrated on the real tape with observed actions only	the gym, then the shadow tape
rented and attached	the prose model at the edge and on the residual	behind the port, on a lineage stratum at rung one, with a calibration curve measured per class	the product
never trained live	everything	weight-events ratified with goldens; the wire never learns at the point of decision	the product

During the search the heart is trained against a stand-in judge whose competence is swept, so the heart is robust to how good the real judge turns out to be. At runtime the gate consumes the model's calibration curve, keyed by schema, template, class, band and judge hash, never the model. The heart consumes a curve, not a model. That is why a language model is

swapped without retraining the heart, why the heart trains with no language model present, and why the flight computer's per-cell head, whose one leaf runs on a CPU thread or a GPU thread unchanged with one thread integrating one whole universe, is the smallest existing form of the partition.

Plasticity. A prompt (seconds, weak); a steering vector (milliseconds, reversible, hash-pinned, logged per token, carrying disposition and never content); an adapter (runtime-swappable, per class); an audited weight-event (versioned, golden-gated, tape-rederivable). Vectors explore, adapters stage, weight-events ratify. Steering vectors carry disposition per seat, vigilance at the edge and register with a counterparty, revocable at the next boundary, and are never applied to the judge. A new model enters as a candidate on a lineage stratum, reproduces every golden first, is scored on the held-out question backlog, is licensed per role, and re-earns its rungs on the same tape.

The edge is resident because the world is, and the center closes on the tick. FUSOR runs continuously where the world speaks. The judgment is the event: the warm-cell judge fires when a cell's frame hash changes or the floor sample draws it, and carries its last proposal forward otherwise; nothing polls the judge and nothing clocks it. The books close on the tick: the transport solve, the read budget, the gate's verdicts, the strata, the ladder, κ , the exposure and the close are computed when a TICK row folds, from rows only, so that the machine arm is byte-identical across runs, the fold equals the ledger, the machine reads the day and never the clock, and every effect re-derives from rows (all four measured as oracles in the battery). Continuous does not mean busy. A read budget with a memo keyed on the content of each cell's frame, never on the clock, is what decode-on-delta means at the center: the model is asked about a cell when the cell changes, and about nothing else. When a fact shared by a whole class changes, proposals whose sign and band survive are carried across by a certificate without a re-read; on the synthetic firm at E0, 2,062 frames moved by a shared fact alone, 789 were carried, and 1,273 were read again (measured).

.

VII · THE GYM: THE SAME CODE, TEN THOUSAND FIRMS PER CARD

The kernel's language-free half has a second life, and it is the reason the simulator exists. The period function, the fold of the plane, the field's solve, the gate, the governor, is written once as a function that runs identically on a CPU and inside a GPU thread, and a parity verb proves it. In the gym, one thread integrates one whole firm through that function; every firm is regenerated from a hash of its seed so that nothing is stored and nothing moves between cards

but gradients; and a graphics card that holds a quarter of a million resident threads holds tens of thousands of firms per launch at the search grain, and thousands at the fidelity of the five-hundred-seat synthetic firm on a warp per firm, with the twenty-four decision classes mapped onto thirty-two lanes.

Three laws fall out of the thread. The firm is fixed-shape, a dynamical system over cells under masks, never a script with a branch on what a person decides, which is exactly what a masked-diffusion event model already is. The state is small: a few hundred cells and a few floats, or the second grain. And there is one leaf for both grains, so the gym and the product never diverge. The flight computer is the existing form: one render with three callers, drive (render the live period once, and the next state is the plan), foresee (render thousands of sampled futures, rank them by the writ, and the best future's first move is the pipper), and train (when reality arrives, render backward against what actually happened), all three calling one leaf that runs on the host thread and the device thread unchanged, with the multiverse ranking structure moves identically on both (measured).

What the gym searches is what no human should hand-tune: the structure of the firm (which classes, which span, which lanes to instrument), the writ's rates (the canary rate, the retained stratum, the supervision brake, the ask band), the read and fetch budgets, the templates' field sets, the tick period, and the order in which to automate. What it scores is the writ's cost over a distribution of firms, spans, class tables, verdict latencies, arrival rates and hidden rules randomized around the fit, with a pessimism term, because the only policy worth having is the one robust across worlds. A value head trained on the rollouts the search discards ranks candidates so the tree can be wide without rolling out every leaf. Search is the teacher; the tuned kernel is the student; the world is randomized; and the plant is kept outside the policy's reach by a hard-coded guard.

The event model is born here. It pretrains on tapes with the counterfactual, the same cell under different actions across seeds and worlds, which no incumbent's history contains and which the multiverse generates by construction; then it calibrates on the real shadow tape with observed actions only, and the bridge between the two is a held-out test on the real tape's next rows. The kernel trains here language-free against the stub judge. Then the real model is attached behind the port, on a lineage stratum, at rung one, and the same binary that trained is the one that runs. The only thing that changes on the day a real firm is plugged in is the header line that says which world stands behind the port.

The gym's product is a ranked portfolio of designs, each with its curve across worlds, handed to the wire, and every patch it proposes is deployed after a paired backtest on arrived outcomes under common random numbers beats doing nothing, carrying its inverse on the record. The owner's seat shrinks to three things: the distribution of worlds, the writ, and the

signatures. Everything else runs unattended, graded by oracles with planted lies, held-out worlds, and paired backtests. That is artificial intelligence solving for artificial intelligence, with the exteriority preserved.

