ML LABS INTELLIGENCE

How to Measure AI Impact That Matters

Teams struggle to defend AI spend when metrics stop at usage or model quality. This article shows how to measure impact in terms the business can act on.

STRATEGIC

Author Omar Trejo

Date 2026-03-17

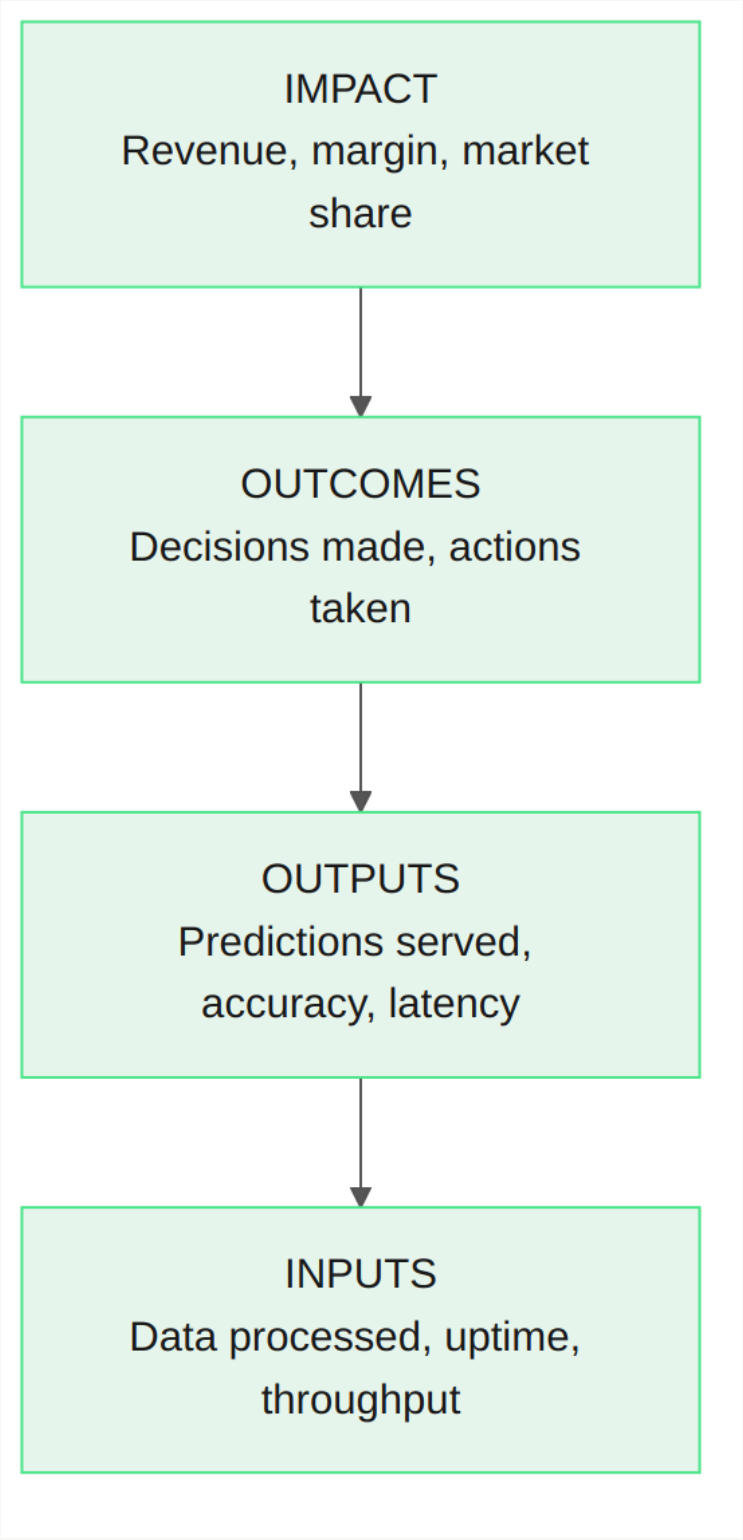
"Our model achieves 94% accuracy." This appears in virtually every AI project update. It tells budget owners almost nothing. Model metrics measure whether predictions are correct, not whether anyone acts on them or whether those actions produce business results. The gap between technical performance and measurable impact is where most AI programs lose executive trust.

A survey of 1,000 C-suite executives (BCG, 2024) found that 74% have yet to show tangible value from AI, even as budgets increase. The model might be excellent while the business outcome remains unchanged — predictions aren't acted upon, arrive too late, or address the wrong problem. **Research on scaling AI** (MIT Sloan, 2024) found that organizations scaling AI define business outcomes before model metrics.

A separate problem compounds the gap: **absence of baselines**. You cannot prove improvement without documenting the pre-AI state, yet most projects begin without them. **Research on AI and the productivity paradox** (NBER, 2018) showed that even national-level statistics fail to capture AI's benefits; organizational measurement gaps are far worse.

