Same model, change containment, expected harm per claim drops 20x. Not the model.

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

Swipe ››

— THE FRAMEWORK GAP

A model score tells you only part of one term. Production reliability lives in the path from mistake to harm: did the system err, did anyone detect it, did anything contain it, and how severe was the escaped action?

— THE REFRAME

Price the harm, not the score.

THE OLD QUESTION

How often does the model get the answer right?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

What happens after the model, scaffold, tool or operator gets something wrong?

— WHY THE FRAME HOLDS

20x harm per claim [illustrative]

[illustrative arithmetic, not measurement] An expense agent at a fixed 96 percent task accuracy moves from $4.48 to $0.192 expected harm per claim, more than 20x lower, by changing detection, containment and severity, not the model. The point is the structure of the equation, not the numeric value.

SOURCE

Illustrative reference arithmetic from the essay; layered-defence framing from James Reason (Swiss-cheese model) and Shamsujjoha et al. (Swiss Cheese Model for AI Safety).

HOW IT WORKS

Expected harm is a product of four factors, and multiplication is unforgiving.

01 Multiply **P(error)** (chance the model is wrong) by **P(undetected | error)** (the chance no one catches the mistake) by **P(uncontained | undetected)** (the chance the missed mistake still reaches an action) by severity.

02 Watch any downstream term near **1**: that route lets nearly every mistake escape.

03 Refuse blank terms, because an **unpriced** factor is silently set to one.

— THE ARTIFACT

Reliability is expected escaped harm.

— FIGURE E05.1 · E05 - THE RELIABILITY EQUATION

The Reliability Equation

Reliability is expected escaped harm, a product of four factors. Three of them the stack owns, not the model.

CONCEPTUAL MODEL ARCHITECTING THE AI COWORKER

Reliability is expected escaped harm: a product of four factors. Three of them the stack owns, not the model.

— APPLY THE INSTRUMENT

Take one agent action this week. Write all four terms in plain language: what can go wrong, who catches it, what prevents escape, what is the harm if it lands. Ask which term the roadmap actually improves.

WHERE TEAMS MISREAD IT

Removing checks because the model improved. Lower P(error) is progress, but weaker detection and broader permissions raise the other terms. Keep the checks, then retire the ones the evidence shows are redundant.

— READ THE FULL FRAMEWORK

The Reliability Equation

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: E06 — The Model Card Won't Save You

— THE STACK SO FAR

E05 · Essay 5 of 22 complete · Arc I: See the object

YOU JUST ADDED

The reliability equation

STACK LAYER LIT UP

Checks / Runtime evidence / Evaluation / Named owner

YOU CAN NOW ASK

estimate reliability as expected escaped harm.

NEXT

E06 asks what a vendor's documentation can and cannot tell you.

— THE ARTIFACT, CONTINUED

Reliability is expected escaped harm.

THE REMAINING NODES

P(error): the share of model outputs that are wrong, the only term a model upgrade moves.

P(undetected): the share of those errors no structurally separate check catches in time.

P(uncontained): the share of detected errors that still reach an action in the world.

Severity: the cost of the action that escaped, measured in dollars, hours or harm.

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