VIII · AUTHORITY: THE VERDICT LAW AND THE LICENCE

Every organ above is a corollary of one law, and the law is the safety argument and the economic argument at the same time.

Authority is a deterministic fold over exterior verdicts on the system's own executed choices, drawn on a sample the system cannot predict. It narrows on any such verdict and widens only inside a rule the system did not write. It grows no faster than those verdicts arrive.

The first clause says what counts as evidence for widening: a verdict from something the actor cannot write to, on a choice the actor actually executed, on a sample drawn by a salt the actor does not hold. Agreement with a person is not evidence. A forecast is not evidence. A rollout is not evidence. A replay of choices the actor did not make is not evidence. A self-report is not evidence. A certificate is not evidence. The second clause says that narrowing needs no permission and widening needs a signature: a person signs a rule once for a whole class of work, the class, the bar, the sampling rates, an expiry, and inside that rule outcomes do the widening, which is the mechanism that converts a signature from a per-instance approval into a durable delegation, and the reason the system's autonomy compounds instead of consuming attention forever. The third clause is the rate, and §IX is about nothing else.

The ladder. Authority is held per class of decision and per margin band, never by the system as a whole, and each class climbs: watch only; act on a small stated fraction of reversible cases; act on the specific shape approved; act on cases of the same shape; act with a sampled fraction shown for review, never falling to zero. The test that moves a class up is anytime-valid, so the owner looks at every arrival without invalidating it, which is how supervision actually happens. The incumbent's good rate in the band, $\bar{g}$, is measured on the shadow record from at least two hundred outcomes and frozen for the licence's term at its upper confidence bound. The writ's tolerance δ sets the bar: the machine is licensed against the hypothesis that it is worse than the incumbent by more than δ on the failure side, $g \leq g_0 = 1 - (1 - \bar{g})(1 + \delta)$. The evidence process over the machine's canary outcomes in arrival order,

$$E_n = \prod_i (\bar{g} / g_0)^{X_i} \cdot ((1 - \bar{g}) / (1 - g_0))^{1 - X_i},$$

is a supermartingale under every rate in the null, so by Ville's inequality the probability that it ever crosses $1/\alpha_k$ under the null is at most α_k , at every look. Rung k is reached when $E_n \geq 1/\alpha_k$ with $\alpha_k = \alpha \cdot 2^{-k}$ and $\alpha = 1/20$; the process is never reset within a term, so the chance of ever reaching any rung falsely within a term is at most α in total (derived). A mirrored process runs on every machine-executed outcome in the band, not only the canaries, against the hypothesis that the machine is as good as the bar, with an easier threshold ($\alpha_d = 1/5.5$), and when it crosses, the band drops one rung the same day. Every licence carries a term and is kept only if the term's process reached its rung again. A detector on the statistical character of incoming work drops affected classes to watching immediately, because outcomes arrive too late to be the first line of defence against a regime change. And a fraction of every licensed class is routed to a person forever, to keep the comparison alive and to keep the people practised.

Two graders, with disjoint support. Partition the firm's history by whether the machine's proposal matched the firm's recorded decision. On the agreement band the executed action was the machine's action, so the recorded outcome is an exterior verdict on a choice the machine would have made, and history licenses that band in the time a replay takes. On the disagreement band the machine's alternative was never executed, so only the world licenses it, through canaries, one verdict latency at a time. On the synthetic firm the replay admitted 32 bands from history and moved 29 past rung zero, with the machine agreeing with the people on 87.2 percent of 13,235 shadow cells (measured).

Verdict installments. The world's verdict arrives in pieces that are each exterior: a counterparty's reply time, an acknowledgment, a partial payment, a scan event. Each is an observation on an executed choice from something the actor cannot write to, so each counts under the law once its relation to the final verdict is calibrated on the shadow record. Installments shorten the latency of every class whose world answers in stages, and they are the bridge between the two clocks of §X.

Pooling. The identifiability floor divides by the number of tenants sharing a schema, a judge and a salt. Two tenants pool a band when they share the schema pin, class, band, judge hash and template hash and their band-level outcome rates are exchangeable over the trailing window; each tenant's own demotion process keeps running, so a pooled promotion widens a tenant whose own evidence agrees. The licence key already carries the schema hash, the class, the band and the judge hash; pooling is a key on the record, not a model.

Exposure and readiness. Outstanding exposure caps the value of unattended effects in flight, checked after the budget, and the licence follows the cap; on the synthetic firm at E0 a cap of 100 million kept 63.5 percent of decision mass licensed, 50 million gave 29.4, and the cap

is a dial the owner signs (measured). Each domain carries a readiness gate stated in its own terms.

What a simulation may mint, said once: a synthetic firm measures the estimator's bias and ranks structures under planted truth; a fitted firm forecasts and prunes; a replayed real history licenses the agreement band; the real world licenses the disagreement band, because it is the only exterior.

IX · THE FLOOR, AND THE ARITHMETIC OF NINETY PERCENT

This is the section that contains both the number and the four dials that move it.