The Measurement Hierarchy

Effective AI measurement follows a hierarchy from activities (what the system does) to impact (what changes in the business). Each level answers a different question and matters to a different audience.


IMPACT
Revenue, margin, market share
OUTCOMES
Decisions made, actions taken
OUTPUTS
Predictions served, accuracy, latency
INPUTS
Data processed, uptime, throughput

Most organizations measure only the bottom two levels. They can tell you the model is running and performing well. They cannot tell you whether anyone acts on its predictions or whether those actions produce results.

Level 1: Input Metrics

- **Data freshness**: stale data produces stale predictions
- **System uptime**: prediction service availability
- **Throughput**: predictions generated per unit time

Level 2: Output Metrics

- **Coverage**: percentage of inputs handled vs. falling back to defaults
- **Prediction accuracy** (precision, recall, F1)
- **Inference latency** (p50, p95, p99)

Level 3: Outcome Metrics

Research on AI and human decision-making (HBR, 2021) found the primary determinant of AI value is whether the system changes actual decision-making behavior.

- **Override rate**: how often humans override the model (higher is worse)
- **Adoption rate**: percentage of intended users consuming predictions
- **Action rate**: how often a prediction triggers a decision

Level 4: Impact Metrics

The only metrics that justify continued investment.

- **Revenue impact**: incremental revenue from AI-influenced decisions
- **Cost impact**: operational savings minus AI system cost
- **Risk impact**: reduction in adverse events from AI detection

Establishing Baselines

Before deploying, document current performance on the target metric, the measurement methodology, and the variance range so you can distinguish improvement from noise. **Research on AI value metrics** (Gartner, 2024) recommends substantial baseline data and a meaningful post-deployment window before claiming impact.

Leading vs. Lagging Indicators

Leading indicators predict future impact; lagging indicators confirm it. Track leading indicators frequently to catch problems early; track lagging at a longer cadence to confirm value. **Research on enterprise AI maturity** (MIT CISR, 2025) found that progression from piloting to scaled AI depended on measuring the right leading indicators consistently.

- **Leading**: adoption, decision speed, confidence, data quality
- **Lagging**: revenue, cost reduction, error rates, satisfaction

Expected Results

Organizations implementing the full hierarchy typically discover that many systems considered successful aren't delivering business impact — they perform well technically but don't change outcomes. Measurement reveals the specific layer where value breaks down, allowing targeted fixes or elimination.

The Attribution Challenge

> Conservative attribution beats optimistic attribution every time. Naive before/after comparisons commonly overestimate AI impact due to confounding factors.

Three approaches, in order of rigor:

1. **Randomized A/B testing** (gold standard): treatment vs. control groups, comparing outcomes directly
2. **Difference-in-differences**: compare before/after change against a comparable metric unaffected by AI
3. **Interrupted time-series analysis**: project what would have happened without AI from pre-deployment trends

When This Approach Does Not Apply

The hierarchy assumes outcomes can be observed and attributed. This breaks with incomplete instrumentation (no capture of post-prediction actions), missing baselines, and high causal complexity. For missing baselines, establish them now — the payoff is forward-looking. For high causal complexity, focus on leading indicators and let lagging data accumulate over longer horizons.

First Steps

1. **Pick one production system.** Map it to the hierarchy. The gaps between levels tell you where to invest.
2. **Establish baselines now.** For deployed systems, reconstruct from historical data or run controlled rollback experiments.
3. **Build a single-page dashboard.** One metric per level, shared with both AI team and business stakeholders.

Practical Solution Pattern

Implement a layered measurement stack linking technical performance to behavioral change to business outcomes, with baselines and attribution rules agreed before deployment. Build a single-page dashboard with one metric per level — input, output, outcome, impact — updated on a consistent cadence.

Value leakage in AI programs almost always occurs at the transition between levels. A technically accurate model can fail to change decisions, and changed decisions can fail to move business results. Tracking all four levels simultaneously makes the failure point visible. Conservative attribution methods established before deployment prevent the overestimation bias that erodes executive trust. Organizations that need to establish whether existing AI systems are delivering business value can get a structured assessment through an **AI Technical Assessment** that maps systems to the measurement hierarchy and identifies where value is leaking.

References

1. Boston Consulting Group. **Where's the Value in AI?**. *BCG*, 2024.
2. MIT Sloan Management Review. **Winning With AI**. *MIT Sloan Management Review*, 2024.
3. Brynjolfsson, E., Rock, D., and Syverson, C. **Artificial Intelligence and the Modern Productivity Paradox**. *NBER Working Paper*, 2018.
4. De Cremer, D., and Kasparov, G. **AI Should Augment Human Intelligence, Not Replace It**. *Harvard Business Review*, 2021.
5. Gartner. **5 AI Metrics That Actually Prove ROI**. *Gartner Research*, 2024.
6. MIT Center for Information Systems Research. **Enterprise AI Maturity Update**. *MIT CISR*, 2025.


BUSINESS OUTCOMES