Escaping Pilot Purgatory

Why Your AI Pilots Stall — and How to Break Through

Martin Rojas · PlayOn Sports

Two Studies. Same Projects. Different Answers.

MIT (2025)

~90% of AI projects fail

Measured immediate P&L impact. If it didn't move the bottom line within the quarter, it didn't count.

Wharton Study (2025)

~75% succeed

Applied broader evaluation criteria — learning, capability building, strategic optionality. Same projects, different lens.

The difference wasn't the work. It was how they measured success. We don't have a consistent measurement framework — and that gap is why pilots stall.

Your AI pilot worked. Leadership loved the demo. And then... nothing.

Not because the technology failed. Because nobody defined what **"ready to scale"** looked like *before* the pilot started. The demo was a success. The transition plan was missing.

The Coca-Cola Freestyle Problem

Two valid technical approaches for the consumer-to-technician interface switch on Freestyle machines. From a purely technical standpoint, who cares about 30 seconds?

Then someone did the math.

95,000

Freestyle machines in field

4×/yr

Technician visits per machine

$30/hr

Average technician labor rate

30 sec

Added delay per visit

~$95K

Per year. In wasted wages.

From a 30-second screen switch that "didn't matter."

The KPI lived outside the technology. That's exactly what's happening to AI today.

Geoffrey Moore's Three Horizons

Credit where it's due — Moore created Three Horizons to classify *types of businesses*, not initiatives. H1 is the legacy core that pays the bills. H2 is the emerging growth engine. H3 is the speculative bet. The tension Moore named: **H1 lives for process and compliance. H2 and H3 live for learning. They clash by design.**

H1 — Core Business

Process & compliance. Revenue-generating. Operationally mature.

