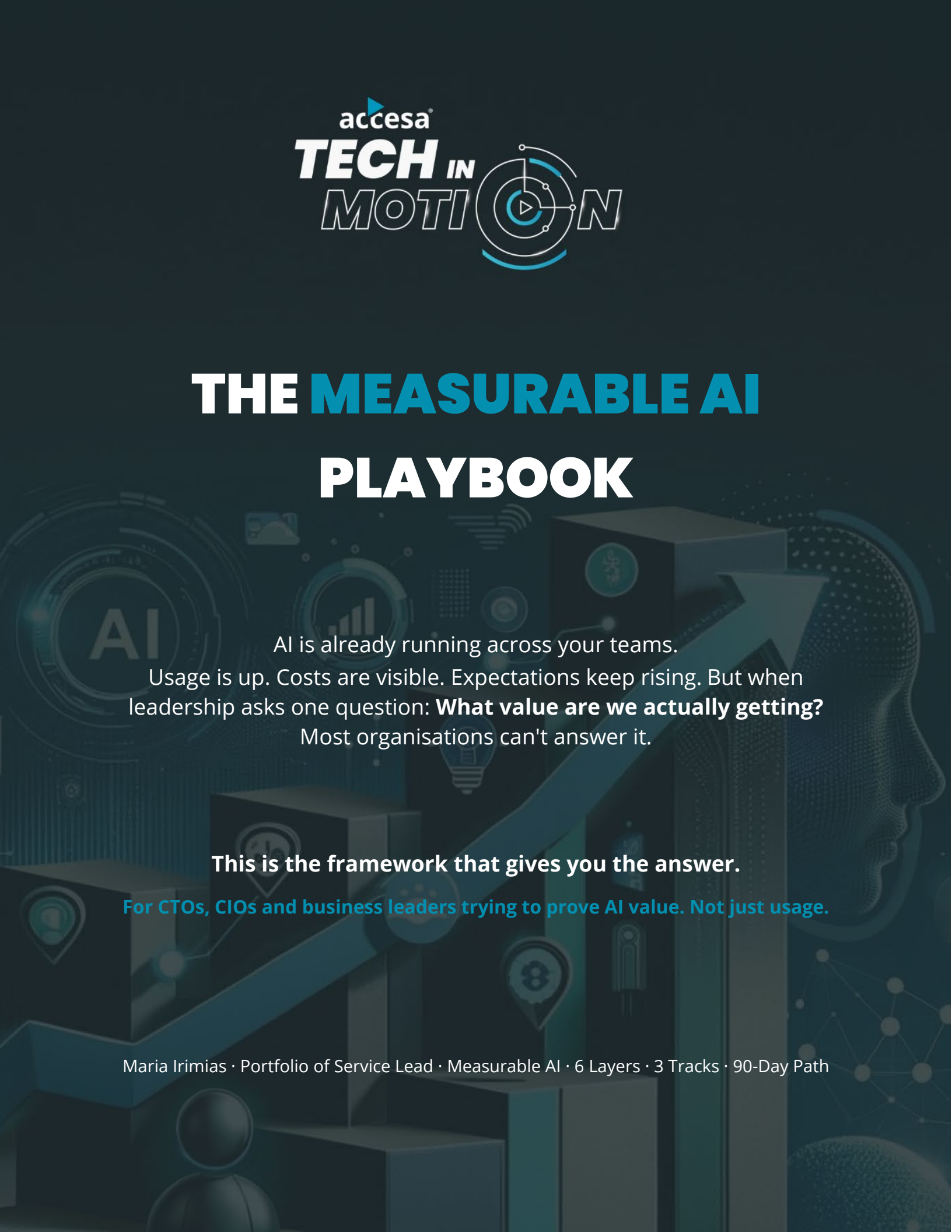
accesa

TECH IN MOTION

THE MEASURABLE AI PLAYBOOK



AI is already running across your teams.
Usage is up. Costs are visible. Expectations keep rising. But when leadership asks one question: **What value are we actually getting?**
Most organisations can't answer it.

This is the framework that gives you the answer.

For CTOs, CIOs and business leaders trying to prove AI value. Not just usage.