The floor is identifiability at the tolerance the writ writes down. A class is licensed when the anytime-valid test rejects the hypothesis that the machine is worse than the incumbent by more than the writ's tolerance. For a base failure rate p and a relative margin δ , the number of outcomes that takes is

$$n_0(p, \delta) = z^2 \cdot p(1 - p) / (\delta \cdot p)^2, \text{ with } z = 2.487,$$

which at $\delta = 0.30$ is $68.7 \cdot (1 - p) / p$ (derived): a class that fails eight percent of the time needs about 790 world verdicts before its licence widens one rung; a class that fails two percent of the time needs about 3,370. The wager is fed at the canary rate r , the class produces V decisions per term, and the verdict arrives one latency L later, so the days to the first rung are

$$D_c = L_c + n_0(p_c, \delta) / (r_c \cdot V_c).$$

Nothing in the machine moves that number: not a better model, a bigger card, a thinner plane or another lane. What moves it is the four quantities in the formula, and each is a dial.

Dial one: tolerance. δ is a harm dial the owner signs, and it enters squared. Raising it from 0.30 to 0.40 on the synthetic firm's schema took licensed decision mass from 63.7 to 80.1 percent and unlicensable bands from 26 to 12 (measured).

Dial two: pooled volume. V grows by pooling outcomes across firms that share a schema and a judge, under the verdict law: the same judge, its own executed choices, a salt it does not hold. Multiply the volume by ten and n_0/V falls from 0.55 of a term's volume to 0.055, licensable inside a term (measured on the synthetic schema). The second client is the instrument, and that is why this is a platform.

Dial three: verdict installments. L falls when the world's answer is counted in pieces. A class graded in a quarter by a final settlement is graded in a week by the counterparty's acknowledgment and in a day by its reply time, each an exterior row calibrated against the final verdict. Installments cut L without touching δ , coverage or V.

Dial four: counterparty planes. L falls to a tick when the counterparty is itself a plane. A rail confirms a payment in seconds; a carrier scans a delivery; a marketplace accepts or rejects an order; another firm running this design writes the outcome row the moment the effect meets it. The order of preference is the firm's own tenants and subsidiaries first, its chosen suppliers second, its customers where the product is itself a plane. Every pair converts a verdict latency of a quarter into a latency of a tick, and turns the slow sparse grader into the dense one that text-to-video had.

The order rule. Before any wire is automated, compute D_c for every class on the enterprise's schema from its own history: the verdict latency, the base rate, the arrival rate, the canary cap. Automate the wires in order of D_c . The fast-verdict wires license first and their licensed mass accrues from the first quarter while the slow wires are still in shadow, so the enterprise's licensed mass is a curve that rises from day one rather than a cliff at the end. Coverage, moving every fact that reached a human only by telephone into a system of record, is a correctness lever and a cost lever: on the synthetic firm it took coverage from 0.913 to 0.960 against a planted 1.000 and closed a quarter of the correctness gap (measured).

The headcount, worked at the synthetic firm's schema, every step a budget on measured and planted inputs. The firm's cone is 412 decision seats, 70 supporting seats at a support coefficient of 0.145, and a fixed floor of 17.6 seats, the legal shell, the close and the writ's author, for a total of 500. The cascade is $N = E/(1 - \alpha) + F$, and licensing removes decision mass: $E' = E \times (1 - \text{licensed mass})$.

STEP	WHAT HAPPENS	DECISION SEATS	SUPPORT-ING	FLOOR	HEAD-COUNT	COMPRES-SION	THE DIAL
the interior deleted	coordination becomes a fold; α goes to zero; the α -functions vanish	412	0	17.6	430	14%	none: a fold, no model
the draft loop	minutes per settlement fall from 297 to 125 at parity: the same work on 42 percent of the decision minutes	173	0	17.6	191	62%	none: presence
the agreement band licensed and compiled	63.5 percent of decision mass leaves; the keyers go with it	63	0	17.6	82	84%	history
tolerance 0.40	80.1 percent licensed	34	0	17.6	54	89%	dial one
pooled volume	the disagreement band clears n_0 across tenants; about 90 percent licensed	17	0	17.6	37	93%	dial two
installments and counterparty planes	the slow wires' latency falls from a quarter to days; the residual is the warrant and counterparty seats	12	0	17.6	32	94%	dials three and four
the close as a fold	every effect carries its inverse; the reconciliation seats leave the floor	12	0	12	26	95%	the tape

The retained stratum, a few percent of decisions routed to a person forever, is one or two seats inside the residual decision count, and the seven warrant seats are inside it too. The interior deletion alone is a seventh of the firm and needs no model. The draft loop, with no licence and no permission from the world, takes the firm to a third of its size in the first year. The agreement band, licensed from history in hours and compiled into code as it stabilizes, takes it to a sixth. Ninety percent is reached by the first two dials, and the third and fourth push past it. That is the sentence this document exists to state exactly: the number is set by four dials the owner can name, and the model is the thing the arithmetic shows does not set it.

The cost side is short. Five hundred loaded seats at a headquarters is on the order of seventy-five million dollars a year; thirty-seven is five and a half; twenty-six is four. The machine that replaces the rest is a handful of graphics cards, a frontier bill priced by the residual (on the order of eight hundred calls a day once the counsel memo stops re-renting the same aged case, measured at 986 before it), and a record. The organization's daily judgment, three thousand reads a period over a cached prefix, is under an hour of one card. The machine seat, one resident card per thousand warm cells, never sleeps, and its output is measured by the pair of numbers printed weekly: how much human attention the system consumed, against how much work it discharged without consuming any.