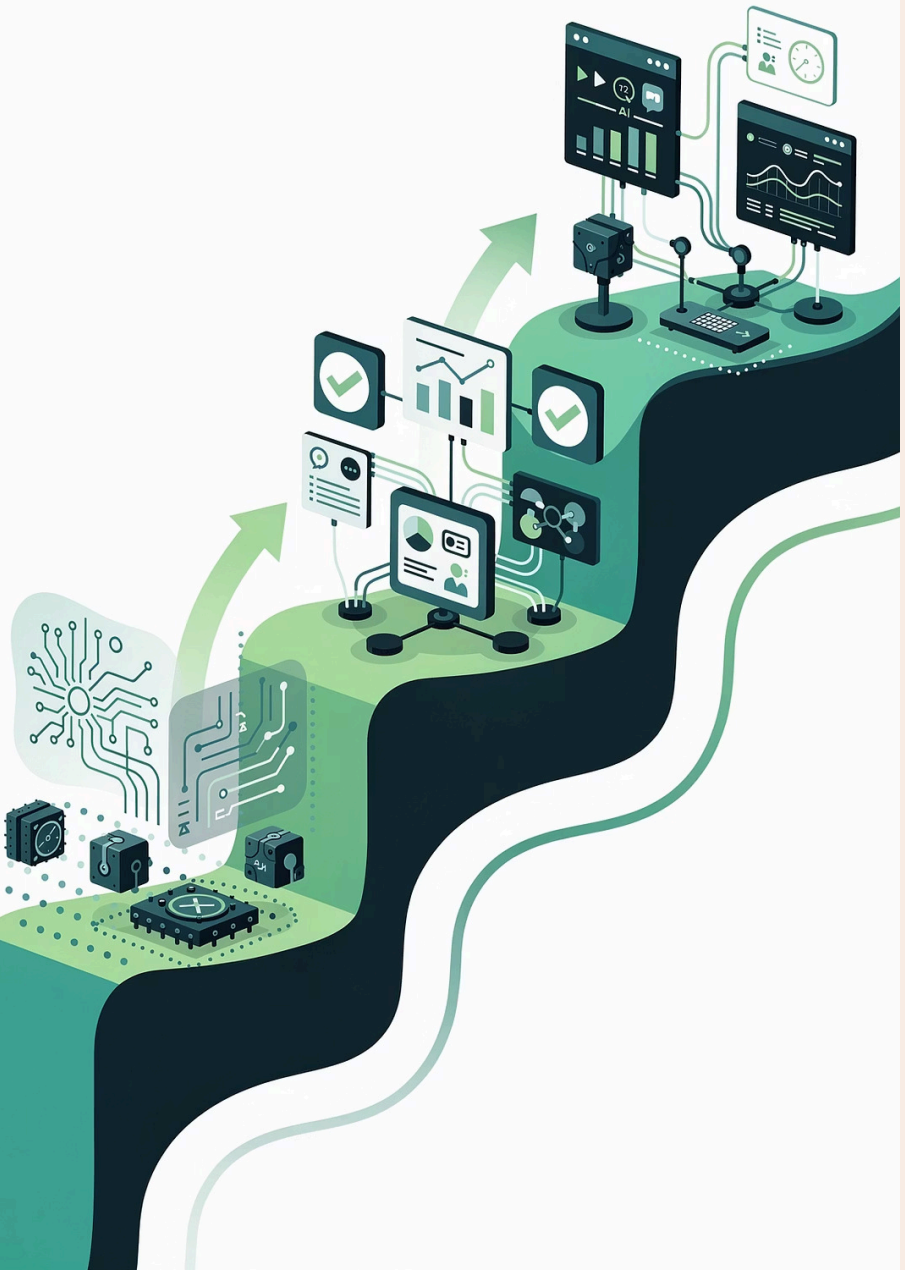
H2 — Emerging Business

Growth bets. Proving value. Bridging exploration and scale.

H3 — Speculative Bets

Exploration. Experimentation. No proven model yet.

Moore used this to classify **businesses**. We repurposed it — and flipped it — for AI **initiatives**.



For AI Initiatives, We Invert It.

Moore described businesses. We use Three Horizons for *initiatives within* a business — and we flip the direction. Every AI idea starts at H3 (exploration). H1 is where it becomes operational. **Numbers represent maturity, not time. You don't start at H1. You earn your way there.**

H3 — Exploration

Where every initiative begins. Prove it's possible.

H2 — Validation

Prove it's valuable. This is where most pilots live — and die.

H1 — Operations

Integrated. Adopted. Delivering durable, measurable value.

Three Horizons isn't a maturity model. It's a metrics migration framework.

Each horizon demands a fundamentally different *type* of measurement. The goal isn't to move through phases — it's to deliberately change what you're measuring as you go.

Pilots stall because organizations never formally transition their metrics. They keep measuring H3 feasibility — *"it works!"* — long after they should be measuring H2 value — *"is it worth it?"*

⚠ The failure mode isn't technical. It's organizational. The measurement conversation never happens.

The Migration Table

This is the table to memorize. Feasibility is binary. Value is conditional. Sustainability is durable. Each gate is a deliberate leadership decision — not a natural progression.

Horizon	Question	Metric Type	Time Box
H3 Exploration	"Can we build it?"	**Feasibility** — binary pass/fail	1-2 sprints
H2 Validation	"Is it solving the problem?"	**Value** — KPIs + KRIs, time saved, $ per outcome	1 quarter
H1 Operations	"Is this how we work now?"	**Sustainability** — costs at scale, adoption, drift	Ongoing

HORIZON 3

Exploration — "Can We Build It?"

Quick Stats

2

Max Sprints

1

Gate Review

Required before launch

Gate Criteria to Advance

- **Technical feasibility:** Can we actually build this with available data, tools, and talent?
- **Strategic alignment:** Does it connect to a real business problem, not just a cool capability?
- **Quantifiable KPI hypothesis:** What number are we trying to move — and where does it live?

⚠ **The Trap:** Staying too long. Endless exploration without a forcing function. Set the gate review *before* you start — not after you're attached to the output.

HORIZON 2

Validation — "Is It Solving the Problem?"