Maria Irimias · Portfolio of Service Lead · Measurable AI · 6 Layers · 3 Tracks · 90-Day Path

01

The Problem

Why Most AI Investments Fail to Prove Their Value

The pattern: Teams ship AI features but nobody ships the measurement alongside them. So when the CFO asks for ROI, there's no number to give. High usage tells you people are clicking; it doesn't tell you whether any value came out the other end. That distinction is the whole point of measuring properly.

Teams measure a lot. They just measure the wrong things — and they're confident they've got it right. AI can generate activity, but it can't generate proof. Report adoption before you've agreed what meaningful use means and you get vanity metrics: usage charts that climb while nobody can name a single outcome they bought. The gap that matters is between tracking how much AI activity happens and tracking what AI impact it actually produces. Activity is easy to show. Impact is the only thing the board pays for.

95%	61%	53%
of enterprise AI projects fail to show financial returns within 6 months — MIT 2025	of senior leaders feel more pressure to prove AI ROI than a year ago — Kyndryl 2025	of investors expect positive ROI in 6 months or less — Teneo 2026

02

Architecture

The Six-Layer Framework

Most teams try to measure AI impact. What they miss is sequence. You don't start with ROI — you earn the right to measure it. Each layer is the prerequisite for the next: Layers 1–3 apply to every team, and Layer 4 only matters once your volume justifies it.

Layer	Name	Question / Formula	Status
L−1	AI Readiness	Are we ready to deploy AI at all? DORA 2025's seven AI Capabilities: AI stance, healthy data, data access, version control, small batches, user focus, internal platform.	**HARD GATE**
L0	Baseline Capture	Do we have our anchor numbers? 5 metrics, 2 hours, before first deployment.	**HARD GATE**
L1	Efficiency	Is AI earning its compute cost? ES = Quality ÷ (cost + latency + Δ)	**REQUIRED**
L2	Stability	Is AI making us better or fragile? Δ = (ΔCFR + ΔRework) ÷ 2	**REQUIRED**
L3	Impact	Can we prove value to the board? IS = (A×0.30) + (Q×0.40) + (P×0.30)	**REQUIRED**
L4	Unit Economics	Optimising at scale — €/outcome · €/inference · TPSO. Activates Month 7+	**OPTIONAL**

Don't treat Layer 4 as a maturity badge. It's a financial optimisation tool, and it only pays off once you're running enough inference volume to justify it. A team at Layer 3 with a stable Impact Score is already doing this right.

03

Layer by Layer

What to Measure and When

Layer −1 + 0 · Before You Deploy — Baseline Capture: 5 Numbers, 2 Hours

Without a baseline, a metric is just a count — and a count proves nothing. Get these five down before any AI tooling goes live; once it's live, the "before" picture is gone for good.

- **01 Deployment Frequency —** How often do you ship?
- **02 Change Failure Rate —** % of deploys causing incidents
- **03 Lead Time —** Commit to production — hours or days?
- **04 Rework Rate —** Unplanned work from production issues
- **05 Developer Trust Index —** Anonymous 1–5 weekly survey. Gate: >60%

Layer 2 · The Hidden Risk — Instability Delta (DORA 2025)

DORA 2025 (5,000 professionals): AI lifts throughput, but it also pushes up instability. Most teams watch one and ignore the other — they bank the velocity and miss the stability cost until rework rate has already crept up. Throughput alone is a trap: it stays green right up until the rework starts.

> *"The more AI was used, the worse the software delivery stability and throughput became — the very attributes software development professionals have been working for the last decade to improve." — Gene Kim, Foreword, DORA State of AI-Assisted Software Development 2025*

Δ = (ΔCFR + ΔRework Rate) ÷ 2

where Δ = current value − pre-AI baseline (Layer 0). Both ΔCFR and ΔRework are the two factors DORA 2025 defines as software delivery instability. Note: the numerical thresholds below are operational guidance derived from practice — DORA 2025 does not publish specific Δ cut-offs.

