

AI BENEFIT REALISATION

Setting the baseline before an AI pilot starts

What to measure, for how long, how much it moves on its own, and who signs the record before the first model call.

Terence Kok with Dr. Christina Chua

<1 in 5

organisations tracked KPIs for their gen AI solutions

McKinsey State of AI, March 2025. Tracking well-defined KPIs had the most bottom-line impact of the twelve practices tested.

In August 2026, 37% of organisations attributed any EBIT impact to AI.

19%

slower, as measured, with AI
tools allowed

20%

faster, as the same developers
believed afterwards

METR randomised study of 16 experienced open-source developers, July 2025, early-2025 tools. METR now says the results no longer reflect current tools. The perception gap is the lesson.

Four ways a late baseline flatters the pilot

1 People misremember the start

Recollection gathered afterwards inherits the perception gap.

2 Pilots go where things are worst

Bad weeks tend to be followed by better ones anyway: regression to the mean.

3 New tools look better at first

Novelty effects fade after the first weeks.

4 Measures shift once the project is known

Teams tidy data and redefine metrics before go-live.

The baseline decides the result



A large, bold, dark blue '3%' is displayed within a white rectangular box. Above the box is a thick dark blue horizontal line.

gain against twelve weeks signed
in advance, including quarter-
end



A large, bold, orange '9%' is displayed within a white rectangular box. Above the box is a thick orange horizontal line.

gain against three busy weeks
reconstructed afterwards

The eight pilot weeks are identical. The first result sends the pilot back for a longer run. The second gets it scaled. Illustrative example.

8.4%

the smallest change a before-and-after can detect

With a weekly spread of 6% and eight weeks either side. A business case that needs 5% cannot be proven that way.

Run longer, add a control group, use the baseline data in the analysis, or accept that the benefit cannot be measured.

What finance signs before the first model call

Use case identifier	Primary outcome, as finance counts it
Data source and owner	A full business cycle of history
Average and weekly spread	Smallest detectable change
Quality guardrail	Expected disbenefits
Current cost to serve	Target and kill threshold
Plan for released time	Signatures, freeze date, change log



The baseline should be taken during project initiation so that performance is not impacted by any element of project implementation.

UK Infrastructure and Projects Authority, Guide for Effective Benefits Management in Major Projects, 2017.

Five questions before approving a pilot

1 Who pulled the data?

Was it recorded before anyone knew the pilot was coming?

2 Can the pilot detect the target?

Compare the week-to-week spread with the gain the case needs.

3 What shows quality slipped?

Every speed measure needs a guardrail.

4 Where does released time go?

Capacity, cost avoided, cost removed or revenue.

5 Has any definition changed?

Every change goes in the log, with a reason.

The baseline is the cheapest evidence an AI programme will ever collect.

Read part 2 of the series at terencekok.com



Terence Kok · AI & Innovation Strategist

SAVE · COMMENT · FOLLOW