Metrics: Value — two halves of the same coin

- **KPIs** – leading indicators of value: unit economics, time saved, $ per call
- **KRIs** – leading indicators of risk: token cost trending up, accuracy trending down, adoption flatlining

Time box: 1 quarter

Gate criteria to advance:

- Defined target users – not "everyone." A specific cohort with measurable behavior.
- **Both** KPIs *and* KRIs defined *before* you start
- Results that can be isolated

The Critical Unlock

Entry criteria for H2 must include **exit criteria to H1.**

Define "ready to scale" *before* you pilot – or you're already in purgatory.

⊗ **The Trap:** Measuring only KPIs. If you're not tracking KRIs, you're flying blind on risk.

HORIZON 1

Operations — "Is This How We Work Now?"

Integrated into workflows. Adopted by users. Delivering measurable value at scale. New metric class: **sustainability.** Are H2 assumptions holding? Are costs tracking? Are users adopting — or just tolerating?

H2 Hit Its Numbers

Not just "it worked" — the value metrics landed. The case was made with evidence, not enthusiasm.

Scale Plan Exists

Deployment path, infrastructure, and rollout sequencing are defined and funded.

Production Owner Named

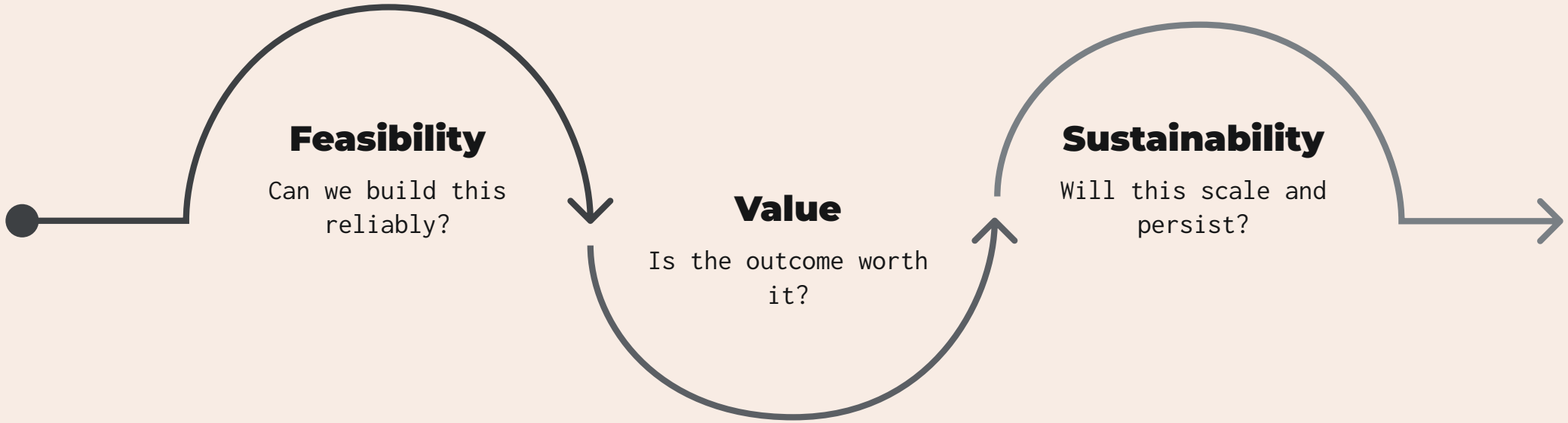
A single person is accountable for this system in production — not the pilot champion.

Reusable by Other Teams

The capability can be leveraged beyond the original use case. That's how you know it's scaled.

⚠ **The Trap:** Governance gaps. Data privacy, bias monitoring, and accountability feel like overhead in H3. In H1, they're non-negotiable — and the org that skipped them will find out the hard way.

The Migration — One Visual to Remember



Each arrow represents a metrics migration — a deliberate decision to stop measuring one thing and start measuring another. That transition is the work most organizations skip. **The metrics migrate. So must the conversation.** Most pilots fail not at the technology layer, but at this handoff — where nobody formally changes what success means.