Δ Range	Meaning	Action
Δ < 0	AI improves speed AND stability	SCALE INVESTMENT
0 – 0.05	Acceptable in first 90 days	MONITOR WEEKLY
0.05 – 0.15	Accelerates but destabilises	FIX COVERAGE FIRST
Δ > 0.15	Amplifies dysfunction	**STOP EXPANSION**

Layer 3 · The Board Number — Impact Score

IS = (Adoption × 0.30) + (Quality × 0.40) + (Productivity × 0.30)

Quality carries 40% on purpose. Going faster while quality drops is how most AI projects come undone — speed shows on day one, the rework shows up weeks later in production. Reward velocity without penalising the rework it creates and you'll greenlight the wrong investments.

Impact Score	Meaning	Action
IS < 0.25	Positive ROI signal not present	FIND THE BLOCKER
IS ≥ 0.25	Positive ROI signal — continue direction	REPORT MONTHLY
IS ≥ 0.50	Strong ROI signal	**SCALE AGGRESSIVELY**

Layer 4 · True AI ROI — The CFO Number

True AI ROI = Business Value Delivered ÷ TCAIO

TCAIO = Licences + Consumption + Fix Costs + Monitoring + Drift Fixing + Hidden

Fix costs are typically 4–6× the budgeted amount. A team budgeting £600 for licences + tokens often has real TCAIO of £2,800+. Reconciliation contract: $|\Sigma \text{ allocated} - \Sigma \text{ invoiced}| \leq 5\%$ per period. Use the full TCAIO, not just token cost. ROI built on licences and tokens alone falls apart the moment finance starts asking questions.

04

Three Tracks

One Discipline — Different Signals

The six layers hold across every role — only the signal sources change. The board number, the Layer 3 Impact Score, stays the same formula on every track. A 200-step agent that fails the goal is worse than a 5-step agent that succeeds. The metric is Goal Completion Rate — not steps, not invocations, not token consumption.

01 Engineering Teams	02 Business Roles	03 Agentic Solutions
ENGINEERS · TECH LEADS · QA	**PM · LEGAL · FINANCE · HR · SALES**	**PRODUCT · PLATFORM · AI TEAMS**
• Efficiency Score (quality ÷ cost) • Instability Delta (Δ) • Code Survival Rate at 30 days • DORA 2025 — 5 metrics	• Task Automation Rate • Rework Avoidance Rate • Decision Cycle Time • Adoption & Trust Index	• Goal Completion Rate >85% • Human Override Rate 20–50% • Hallucination Rate ≤ 2% • Unit Cost per Outcome
Does AI make us better — or just faster?	**Is AI removing friction — or adding steps?**	**Does the agent actually do the job?**

05

For the Board

The Break-Even Conversation

Present this curve before deployment — not after the CFO asks. 90% of AI programmes die at Month 1–2 because nobody showed the board that negative returns are expected, not failure.

Phase	What's happening	Return
Month 1–2	DANGER ZONE — 90% of programmes quit here. Near-zero or negative returns. Learning curve. Expected — not failure.	0.1×–0.4×
Month 3–6	Inflection point — your proof window. Team proficient. Break-even achieved. Reinvest early savings.	**4×–9×**
Beyond 6m	Model direction, not destination. Landscape shifts too fast for multi-year projections. Measure: are we on the right side of break-even?	→ DIRECTION

Teams that quit at month two never reach the inflection point. The model is the argument — show it before the investment starts, not in the post-mortem.

06

Implementation

90-Day Roadmap

This is how we've done it at Accesa: start with one delivery team, instrument before you change anything, and only widen once the numbers hold. Ninety days is enough to get from a blank baseline to a board-ready Impact Score — provided you resist the urge to measure everything at once.

Phase 1 · Instrument + Baseline (Days 1–30)

- Activate logging — Helicone (free ≤100k req/month) or Langfuse open-source. 30-minute setup.
- Capture DORA baseline — 5 numbers written down before any AI tooling goes live. This is your anchor.
- Developer Trust Index — anonymous 1–5 weekly survey. Set up now, run weekly thereafter.
- Define meaningful use — agree what counts as value-adding use before you report adoption numbers.

Deliverable: 8–10 reference numbers written down. Skip these and you'll have no way to prove anything in Phases 2–3 to finance.

