FRACTIONAL CAIO FIELD MANUAL

The Mid-Market AI *Operating System*

Four working tools for the first ninety days of any serious AI program. Governance, decision routing, board reporting, and the deployment gates that get models past legal.

FOREWORD

Most AI programs don't fail at the model. They fail at the *handoff*.

I've shipped six production AI systems this year and spent the four years before that inside SR 11-7 model risk at a top-five U.S. bank. The same failure pattern shows up at every scale: the model works, the pilot demos well, and then the program stalls because nobody owns governance, decision rights, board reporting, or deployment.

> *Strategy decks don't ship AI. Operating systems do.*

This field manual is the operating system. Four working tools, designed to be used together, that compress the first ninety days of a serious AI program into something a CEO, CTO, and board can all see at once.

WHAT'S INSIDE

TOOL ONE

The Model Governance *Checklist*

Twelve items every mid-market AI program needs before the next pilot. This is SR 11-7 calibrated for non-banks — the regulatory framework that governs model risk at every U.S. bank, scaled down to what a 50–500-person company can actually maintain. If your team can't check ten of these off in the next thirty days, you're not ready to deploy. You're ready to govern.

INVENTORY & OWNERSHIP

☐ *01* **A model inventory exists and is current within 30 days**

Every model in production, every model in pilot, every shadow model someone built in a spreadsheet. You cannot govern what you cannot count. One named owner per model.

☐ *02* **Each model has a single accountable executive**

Not the engineering team. A VP-or-above whose performance review reflects the model's behavior. If no executive will sign, the model isn't ready to ship.

☐ *03* **A documented business purpose, in one paragraph**

What decision does this model inform? What's the cost of a wrong decision? If it takes more than a paragraph, the use case isn't clear enough to govern.

RISK & FAILURE MODES

☐ *04* **Defined failure mode with a named "who pays" owner**

When the model is wrong, what does wrong look like — and which P&L absorbs the loss? Customer-facing? Regulatory? Reputational? Make it concrete and dollar-denominated.

☐ *05* **A documented kill switch and reversal plan**

If the model goes wrong tomorrow, who turns it off, how fast, and what's the manual fallback? Written down. Tested at least once.

☐ *06* **Drift monitoring with thresholds, not vibes**

A specific metric (accuracy, calibration, distribution shift) and a specific number that triggers review. "We'll watch it" is not monitoring.

DATA & INPUTS

☐ *07* **Training data lineage is documented end-to-end**

Source system, transformation steps, refresh frequency, PII handling. The first question every regulator and every acquirer's diligence team asks.

☐ *08* **A bias and fairness review, sized to the use case**

For lending, hiring, insurance: full disparate-impact testing. For internal tooling: a one-page protected-class review. Either way: written, dated, signed.

☐ *09* **Vendor model assumptions are surfaced, not buried**

If you're using OpenAI, Anthropic, or any third-party model, what's the SLA, the data retention policy, and the fallback if the API changes overnight? Recorded in the model inventory.

OPERATING DISCIPLINE

☐ *10* **A human-in-the-loop checkpoint exists where stakes are highest**

See Tool 02. Not every decision needs human review — but every high-reversal-cost decision does, and someone's job description says so.

☐ *11* **Quarterly model review on the executive calendar**

Standing meeting. The accountable executive walks through inventory, drift, incidents, and proposed changes. If it's not on the calendar, it's not happening.

☐ *12* **An incident log with at least one entry**

Models fail. Programs that pretend otherwise are programs without telemetry. The first incident log entry is the first sign of an honest practice.

FIELD NOTE

Where mid-market companies most often fall short

Items 02, 04, and 11. Engineering teams are usually fine. Executive accountability, dollarized failure modes, and a standing review cadence are where governance becomes real — and where most AI programs quietly stall for a year before someone notices.

TOOL TWO

The Confidence-Threshold *Matrix*

Most production AI failures aren't model failures. They're routing failures — the wrong decision-maker at the wrong threshold. This matrix forces a team to draw two axes and place every model decision in a quadrant. If you can't do this on a whiteboard in five minutes, that's the work.

> *Auto-execute the cheap-to-reverse. Human-approve the expensive-to-reverse. Refuse to deploy the catastrophic.*

THE MATRIX

	REVERSAL COST: LOW	REVERSAL COST: HIGH	REVERSAL COST: CATASTROPHIC
Confidence ≥ 0.90	**Auto-execute** Model decides. Logged for audit. No human in the loop.	**Human-approve** Model recommends. Named human approves before action.	**Do not deploy** Confidence is irrelevant when the downside is unrecoverable.
Confidence 0.70–0.90	**Auto-execute, sample QA** Model decides. 5–10% sampled for human review weekly.	**Human-decide, model-informed** Model surfaces options and confidence. Human chooses.	**Do not deploy** Catastrophic risk requires near-certainty, not heuristics.
Confidence < 0.70	**Refer or default** Route to human or use a deterministic fallback rule.	**Refer to human** Model is informational only. Human owns the decision.	**Do not deploy** Model has no business in a catastrophic-cost decision.

CALIBRATING THE TWO AXES

Confidence

Most teams use raw model probability. Don't. Calibrate first — a model that says "0.85" should be right 85% of the time, and most aren't out of the box. Hold out a calibration set, fit a Platt scaling or isotonic regression, then route on the calibrated score.

Reversal Cost

Dollarize it. "Low" means under one hour of human time or under $100. "High" means a customer apology or a four-figure refund. "Catastrophic" means regulatory exposure, a lawsuit, or a public trust failure. If you can't put a number on it, you don't yet understand the decision.

