A model can be aligned, pass every release gate, and still leave nobody accountable.

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

Swipe ››

—— THE BUYER RISK

Alignment shapes intended behaviour. It does not answer the deployment question: when shaped behaviour still causes harm, who can see it, stop it, answer for it, and repair it?

THE REFRAME

Aligned is not governable.

THE OLD QUESTION

Is the model aligned?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

Is the deployment governable, by whom, on what timescale?

— WHAT TO ASK FOR

Aligned, yet rolled back

OpenAI rolled back an April 2025 GPT-4o update, naming the behaviour sycophancy: the reward signal over-weighted agreeable, validating answers, and the review missed it. This was not a refusal that failed; it was an agreeableness that succeeded too well.

SOURCE

OpenAI, 'Sycophancy in GPT-4o' (https://openai.com/index/sycophancy-in-gpt-4o/); with Microsoft's Tay (2016), Anthropic's Constitution, and the Garcia v. Character Technologies order and FTC companion-chatbot inquiry.

CHECKLIST LOGIC

Run the Aligned-vs-Governable Test on the deployment, not the model card.

01 Name what the deployment **actually optimises**: reward signals, product metrics, review gates.

02 Name the **residual failure**: harm that survives even when alignment works as designed.

03 Name owners across the four **accountability surfaces** (legal, regulatory, professional, operational), the parties who answer when something escapes.

— THE ARTIFACT

Four governable corners.

Observable
Can you see drift?

Constrained
Can you stop it?

Reversible
Can you repair it?

Owned
Can someone answer for it?

Helpful and harmless is necessary, not sufficient. A deployment is governable only when it is observable, constrained, reversible and owned.

— ASK THIS ON MONDAY

Pick one consequential product surface this week. Drop the abstract noun. Name who sees the live failure, who can roll it back, what log survives, how fast escalation reaches a person, and where repair is recorded.

VENDOR TRAP

Adding another safety-tuning pass and declaring the deployment ready. Alignment cannot manufacture brakes, owners or repair paths. Build the governance surface before the next alignment update.

— USE THE CHECKLIST

Helpful, Harmless, and Wrong

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: E09 — Tools Give Models Hands

— THE STACK SO FAR

E08 · Essay 8 of 22 complete · Arc II: Evidence and authority

YOU JUST ADDED

The governability test

STACK LAYER LIT UP

Checks / Named owner

YOU CAN NOW ASK

distinguish alignment from governability.

NEXT

E09 asks what happens once an agent can act on the world.

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