The vendor says the agent completed the task. You say: show me the run.

Dr Peter McCann Strain
CTO, DPhil/PhD in AI from Oxford University

Swipe ››

THE OPERATING RITUAL

If a vendor can only produce the final output and a latency chart, you are being asked to buy a system whose failure modes are invisible to you. The output may be good and the dashboard green, but none of that tells you what happened in the run.

— THE REFRAME

What the trace MUST contain.

THE OLD QUESTION

What is your accuracy, and what model is under the hood?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

Show me one complete trajectory (the path of decisions, tool calls and verifications), and does the run ID resolve to the whole tree?

— THE CONTROL SIGNAL

Half a sentence

A completion percentage with no version, date, scaffold or scoring method attached is only half a sentence. TheAgentCompany's headline rate shifts with all four, and its published versions do not agree: a warning label, not a settled fact. Benchmarks are population priors; they cannot say where this run succeeded, where it failed, which evidence existed at the time, or who had authority to let it continue. The trace can.

SOURCE

TheAgentCompany benchmark (arXiv and OpenReview versions disagree); METR time-horizon work, reported with wide confidence intervals; OpenTelemetry GenAI and OpenInference trace specifications.

CADENCE

A trace tree is a procurement test.

01 Pull **one trajectory** end-to-end: decisions, tool calls, verifications, artefact, run ID.

02 Answer the **eleven plain questions** the trace must answer, end to end, and confirm one run ID resolves to the whole governance trail.

03 Demand **portability**: three exported runs, viewable in a tool your team controls.

Eleven nodes a complete run leaves behind.

Input

What entered the run?

Plan

What path did it choose?

Tools

What did it call?

Checks

What passed or failed?

Outcome

What changed?

Eleven nodes drawn as a tree, rooted at the user task: sandbox, tool gateway, planner, the action and observation branch, memory, sub-agents, verification, human checkpoint, and final artefact, with the governance trail joining back to the run ID. Each node carries its name and the failure to inspect when its span is missing. Open formats: OpenTelemetry GenAI and OpenInference, viewable in a tool your team controls such as Arize Phoenix.

— PUT IT ON THE CALENDAR

Email your top agent vendor this week for three exported runs: success, failure, near-miss. Open them without the vendor narrating. Ask what it read, which tools it touched, where the human entered, and whether the run ID resolves to the whole tree.

RITUAL DRIFT

Accepting a green uptime dashboard as the audit trail. It answers how often, how fast and how many, never what this run did or where the human approved it. Require trace export in an open format and reconstruct three runs yourself before signing.

— RUN THE OPERATING LOOP

Show Me the Run

Read the full essay — the argument, the sources, the figures and a reader-ready working artifact.
Substack petermccannstrain.substack.com · Medium @peter.mccann.strain · LinkedIn peter-strain-dphil-15a607128
New essays twice weekly, 2 June – 21 July 2026.

Next: E16 — The Dashboard Is Green

— THE STACK SO FAR

E15 · Essay 15 of 22 complete · Arc IV: Proof and accountability

YOU JUST ADDED

The trace tree

STACK LAYER LIT UP

Runtime evidence

YOU CAN NOW ASK

ask a vendor to show the run.

NEXT

E16 asks how a green dashboard can still be quietly wrong.

— THE ARTIFACT, CONTINUED

Eleven nodes a complete run leaves behind.

THE REMAINING NODES

Memory
what did the run read, write or forget?

Sub-agents
who else acted under this run ID?

Verification
which independent checks ran, and which passed?

Human checkpoint
where did a person actually approve?

Governance trail
which policies and reviews are linked to this run?

Sandbox reach
what could the agent touch, and what was off-limits?

Dr Peter McCann Strain

CTO, DPhil/PhD in AI from Oxford University

I build production AI systems and write about making agentic AI useful, inspectable, governable and safe enough for real work.

Follow on Substack for the full 22-essay series
petermccannstrain.substack.com