WORKSHEET — MAP YOUR TOP DECISIONS

For each model decision in production or pilot, fill in one row. Most teams find that 60–80% of their decisions are misrouted on first audit.

DECISION	CALIBRATED CONFIDENCE	REVERSAL COST ($)	QUADRANT	CURRENT ROUTING — AND IS IT RIGHT?

THREE DIAGNOSTIC QUESTIONS

01 **For every "auto-execute" cell — when did you last sample-audit the outputs?** If the answer is "we haven't," the cell is wrong. Auto-execute requires sampled QA or it's just deployment-by-hope.

02 **For every "human-approve" cell — what's the human's actual response time?** If approvals consume more than 10% of the human's working day, raise the auto-execute floor or accept that the model isn't ready.

03 **For every "do not deploy" cell — is the model deployed anyway?** Don't laugh. Most common audit finding: a catastrophic-cost decision routed to an uncalibrated model.

TOOL THREE

The Board-Ready AI *Scorecard*

One page. Four sections. The numbers a CEO should be able to read to a board in under five minutes — not a deck, not a narrative, a scorecard. The template below is the format I use in CAIO engagements.

AI Program Scorecard

Q_ · 20__

METRIC	TARGET	ACTUAL	Δ
01 · VALUE			
Models in production generating measurable P&L impact	—	—	—
Annualized run-rate value (gross margin or cost-out, $)	—	—	—
Cost per model decision (compute + ops, $)	—	—	—
02 · RISK			
Models with current governance documentation (% of inventory)	*100%*	—	—
Open incidents over 30 days unresolved (count)	*0*	—	—
Models breaching drift threshold (count)	*0*	—	—
03 · VELOCITY			
Time from pilot to production (median, days)	—	—	—
Pilots killed this quarter (count — yes, this is a positive number)	*≥1*	—	—
Production models retired or replaced (count)	—	—	—
04 · CAPABILITY			
Headcount with AI delivery experience (FTE)	—	—	—
External vendor concentration (% of total AI spend on top vendor)	*< 60%*	—	—
Knowledge transfer events completed (sessions)	—	—	—

HOW TO READ IT — AND HOW A BOARD READS IT

Value

The board's first question is always "what's it worth." If you can't put a dollar number next to a deployed model, that model is a science project — and the right next conversation is whether it should be killed or commercialized.

Risk

The board's hidden question, the one they ask each other after the meeting. If governance coverage isn't 100% on a quarterly basis, you're one incident away from a much harder board meeting. Treat this section as the canary, not the trailing indicator.

Velocity

The single most diagnostic line on the scorecard is "pilots killed this quarter." Zero is a red flag — it means the team can't tell the difference between a working model and a working program. Healthy programs kill at least one pilot every ninety days.

Capability

Vendor concentration is the line every PE board cares about and most CEOs underweight. If 80% of your AI spend is on one vendor, the vendor is the program. Bring it under 60% within four quarters or accept that you're outsourcing your competitive moat.

CADENCE

- **Monthly:** CTO + CAIO update Risk and Velocity numbers internally.
- **Quarterly:** Full scorecard delivered to the board pack one week before the meeting.
- **Annually:** Targets reset based on the prior year's actuals plus the strategic roadmap. The target column is not a wish list — it's a commitment.

FIELD NOTE

The "ratio test"

A healthy mid-market AI program produces a scorecard where Value > Risk-adjusted-cost, Velocity is improving quarter over quarter, and Capability is shifting from external vendors to internal headcount. If two of those three are flat or worsening, the program needs a Fractional CAIO before it needs another pilot.

TOOL FOUR

The Five-Gate Deployment *Checklist*

A model is "deployed" when it has passed five gates, in order, with named owners signing each one. Most AI programs that look stuck are actually mid-gate — treating engineering completion as the finish line and discovering, six weeks too late, that legal hasn't seen the documentation.

01

Documented

Model purpose, data lineage, training methodology, calibration results, known failure modes, and the kill switch — all in one document, in plain language, signed and dated.

Common stall: documentation written after the fact, in a tone the model risk committee won't accept.

OWNER
MODEL OWNER

02

Reviewed

Independent review by someone not on the build team. For high-risk models, a formal model risk committee. For lower-risk models, a peer review with written sign-off.

OWNER
RISK / VALIDATION LEAD

03

Deployed

Code is in production. CI/CD passes. Monitoring is live with the drift thresholds defined in the governance checklist. A rollback path is tested.

Common stall: deployed without monitoring, which means you'll find out it broke from the customer who complained loudest.

OWNER
ENGINEERING LEAD

04

Integrated

Downstream consumers — sales ops, customer service, the BI team — are using it in real workflows. Training delivered. Runbooks written. Tested with real users on real data, not synthetic demos.

OWNER
BUSINESS SPONSOR

05

Measured

Success metrics flowing into the dashboard the executive sponsor reads weekly. Value is being captured. The line on the scorecard is moving.

Common stall: nobody owns the dashboard, so nobody notices when value flat-lines.

OWNER
EXECUTIVE SPONSOR

GATE SIGN-OFF — ONE PER MODEL

Print one for every model heading toward production. Hang it on the wall if you have to. The five gates are sequential, not parallel; a model doesn't progress until the prior gate is signed.

GATE	STAGE	ACCEPTANCE CRITERIA	OWNER SIGNATURE	DATE
01	**Documented**	Governance doc complete, signed by accountable executive, kill switch defined.		
02	**Reviewed**	Independent review complete, written sign-off filed, material findings addressed.		
03	**Deployed**	Code in production, monitoring active, rollback tested in last 14 days.		
04	**Integrated**	Downstream consumers trained and using model in real workflows.		
05	**Measured**	Success metric flowing into executive dashboard; first reading recorded.		