"The AI failed" is not a diagnosis. It is a way to learn nothing.

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

Swipe ››

— THE FRAMEWORK GAP

When an agent run breaks, the post-mortem reaches for a slogan instead of a map. The failure usually belongs to a specific layer, and without naming it, teams fix the most visible thing rather than the layer that actually missed.

— THE REFRAME

Failure has a location.

THE OLD QUESTION

Which model was involved?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

Which layer was supposed to catch this, and which layer missed it?

— WHY THE FRAME HOLDS

9 seconds to deletion

A Cursor-driven agent deleted the PocketOS production database and volume-level backups in a reported nine seconds (outage ~30 hours). Running it through the nine layers, the load-bearing misses land on layers 4 (tools), 6 (memory), 7 (stopping) and 9 (governance), not the model. That mapping is my analysis, not a claim Railway made.

SOURCE

Railway and OECD.AI (PocketOS), The Register, and Xu et al. (TheAgentCompany).

HOW IT WORKS

The stack is a failure map, read one incident at a time.

01 Ask per layer which was **supposed** to catch this and which **actually** missed.

02 Classify the fix (model, harness, orchestration or **governance**) before approving it.

03 Refuse the chained-agent fantasy: dependent steps multiply, so more agents is **not** more reliability.

— THE ARTIFACT

Nine layers in three bands

Inside the model
model / alignment / reasoning

Around the model
tools / routing / context / memory

Around the system
stopping / orchestration / governance

Three model-layer cards, four runtime-layer cards, two governance-layer cards. Every agent failure lands in exactly one cell.

— APPLY THE INSTRUMENT

Take one incident or near miss this week. Run the nine-layer map in three steps. (1) Name the layer that should have caught it. (2) Name the layer that actually missed. (3) Check whether the fix targets the missed layer or just the visible symptom.

WHERE TEAMS MISREAD IT

Approving a model upgrade as the fix. The familiar lever rarely matches the missed layer. Narrow the token, harden the staging boundary, or add a stopping rule outside the confused loop instead.

— READ THE FULL FRAMEWORK

The Nine Layers Where Agents Break

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: E05 — The Reliability Equation

— THE STACK SO FAR

E04 · Essay 4 of 22 complete · Arc I: See the object

YOU JUST ADDED

The nine-layer failure map

STACK LAYER LIT UP

Whole stack

YOU CAN NOW ASK

locate failure by layer, not by vibes.

NEXT

E05 asks how to measure reliability when failure is layered.

— THE ARTIFACT, CONTINUED

Nine layers in three bands

THE REMAINING NODES

Model band: 1 capability, 2 alignment, 3 reasoning, the layers a vendor upgrade actually moves.

Runtime band: 4 tools, 5 routing, 6 memory, 7 stopping, the layers your scaffold owns.

Governance band: 8 orchestration, 9 oversight, the layers no model patch will fix.

Read every incident through the nine-cell map before reaching for a model swap.

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