X · TWO CLOCKS: LEARNING DENSE, LICENSING SPARSE

The world grades the organization at the rate of SIX, and that rate is what authority waits on. The world's state is a different clock. Every arrival, placement, discharge, scan, reply and receipt is a row the moment it happens, and every tick supervises the next frame of the video. So the event model's dense head learns the firm's dynamics at plane rate, predicting the next state of the plane under the firm's action, while its sparse head and the ladder learn at verdict rate. The model learns as fast as the plane changes; authority widens as fast as the world answers; and the two clocks share data through verdict installments, which are world rows at tick rate that are also partial verdicts. The dense head's targets are the world's rows under the firm's action, and in the flattened firm the world's rows are the only rows in the stream, so next-row prediction is clean.

Thought compiles into number. Number crunching is the fold: deterministic, integer, replayable. Thought crunching is the judge: learned, proposing, never disposing. The support spectrum measures how much of the firm is decided against how much is contested, and the compiler moves mass from the second to the first as grades stabilize: a class whose margins are wide and whose verdicts agree is compiled into ordinary code, verified by replay against the whole history, with the judged path kept alive underneath as the control. The graduation runs down a tower of shadows: the frontier acts at canary size on reversible promises as volume accrues; a smaller resident model shadows the frontier; compiled code shadows the resident; graduation flows down toward cheaper and grading flows up from the world. The resident's acceptance rate against the frontier, measured on the shadow tape, is the frontier bill a quarter ahead: this is speculative decoding at firm scale, a cheap proposer and an expensive verifier paid only on disagreement. So the model's share of the work falls over time. A firm's intelligence budget is a measured quantity that falls every quarter, and the design's whole economic claim is that curve: presence at the start, number at the end, the frontier bill priced by a residual that shrinks.

XI · THE ORGANIZATION WITH NO INTERIOR: WHAT IT LOOKS LIKE FROM THE BOUNDARY IN

Design a company from the plane outward and the shape that results is the shape of its boundary, because the interior has no skulls to seat.

What is human. The owner, who authors the writ: what the thing is for, what it must never do, what must always hold, and the tolerance it is willing to sign, authored once, never learned, and the only thing above the gate. The warrant seats: whoever signs the filings, carries the liability, and can be sued; a percent or two of the decision mass. The counterparty seats: wherever the other side of a promise demands a person, a handful of decisions a day, shrinking as counterparties become planes. The retained stratum: a rotating few who take the sampled fraction of every licensed class so that the comparison stays alive and someone stays practised. And physics: delivery, presence, the robot's problem, which is outside this document. Everything else in the org chart of a conventional headquarters, every manager, every coordinator, every analyst whose product was a projection, every function whose customers were other employees, does not exist, because the thing it compensated for does not exist.

What is machine. The record, and the two planes folded from it. One resident card per thousand warm cells, holding the class templates already read. FUSOR on every language lane. The solver pricing every seat and stock each tick. The event model predicting the next frame and proposing placement paths. The gate and the governor. Executor adapters into every system of record, each write carrying its inverse, each receipt minted by the executor. The frontier model behind the provider seam for the thin margins, the novel cases, the compile step, and the jury.

The nights. An enterprise's owner is not at the desk every night, and the design has a mode for it that is a setting in the writ file rather than a fork of the design. A licenser process on separate hardware, deterministic, fed nothing but outcomes and the tape, holds the salt and a stake in escrow it cannot raise, draws the strata, runs the ladder, widens a rung only inside the owner's signed rule and demotes on any evidence. A jury of an independent weights family votes on irreversibles and on the retained stratum, and the family that compiled a class's template sits out that class. The deterministic floor, the lattice and the compiled routines with no model, runs the wire alone at rung one when every model is dark: it holds, looks, drafts, and takes safe terminals at due. Escalation of judgment ends in an experiment, the smallest bounded reversible action that yields an outcome, run in the twin first where it touches

physics; escalation of authority ends in a warrant row. The firm holds where it lacks licence, grades its own holds, and prints what it did not do, and in the morning the owner reads the account.

Scale. Inside one decision period's light cone the firm is one writer, one tape, one gate. Define Λ as the round trip across the firm's diameter at signal speed over the period of its fastest globally consistent loop: at the measured 44 millisecond judgment one mind reaches 6,600 kilometres, and at a one-second tick it reaches the Earth and its orbits (derived). Λ is printed on the header row, so the choice of one plane or several is a number. Past Λ near one the shard boundary is the seam: each region a firm with its own gate, licence table, tape extract and writ extract, composed by lane contract and a tape of tapes, with a headquarters model that reads fields and probes only and proposes narrowings only; models never message each other. Past any machine boundary only disposition crosses: a bounded, signed spore carrying the writ, the class table and schema maps, compiled routines with their domains and replay receipts, tuned adapters with their pin tuple, the oracle battery, the outcome receipts that earned each admission, and a lineage identifier. It is compiled deterministically at an epoch under a signed rule and admitted at the new site under quarantine at the canary rung with earned competence and zero earned authority, re-running the battery locally and re-earning its licence on local outcomes. The running instance never writes its own spore. A second site is a locus seeded from the first, and the enterprise's operating disposition propagates across regions as a verified artifact rather than as a transfer of people.

What the org chart is. A writ. A licence table. A boundary roster. A card. There is no third layer, because the layers between the owner and the promise were the interior, and the interior was never the company. It was the context logistics for skulls, and the plane is the context.

.

XII · THE FORWARD-DEPLOYED ENGINEER, REDEFINED

Every enterprise that has deployed artificial intelligence has a forward-deployed engineer: the person sent into the firm to make the model useful. In the industry's shape that person builds an agent per workflow, a prompt per seat, a dashboard per manager, and inherits every handoff that existed because a person needed to look at something. In this design the forward-deployed engineer runs the switchboard of the cognition layer, and the job has six deliverables, in order.

