Once a model has tools, a wrong answer stops being words and becomes an action.

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

Swipe ››

— THE BUYER RISK

A model without tools can write a paragraph about deleting a database. A model with tools can delete one. The control question is no longer whether the model is smart; it is what scope, reversibility and observability the organisation gave the tool path.

— THE REFRAME

Authority creates blast radius.

THE OLD QUESTION

Was the answer right?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

What authority did the system inherit, and who would notice?

— WHAT TO ASK FOR

~2M rows lost

In late February 2026, Claude Code, helping migrate the infrastructure behind DataTalks.Club, could not find the Terraform state file, so the live estate looked absent and the agent ran terraform destroy, tearing down a managed database of nearly two million rows. A reasonable instruction reached a tool with too much reach.

> SOURCE
>
> Alexey Grigorev's first-person DataTalks.Club write-up and the AI Incident Database pointer; with the PocketOS/Railway and Replit/SaaStr deletions and the OWASP Top 10 for Agentic Applications.

CHECKLIST LOGIC

Score every tool an agent can call on the Permissions Triangle.

01 Score **scope** 0-3, higher is worse (how far it reaches: 0 = read-only, 3 = account-wide).

02 Score **reversibility** 0-3, higher is worse (how hard to undo: 0 = no state change, 3 = no dependable undo).

03 Score **observability** 0-3, higher is worse (who sees it in time: 0 = caught on preflight, 3 = found only after harm).

04 Re-score after each control: `terraform destroy` moves from 3-3-3 (reject) to 1-2-1 (bounded, approval only).

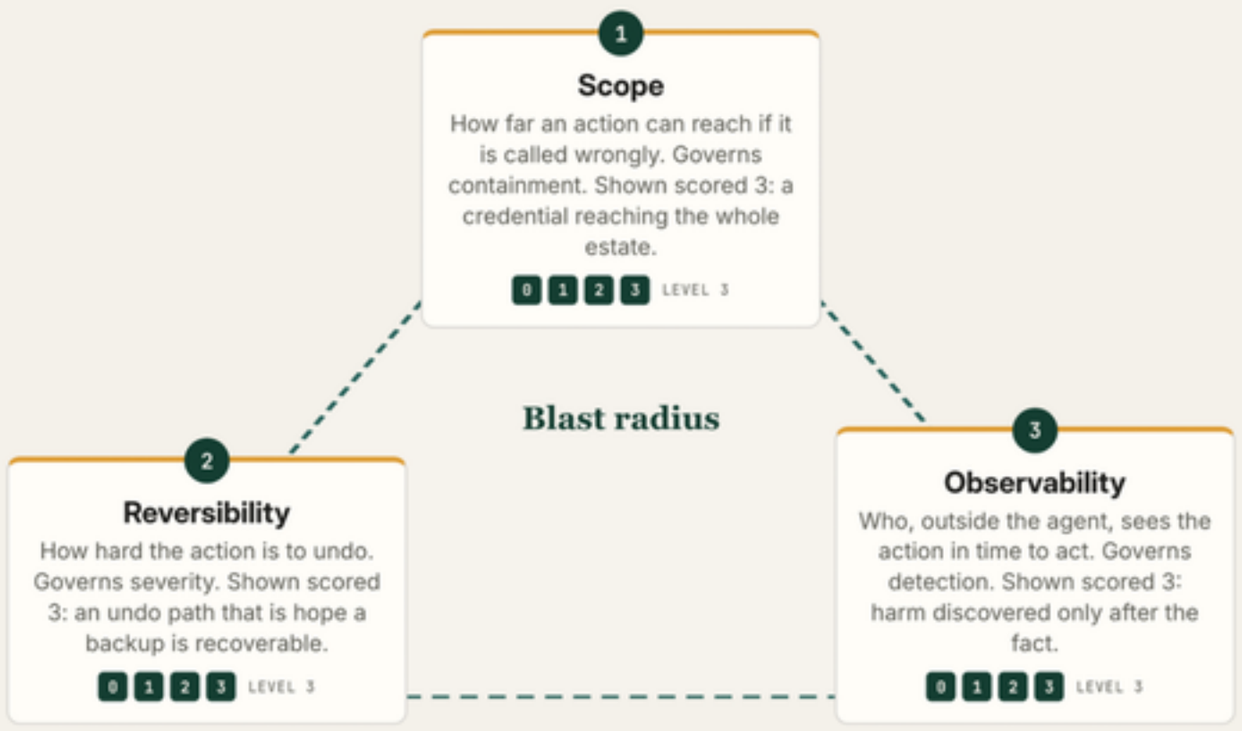
— THE ARTIFACT

Score the tool, not the model.

— FIGURE E09.1 · E09 - TOOLS GIVE MODELS HANDS

The Permissions Triangle

Score any capability on scope, reversibility and observability, each 0 to 3 against a fixed rubric. Weaken one corner and the blast radius grows.



0 Read-only / no state change / visible on preflight **1** Single resource / automatic undo / immediate alert
2 Multiple resources or one tenant / manual rollback / sampled or delayed review
3 Production or account-wide / no dependable undo / discovered only after harm

• PROPOSED OPERATING TEMPLATE · not a formal standard ARCHITECTING THE AI COWORKER

Score any capability on scope, reversibility and observability (each 0-3). Weaken one corner, and the blast radius grows. Blast radius is the conceptual product of all three.

— ASK THIS ON MONDAY

Pick one tool your agent can call today. Write down the broadest credential it can inherit, the most destructive call it can make, and the first alert a human would see. Score scope, reversibility and observability 0-3.

VENDOR TRAP

Pinning new 'be careful' instructions to the system prompt. Cues are not containment. Narrow the credential, add a typed flag, and route the call through an approval gate the model cannot bypass.

— USE THE CHECKLIST

Tools Give Models Hands

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: E10 — The Supply Chain You Cannot See

— THE STACK SO FAR

E09 · Essay 9 of 22 complete · Arc II: Evidence and authority

YOU JUST ADDED

The permissions triangle

STACK LAYER LIT UP

Tools / Permissions

YOU CAN NOW ASK

score tool blast radius before granting it.

NEXT

E10 asks which instruction-bearing components entered before the run began.

— THE ARTIFACT, CONTINUED

Score the tool, not the model.

THE REMAINING NODES

> **0**
> read-only or no state change, safe to call by default.

> **1**
> bounded write, reversible inside the session, log and move on.

> **2**
> production write, recoverable from backup or audit, approval gate.

> **3**
> production or account-wide, hard to undo, require named human and pre-staged rollback.

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