LIVE EXAMPLE

Scout AI at PlayOn Sports

Our sales staff cover entire states worth of high schools — a massive, fragmented territory with constantly shifting contacts, programs, and budgets.

Before Scout AI

Manual research: search Salesforce, cross-reference multiple websites, manually compile school profiles before every sales call. Time-consuming. Inconsistent. Frustrating.

The Principle

"Drink your own Kool-Aid." If we're helping clients use AI to know their customers better, we should be doing it ourselves. Scout AI was born from that conviction.

Scout AI: Mapped to the Framework

Here's how our own initiative maps — honestly, not in hindsight. H3 was the easy part. H2 is where we are now. That's the whole point: H2 metrics aren't a victory lap. They're work.

Horizon	Status	What We're Measuring
H3	✅ Done	Could we pull structured school data from internal systems and the web? Output usable by a rep in the field? **Yes.** Gate passed.
H2	🔄 Current	How much time is saved per call prep? Is the right information actually surfacing? Are coverage rates improving? Still measuring — honestly.
H1	🎯 Target	Standard tooling for the full sales org. Conditional on H2 numbers holding — not assumed.

ⓘ H1 is conditional on those H2 numbers holding. That's not a hedge — that's the framework working as designed.

Three Killers of the H2 → H1 Transition

1 No Defined Blast Radius

If you haven't mapped which teams, workflows, and dependencies are affected, you'll hit resistance you didn't anticipate. Scope the impact before you scale.

2 Metrics That Don't Translate

"40% faster" isn't a business case — it's a technical metric. Remember Freestyle: 30 seconds was meaningless until someone attached $95K/year to it. Attach the number.

3 No Owner on the Other Side

Someone championed the pilot. Who owns it in production? If the answer is "the same person" — you haven't scaled. You've given someone more work and called it a win.

Killing a pilot ≠ failure.

It's a successful evaluation — *triggered by a KRI before the budget was gone.*

The Reframe Leaders Need

Pilot owners often become pilot *defenders* — they advocate for continuation because it's their project, their identity, their effort. The framework gives you permission to stop. The metrics give you the answer.

You learned what you needed to learn. You reallocated resources to something with a better evidence base. That's not a failure — that's exactly how a healthy innovation portfolio works.

Orgs that can't kill pilots can't scale them either. A pilot without kill criteria isn't a pilot. It's a science project.

The Mindset Shift

Each question demands a different evidence base. The migration between them isn't automatic — it's leadership work. Most pilots never formally migrate their metrics to answer question two. So they can never honestly answer question three.

H3 — "Can we build it?"

Technical feasibility. Binary answer. Time-boxed to 2 sprints. If you can't prove it's possible, stop.

H2 — "Is it worth it?"

Value metrics. Unit economics. The question most pilots never formally ask — they just keep running.

H1 — "Is this how we work now?"

Sustainability at scale. Adoption. Drift. The question you can only answer if H2 was done right.

Three Questions for Monday Morning

Run these against every active AI initiative. Not "the team." A name. Not "soon." A date. Not just success metrics — kill criteria too.

1

What horizon is it in right now?

H3 exploration? H2 validation? Stuck somewhere without a name? Naming it forces clarity on what you should actually be measuring.

2

What KPI triggers advancement — and what KRI signals to pivot or stop?

Not "we'll know it when we see it." A specific number. A specific threshold. Defined before the sprint starts. Make at least one metric live *outside* the technology — that's the Freestyle lesson.

3

Who is the single named person accountable for that decision, by what date?

One human being. One calendar date. No committees, no "the AI team," no "TBD." If you can't answer this, you've just found your bottleneck.

A pilot without an end date isn't a pilot. It's a science project.

Let's Connect

Slides

https://martinrojas.dev

Blog

nextsteps.dev

Get the Slides

Connect in LinkedIn