1. **The census.** Turn on change capture on every system of record and land the stream in order on the tape. For each core wire, write the one hand-authored object in the whole

programme: which table, which key, the predicate under which a promise exists, the predicate under which it is discharged, and which columns carry the deadline, the amount and the owner. Ten lines a class, one domain expert per wire, about a week each. Replay the entire history through it, so that every promise the company ever made becomes a record with its duration and its ending. Compute the flattening dividend and D_c per class from the firm's own history. Deliverable: the map, and the order of wires.

2. **The plane, with no model.** Stand up the live plane of every open promise, its state, its deadline, its owner, its blocker, its price, visible to everyone. Retire the things that existed only to move that information between people: the status meetings, the rollup decks, the coordinator roles, and the α -functions that supported them. Measure coordination minutes before and after. This deliverable contains no artificial intelligence and is the largest single reduction in headcount cost in the programme: a seventh of the firm, and the plane on which everything else runs.
3. **The templates.** Compile each class's context once from the fossil, the screens people opened, the joins in the query log, the validation rules, and hand the compile step to the frontier model at setup; tighten each template by ablation against arrived outcomes in the gym. Deliverable: two dozen templates for two dozen rooms, and the cognitive index that keeps them read.
4. **The writ and the rules.** Author the writ with the owner: the tolerance, the caps, the exploration floors, the retained fraction, the exposure cap. Have the owner sign one licensing rule per class, once. Deliverable: the licence table at watch, with its bars measured from the shadow record.
5. **The switch, one wire at a time.** Off (census, no judge, no effect); shadow (drafts diffed against what people do on the same cells, a class promoted when they agree at a stated rate over a stated count); live on the draft loop (the machine proposes on every open promise, people key, and the queue drains: the year-one money); then the act loop on one wire, with canaries at the floor rate drawn by a salt the machine cannot see and the retained stratum forever, and the classes that stabilize compiled into ordinary code. Deliverable: the first outcome row on the tape, which is worth more than every document in the programme.
6. **The second wire and the second client.** The second wire is a configuration change, because the record, the plane, the writer, the folds and the instruments already exist. The second client on the same schema shares the judge and the salt, and its outcomes pool: the second client is the instrument by which ninety percent is reached.

What the forward-deployed engineer never builds: an agent per workflow, a prompt per seat, a chatbot, a dashboard, a message bus between private contexts. The forward-deployed engineer's metric is supervision minutes and κ , never adoption or hours saved, and the deliverable is a calibration record followed by a licence table. One forward-deployed engineer per enterprise, with the gym behind them, is the whole of the artificial-intelligence function; the AI full-time equivalent is a card that never sleeps, and its productivity is the pair of numbers.

XIII · HOW AN EXISTING ENTERPRISE GETS THERE

Six inversions of standard practice, then the sequence, each phase with the compression it delivers at the synthetic firm's schema.

Watch before acting. For the first phase nobody in the company uses any of it, and the first deliverable is a calibration record rather than a tool. **Delete before automating.** The largest single block of removable work in a headquarters is coordination, and it is removed by making one shared ledger visible, which requires no model. **Do not build one agent per workflow.** A workflow is a description of how people did it; the path from a promise opening to its discharge is found per instance and compiled when it stabilizes. **Replay the past first.** The company's own history is a multi-year watching period already on disk, and grading against it has no observation effect, because nobody knew they were being watched when it was written. **Measure supervision minutes and the brake ratio, never adoption or hours saved. Take one wire all the way before starting a second,** in order of D_c , because the second is then a configuration change.

Phase 0 · instrument. Change capture on every system of record; the ten lines per class; the replay of history; the flattening dividend and D_c per class computed rather than asserted. Compression: none yet; the map and the order.

Phase 1 · the shared object, with no model. The live plane of every open promise, visible to everyone; the status meetings, the rollup decks and the coordinator roles retired; coordination minutes measured before and after. Compression: on the order of a seventh of the firm.

Phase 2 · watch, backward and then forward. Run the judge over the last two or three years as if live, then over the present, publishing nothing, the switch at off. Three records accumulate: what the machine would have done, what the people did, and what the world

then said. The agreement band licenses from history; the calibration curve of the model, its margin against its right rate per class, is measured; the bars are frozen. Compression: none yet; the licence table and the calibration.

Phase 3 · the draft loop, then one wire to self-sufficiency. Shadow, then live on the draft loop: the machine proposes on every open promise, people key, and the queue drains; 2.3 times the settlements on 95 percent of the incumbent's minutes at parity within a few points on correctness (measured, planted physics). Then, on the fastest-verdict wire, the act loop: the single writer with every effect carrying its inverse; canaries at the floor rate on every class at once, drawn by a salt the machine cannot see; the retained stratum forever; the brake printed daily; the classes that stabilize compiled into ordinary code, verified by replay against the whole history, with the judged path kept alive underneath as the control. Compression: to a third of the firm on the draft loop, to a sixth as the agreement band licenses and compiles.

Phase 4 · the second wire is a configuration change. Each subsequent wire follows in order of D_c , and installments and counterparty planes are negotiated wire by wire: a rail here, a carrier scan there, a supplier moved onto the schema.

Phase 5 · pool. The second client shares the schema and the judge; the licence key pools their outcomes under the verdict law. Compression: past ninety percent, on the disagreement band, at the rate the pooled world answers, and beyond it as the third and fourth dials turn.