Phase 2 · Measure + Validate (Days 31–60)

- Calculate Efficiency Score — per task type. Identify where AI earns its compute cost.
- Compute Instability Delta — compare current DORA vs baseline. Apply threshold table.
- Track adoption funnel — exposure → meaningful use → reliance. Find where users drop off.
- First dashboard — 5 metrics only — IS components + Instability Delta + DTI. Keep it to five, not fifty.

Deliverable: Efficiency Score per task type + Instability Delta. Can you answer: is AI accelerating us or destabilising us?

Phase 3 · Automate + Communicate (Days 61–90)

- CI gate on Instability Delta — block the PR automatically when Rework Rate crosses the threshold. This will shift behaviour faster than any process document ever could.
- Calculate first Impact Score — IS = (A×0.30) + (Q×0.40) + (P×0.30). One number per team.
- First board report — one page, three numbers: IS, True AI ROI, Instability Delta. One decision recommended.

Deliverable: Live dashboard + board report + one decision. Scale, redesign, or fix foundations first.

07

Measurement Discipline

Stop Measuring These

Run every metric you currently report through three questions. Can it climb while actual outcomes get worse? Does anyone make a decision based on it? Would a 10% swing change how you allocate resources? If you answer no to any of them, drop the metric.

STOP measuring	Measure instead
Lines of code generated	Code Survival Rate at 30 days
AI queries per month	AI-assisted task completion rate
Total chatbot interactions	Sessions resolved + Containment Rate
Features shipped count	Features with ≥5% productivity gain vs baseline
Agent steps / invocations	Goal Completion Rate + Unit Cost per Outcome
Model benchmark accuracy	Human Correction Rate in production

GitClear 2025: AI-assisted code churn has climbed from 3.3% before AI to 7.1% after. Teams writing more code are throwing more of it away too. When the volume doesn't survive, what you've built up is technical debt rather than productivity.

08

Quick Reference

All Formulas

Metric	Formula	Layer
Data Readiness	AI-accessible records ÷ total required × 100 · Gate: >80%	**L−1**
DTI	Σ trust scores ÷ (respondents × 5) × 100 · Gate: >60%	**L0**
Activation Rate	Users with ≥ threshold meaningful uses ÷ invited × 100	**L0**
ES_simple	Quality ÷ (token_cost_norm + latency_s + instability_Δ)	**L1**
ES_full	Quality ÷ (token_cost_norm + latency_s + instability_Δ + fix_cost_norm)	**L1**
TCAIO	Licences + Consumption + Fix Costs + Monitoring + Drift Fixing + Hidden	**L1**
Instability Δ	(ΔCFR + ΔRework Rate) ÷ 2 · Δ > 0.15 → stop expansion	**L2**
HOR	Interventions ÷ total AI actions × 100 · Target 20–50%	**L2**
GCR	Goals achieved ÷ goals initiated × 100 · Target >85%	**L2**
Impact Score	(Adoption×0.30) + (Quality×0.40) + (Productivity×0.30) · IS ≥ 0.25 = positive ROI	**L3**
True AI ROI	Business Value Delivered ÷ TCAIO · Always use full TCAIO	**L3**
Cost/1k tokens	(GPU + API + storage + egress) ÷ (tokens ÷ 1,000)	**L4**
TPSO	Total tokens consumed ÷ successful outcomes	**L4**

The Real Question

How do we prove AI value — not just activity?

If that's the question your organisation is trying to answer, this is exactly the work I do. Let's talk.

About

Maria Irimias

Portfolio of Services Lead · Accesa

maria.irimias@accesa.eu

/in/mariairimias

Presented at Tech in Motion 2026. - This playbook is a companion to the talk Measurable AI.

Sources

- **DORA 2025 Report —** Instability Delta · 5,000 professionals
- **McKinsey State of AI 2025 —** 94% no significant value
- **BCG AI Impact 2025 —** Compounding ROI model
- **MIT 2025 —** 95% failure rate
- **Kyndryl 2025 —** 61% ROI pressure
- **Teneo 2026 —** 53% investor expectations
- **GitClear 2024–2025 —** Code churn 3.3% → 7.1%
- **GitHub + Accenture 2024 —** 55% faster, 4,800 devs
- **NAV IT / arXiv 2025–2026 —** Longitudinal Copilot study
- **Perumal 2025 —** TCAIO · FOCUS-aligned