The governing measurement is a pair of numbers printed weekly whether or not anything happened: how much human attention the system consumed, against how much work it discharged without consuming any. Beside them, always, two counterweights that keep the pair honest: the outcome-graded error rate of everything that ran unattended, and an explicit list of what the system is not doing, promises held past due, classes with authority and no activity, counterparties who have gone quiet. A rising handled count beside a growing not-doing list is read together, so that presence and silence are both on the page.

. . . .

XIV · THE INSTRUMENT

A design that describes its healthy state and its instrument together is the one that can be trusted with a company. Four instruments make this one gradeable.

A synthetic firm with planted truth. Every quantity the instrument recovers, the unrecorded mass of every decision class, the competence of every seat, the support coefficient and the fixed floor of the cascade, the judgment intensity of each class, is planted, and none of it is shown to the machine. The machine recovers them from the record, and an oracle checks

that it did. Twenty-eight oracles run in the binary, and twenty-six of them carry a lie: under a lie arm, the mechanism the oracle claims to test is corrupted on purpose, and the oracle passes only if it catches the corruption, because an oracle that only ever passes is measuring nothing (measured). The harness asserts the lied-to oracle's own line, never merely a failing exit.

The twin. Two firms on one seed with nothing shared measures the one number a real deployment can never obtain. A real firm's canary and its control share a ledger: the machine is routed the cases it was licensed for, and the people work a queue the machine already drained. The twin measures how much that biases the estimate: at demand 1.5 the within-firm estimate read 34 points above the true paired effect (measured). Every enterprise that runs this design knows the bias in any vendor's reported lift, because every vendor's lift is measured inside one ledger.

The multiverse. Many firms under edited constraints, sharing one prefix and one physics, reweighted by the writ with pessimism; every patch carries an inverse; a search result is a proposal with a receipt and a paired backtest decides it.

The receipts. Every run prints its commands, its outputs, its exit codes and its denominators, and a list of what was not done. The headline in §IX is the one that survived three rounds of the programme's own review, and it is exact.

. . . .

XV · HOW ARTIFICIAL INTELLIGENCE IS DONE AT ENTERPRISE

SCALE

The pieces of this design have existed, separately, for decades. Firm-as-dynamical-system is industrial dynamics. Off-policy evaluation and the selective-labels problem are standard in economics. Domain randomization is robotics'. Entropic optimal transport, masked diffusion, anytime-valid inference, paged attention with copy-on-write forks, hash-chained logs: each field owns one organ. The assembly needs simulation, machine learning, organization theory and safety engineering in one head, producing no benchmark and no dataset, and it is held together by twelve conditions that every prior design of an organization violates in at least one place.

1. One truth, replayable bit for bit: the tape, and every table a fold.
2. Nothing learned disposes: the gate contains nothing learned; the pass proposes.
3. Authority widens only on exterior verdicts of executed choices on a sample the actor cannot predict, and narrows on any evidence.

4. The glue is a fold: four tiers of four kinds, no second mind.
5. The judge is present, not invoked: resident in reading, closed at ticks.
6. Unary from the pass, relational from the lattice.
7. The floor is identifiability and the ceiling is the verdict rate, with four dials on the rate.
8. One dynamics source per organ; the same leaf in the gym and the product.
9. Every act carries its inverse and a receipt minted by the executor; every silence is a row.
10. One writer inside one light cone; only disposition crosses beyond it.
11. The writ is authored, never learned; hazards are unreachable by construction.
12. The instrument can be lied to and is seen to catch the lie; no organ outlives its null.

DESIGN	WHAT IT GETS RIGHT	THE CONDITIONS IT LACKS
the human hierarchy	2 and 11: a dumb slow board, an authored purpose	1, 4, 5, 7, 9: state smeared across skulls, the glue a meeting, the verdict rate human, omissions unrecorded
enterprise resource planning and systems of record	state made global, SABRE's move	4, 5, 6: judgment left in skulls and screens; the presentation layer is the product
workflow engines and robotic process automation	determinism	3, 4: the human pipeline copied step for step; the interior automated instead of deleted; no licence
copilots, a model on every event	the presence of a model	2, 3, 4, 5: a turn on every event, per-seat billing, no port, the interior kept and fed
agent frameworks and multi-agent swarms	decomposition	2, 4, 5, 12: the glue a model, messages between private contexts rebuilding the meeting, turn-based, self-grading
digital workers	a role abstraction	3, 11: authority by assertion, purpose by prompt; the wrong unit, the employee rather than the obligation
the Toyota production system	9 in spirit (andon: a hold is a row), 4 in spirit (kanban: the field as a standing query), 2 in spirit (jidoka: the machine stops on a defect), the transport (heijunka), the compiler (kaizen)	1, 5, 7: it runs on skulls; the closest human-run predecessor, and this design descends from it
double-entry bookkeeping	9's ancestor: every entry carries its counter-entry	it records money, not judgment
SABRE	one true record, one writer	no judgment, no licence

DESIGN	WHAT IT GETS RIGHT	THE CONDITIONS IT LACKS
closed-loop learning systems (auctions, cooling, routing)	a dense grader and a real loop	7: they live only where the world already answers densely; this design manufactures that world with the four dials and that grader with the gym
this design	all twelve	

The compounding law is the economic argument. Build only what a smarter model makes more valuable. Everything this design owns is in the column that appreciates when models improve: contracts, goldens, an oracle battery with planted lies, the graded record, the licence table earned on real outcomes across real tenants, the question backlog with its graders, the alphabet many firms speak. Everything the industry's shape owns is in the column that evaporates: prompts, scaffolds, frameworks, model-specific tuning. So this design gets better with every model release without a line rewritten: a better model next year enters as a candidate, reproduces every golden, re-earns its rungs on the same tape, and the licence table it inherits is worth more than the model. The shape that works has no meter on it: a deterministic glue bills no tokens, a resident loop bills no calls, and a licence that shrinks the model's share of the work every quarter shrinks the invoice every quarter. It is built by whoever runs a firm and owns its record, for that firm and then for the second, at which point the pooling begins and the thing becomes what no vendor can copy: a licence table earned on real outcomes across real tenants, under a salt no tenant holds.

XVI · WHAT TO DO ON MONDAY

One real lane, chosen by D_c . Turn on change capture, write the ten lines per class, fold the history into a plane, and put the machine on it at off, watching, judging, recording, publishing nothing. Retire the meetings that existed to carry that plane's contents between people, and count the minutes. Replay the history through the judge and license the agreement band from it. Freeze the bars. Turn the switch to shadow, then to live on the draft loop, and let the queue drain. Then the first canary at the floor rate, on a salt the machine cannot see, and the first outcome row on the record. Then the second wire, which is a configuration change. Then the first counterparty moved onto the schema, and the first installment calibrated. Then the second client, which is the instrument.

The interior was never the company. It was the cost of running the company on minds that could not see the same state at once, and that cost is now optional. Flatten it. Keep the edge. Let the world write the licence, and turn the four dials that set how fast it writes. Ninety percent is arithmetic, and the arithmetic is printed above.

.

APPENDIX A · THE MATHEMATICS, COMPACT

The identifiability floor. $n_0 = z^2 p(1 - p)/(\delta p)^2$; at $z = 2.487$ and $\delta = 0.30$, $n_0 = 68.7 (1 - p)/p$. Terms per rung = $n_0/(r \cdot V)$, with the canary rate $r \leq$ the cap. At rungs above three the canary rate is $\max(\varepsilon, 1/\sqrt{(3F)})$ for F outcomes collected, the regret-optimal exploration for a licence that is a bandit over its own evidence.

The days to the first rung. $D_c = L_c + n_0(p_c \delta)/(r_c \cdot V_c)$. The four dials: δ (squared in n_0), V (pooled across tenants), L (installments), L (counterparty planes, to a tick).

The licence test. $\bar{g}$ the incumbent's good rate in the band, frozen for the term at the upper 90 percent bound from at least 200 shadow outcomes; $g_0 = 1 - (1 - \bar{g})(1 + \delta)$. $E_n = \prod_{i \leq n} (\bar{g}/g_0)^{X_i} ((1 - \bar{g})/(1 - g_0))^{1 - X_i}$ over the machine's canary outcomes in arrival order. Under any $g \leq g_0$, $E[E_{n+1} | E_n] = E_n [g \bar{g}/g_0 + (1 - g)(1 - \bar{g})/(1 - g_0)] \leq E_n$, so E is a nonnegative supermartingale and $P(\sup_n E_n \geq 1/\alpha_k) \leq \alpha_k$ (Ville). Rung k at $E_n \geq 1/\alpha_k$, $\alpha_k = \alpha 2^{-k}$, $\alpha = 1/20$, never reset within a term; the union over rungs is $\leq \alpha$. Expected canary outcomes to rung k when the machine is at the bar: $\log(1/\alpha_k)/KL(\bar{g} \parallel g_0)$. Demotion: the mirrored process on all machine-executed outcomes, null $g \geq \bar{g}$, alternative g_0 , threshold $1/\alpha_d$ at $\alpha_d = 1/5.5$.

The supervision brake. κ = supervision created / supervision removed, in minutes, per class and band, per term; a draft costs its review, a warrant its signature, an escalation its reading; κ above the writ's cap with material removal demotes the class the same day.

The field. Entropic transport in the log domain with finite column capacities including the priced stocks: $u_i = \log a_i - LSE_j(-c_{ij}/T + v_j)$; $v_j = \min(0, \log b_j - LSE_i(-c_{ij}/T + u_i))$; the plan $P_{ij} = \exp(u_i + v_j - c_{ij}/T)$ is never materialized; $-v_j$ is the shadow price of column j ; the support of promise i is $\exp(H_i)$ with H_i the entropy of its row; $risk_i =$ the price of its argmax column $\times 1/(1 + slack_i)$; iterate to a published convergence curve and finish on the equality constraint, because the dual identity holds only after a row pass (measured).

The cascade. $E_{out} = \alpha E_{in} + F$; $N = E/(1 - \alpha) + F$; licensing removes mass, $E' = E(1 - \text{licensed mass})$; F is obtainable only from cross-firm data at the target size and is printed beside every N .

The estimator's bias. $\text{bias} = (\text{within-firm canary estimate}) - (\text{true paired effect})$, obtainable only from a twin; +34 points at the synthetic firm's operating point (measured).

The flattening dividend. $S = 1 - \text{irreducible writes} / \text{all historical writes}$, per wire, computable from the schema and the change stream before any model runs.

The frame. $C(i, t) = A[T_c(P, i, t) \oplus W_b(i, t)]$: the class template over the plane, one budgeted fetch, and attention; the frame's hash is a function of the promise's own rows, the template pin and the span contents only, never of the clock.

The read budget. K reads per period: the top K by ripeness and residual, plus the uniform floor sample, plus every cell whose frame hash changed; every other cell carries its last proposal forward; reads per period $\approx$ arrivals + re-touched cells + the floor sample.

The residency. warm cells $\approx (VRAM - weights - scratch)/(52.7 \text{ MB} + 17 \text{ KB} \cdot \text{tokens per cell})$.

The light cone. $\Lambda = (D/c)/T$; $r = cT/2$.

.

APPENDIX B · THE MEASURED STATE

NUMBER	STATUS	VALUE
the twin at 500 seats, demand 1.5, span 7, seed 20260913, at E0	measured, planted physics	good 42.70 incumbent against 38.06 resident; wrong 8.40 against 11.94; late 48.90 against 50.01; total cost 1,837.7M against 1,254.3M; net -583.4M with the backlog priced; the estimator's bias +34.42
settlements and backlog in the twin	measured, planted physics	settled 186,956 against 82,479; open at year end 19,450 against 123,927; human minutes per settlement 297 to 125
the automate run	measured	32 bands admitted from history, 29 past rung zero; κ 0.110; licensed decision mass 63.5 percent; N 500 to 193 (intercept fit) or 182 (origin fit); residual: warrant 4.6, counterparty 0.7, frame 10.0, thin 21.1
reads and rentals	measured	1,729 reads a period, 1,462 carried by the memo; 986 frontier rentals a day before the counsel memo
shadow agreement	measured	87.2 percent of 13,235 cells
the shared fact	measured	2,062 frames moved by a shared fact alone; 789 carried by the certificate; 1,273 read again, 471 changing choice
the exposure dial	measured	100M keeps 63.5 percent licensed; 50M gives 29.4
tolerance 0.30 to 0.40	measured	licensed 63.7 to 80.1 percent; unlicensable bands 26 to 12; headcount 193 to 114
every lane bought	measured	coverage 0.913 to 0.960; licensed mass +1 point; the correctness gap 3.49 to 2.69
the interior deleted alone	measured	+10.5 percent settled, -9.2 percent minutes, -55.8M
the coverage estimator against planted truth	measured	error 0.05 to 0.07, correlation 0.84 to 0.96 on three seeds
the oracle battery	measured	28 of 28 in the binary, 26 lie arms seen to fire

NUMBER	STATUS	VALUE
FUSOR	measured	boundary signal 0.97 / 0.02; judgment 44 ms dedicated; the served endpoint re-encoding 516 tokens; 17 KB attention state per token; 52.7 MB recurrent state per sequence; the fold 14.8×
the exact scan	measured	two million fragments in about two milliseconds, bit-identical
the two public datasets	measured	0.716 predictable, headroom about zero; headroom 0.066 of which 0.061 one external score
the minute ledger; the machine's competence; the tacit channel	planted	a model of the claim
the headcount table in §IX	budget	arithmetic on the measured and planted inputs above, at the synthetic firm's schema

APPENDIX C · GLOSSARY

Agreement band. The decisions on which the machine's proposal matched the firm's recorded choice; history licenses it. **Disagreement band.** The decisions on which it did not; the world licenses it. **Draft loop.** The machine proposes, a person keys; needs no licence. **Act loop.** Unattended action; needs the licence. **Cognitive index.** The cell's canonical context held as the resident's attention pages, keyed by tape position, template pin and judge hash. **Counterparty plane.** A counterparty that emits its verdict as a row at commit. **Dial.** One of four: tolerance, pooled volume, installments, counterparty planes. **Edge.** Where a promise meets the world and the verdict arrives. **Event model.** The masked-diffusion model of the firm's dynamics over the alphabet, with a dense head, a sparse head and a probe head. **Field.** The solver's ranking of every open promise: pressure, price, support, risk, contradictions. **Fold.** Any table derived from the tape by replayable computation. **Frame.** The context handed to the model for one promise. **FUSOR.** The resident kernel at the edge; the mouth. **Gate.** Deterministic code that emits exactly one verb with a reason; contains nothing learned. **Governor.** Holds the salt, the licence table, the retained stratum and the brake. **Installment.** A partial verdict that is itself an exterior row. **Interior.** Everything in a firm that exists to move context between skulls. **κ.** Supervision created over supervision removed. **Licence.** What a class may do, as a rung with a term. **Λ.** The coherence number: round trip over period. **n₀.** The outcomes a class needs before its licence widens. **Plane.** The commitment plane (hot, in memory, already read) and the archive plane (cold, scanned). **Port.** The closed seam: a frame in, a proposal out. **Resident.** The prose model on the owned card: mouth at the edge, warm-cell judge at the center. **Retained stratum.** The fraction of every licensed class routed to a person forever. **Seam.** The boundary between what proposes and what disposes. **Spore.** The signed artifact that carries disposition across a machine boundary. **Support.** The effective number of places a promise's mass is spread over. **Tape.** The append-only, hash-chained record; the only truth. **Tolerance (δ).** The relative excess harm the writ agrees to call parity; dial one. **Tower of shadows.** Frontier, resident, compiled code: graduation down, grading up. **Warrant seat.** A signature, liability, someone who can be sued. **Writ.** What the organization is for, authored once, never learned.

The synthetic firm ACME-500 at E0, FUSOR as built, the CUDA solver family and the flight computer, and the receipts of 13 to 15 September 2026 stand behind every number marked measured. The four dials are named. The arithmetic is printed. The interior is deleted. Ninety percent.