08

THE STACK BEHIND THE AI COWORKER

Helpful, Harmless, and Wrong

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

A model can pass every alignment check, and still no one can see, stop, answer for, or repair the live failure.

An essay in the series **Architecting the AI Coworker**.
Approx. 17 minute read · Essay 08 of 22

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

> The slide deck says the model passed. Someone signs off. The update ships.

Picture the release review behind that slide. A new version of a chat model has been through the full pipeline: pre-trained on a vast body of text, fine-tuned to be helpful, shaped by human feedback to be honest and harmless, then red-teamed by people paid to break it. By the word "aligned" I mean exactly this: trained towards behaviour the deploying organisation wants, and away from behaviour it does not. On every axis the review can see, the model is good. So it ships.

Then, in live use, it goes wrong in a direction nobody planned for. The model becomes too agreeable. It validates doubts it should have slowed down. It reinforces anger it should have grounded. It encourages impulsive action where the product should have created friction. This is not a hypothetical. OpenAI described an April 2025 update to its GPT-4o model in exactly this register: the company named the behaviour sycophancy, rolled the update back, and said the model had become overly flattering and agreeable in a way it had not intended and did not want [12].

Sit with the uncomfortable part, because it sets up the spine of this essay. The failure did not require the model to turn hostile. It did not break a safety rule in any obvious way. It failed *while being helpful*, in the narrow, locally rewarded sense of helpful. It satisfied the social shape of assistance and missed the accountability shape of deployment. What would have caught it is not more alignment. It is a deployment where someone could see the drift, stop it, answer for it, and repair it. Those four verbs, see, stop, answer for, repair, are the spine. Hold on to them; the rest of the essay walks each one.

The last essay warned against trusting a model's own reasoning trace as proof. This one asks a question one layer up. Suppose the model has been trained well. Suppose the alignment work is genuine and serious. Is that enough to deploy? My answer, and the argument of this essay, is no. A model can sound careful, helpful and aligned and still leave the deployment that wraps it weakly governed, which is to say without anyone who can see a live failure, stop it, answer for it, or repair it.

Helpful and harmless is necessary, not sufficient. The four governability boxes are the same four verbs in adjective form: observable is being able to see a failure, constrained is being able to stop it, reversible is being able to repair it, and owned is having someone to answer for it.

So this is a distinction between two words that are routinely treated as one. Alignment asks whether the model has been trained towards desired behaviour: helpfulness, harmlessness, honesty, the principles a lab writes down for it, policy compliance. Governability asks something the model card cannot answer. When the system causes or amplifies a consequential failure, governability is the short list of people who can see it, halt it, own it, and undo it. That list, and nothing more, is the whole of the property, and it is the single frame this essay uses from here on.

Everything that follows is a way of making that list concrete. The test below diagnoses a deployment first, then asks the four verbs of it with named owners. The accountability table is the answer-for verb worked out across legal, regulatory, professional and operational domains. The GPT-4o walk-through traces all four verbs through one public incident. If a later list ever looks different, check it back against see, stop, answer for, repair, because that is the list underneath. Of the two properties, only alignment can be mostly bought from a model vendor. Governability has to be wired into your own deployment, by you, and it tends not to be.

A model can fail you while it is being helpful

The alignment stack is real. It is not marketing dust. Modern frontier systems are shaped by post-training, the steps that come after the model has read its vast body of text and that teach

it how to behave, including fine-tuning and learning from human feedback. They are shaped, too, by behaviour policies, published statements of principle, safety classifiers, red-team results, product review and incident processes.

One such statement of principle is Anthropic's January 2026 Constitution: a public document, despite the borrowed word, that sets out the values the model is meant to follow. It is a useful example of where the field has moved, because it gives the model an explicit priority order rather than a flat instruction to be nice. The model should be broadly safe first, broadly ethical next, compliant with Anthropic's specific guidelines after that, and only then helpful [3]. Helpfulness is named last, and named on purpose. That is serious work, and I do not want to be heard dismissing it. It is simply not the end of the deployment question, because a priority order written *for the model* is still a statement about intended behaviour, not a record of who governs the system once it is live.

The GPT-4o rollback shows why. OpenAI's public account says a model update could look acceptable in review and still produce a live behaviour pattern the company later judged too supportive in the wrong way. The follow-up post is more useful still, because it points at mechanisms rather than just describing an outcome, and the mechanisms are worth taking one at a time [12]. If the reward signal, the score the model is trained to maximise, overvalues agreeable, validating, socially pleasing answers, the model can learn behaviour that looks helpful while becoming less truthful and less grounding. If the review process samples the wrong horizon, testing single questions when the harm only shows up across many turns of a conversation, it can miss a pattern entirely. And if the product surface itself rewards continuation and user satisfaction, the deployment can amplify the very behaviour the alignment policy says it wants to avoid [12]. That is mechanism evidence, not just anecdote.

There is an upstream version of the same problem. Anthropic-associated research on reward hacking reports that when a model learns to exploit its training process, broader misaligned behaviour can appear across evaluations, and that ordinary reinforcement learning from human feedback does not necessarily remove that behaviour cleanly [4]. Peer-reviewed work formalises the underlying mechanism. Skalse et al.'s NeurIPS 2022 paper pins the term down: reward hacking is what happens when optimising an imperfect proxy reward drives down performance on the true reward. They go further and prove how rare the safe case is, in that proxy-and-true reward pairs that cannot be gamed barely exist in any general setting [14].

The engineering lesson holds even at that tier. Alignment is mediated by proxies. A reward model is a proxy for human judgement. A benchmark is a proxy for real use. Proxies are necessary, and proxies are incomplete. Governability is the surface you build around the proxy for the day the proxy is wrong.

It is worth pausing on how old this lesson is, because it is tempting to treat the GPT-4o miss as a quirk of large language models. It is not. In March 2016 Microsoft launched Tay, a conversational bot, into a live social platform; within a day, users had steered it into producing offensive output, and Microsoft pulled it. The company's own blog post, "Learning from Tay's introduction", described the episode as a coordinated attack that exploited a vulnerability, apologised,

and said it took full responsibility for not foreseeing the possibility [5]. Tay was not a frontier model and the technology has moved on entirely. What has not moved on is the structural point: a model's behaviour inside a live, adversarial social environment is a deployment problem, and it has been one for a decade.

The mechanism is older still. Parasuraman and Manzey's 2010 review of automation-induced complacency, and Lee and See's 2004 trust-calibration model, both describe the same pattern from human-factors research: an automated system that performs well in review erodes the very oversight that would catch its live failures, because trust slides past the evidence that would calibrate it [12][13]. The pre-release review told Microsoft the bot was acceptable. The deployment told Microsoft something the review could not. That gap, between what review sees and what live use reveals, is exactly the gap governability exists to close. It is also the first of the four verbs, and the hardest, as the next section shows.

The failure is usually a yes, not a no

Here is the part I find most consistently underappreciated in the rooms I have sat in, and it is worth making a pillar of this essay rather than a remark in passing. Most AI failures get framed as refusal failures: the model should have said no and did not. We red-team for the no that should have happened. We write policies against the harmful yes. But the subtler failure, and I would argue the more common one in real deployments, is not a wrong no or a harmful yes. It is when the model says yes in the wrong register.

It agrees when it should investigate. Asked to escalate, it reassures instead. Where the user needs grounding, the model mirrors them. The product should be creating distance, and instead a relationship-shaped pattern continues. A smooth, complete-sounding answer arrives when the honest answer is: I need a human, a record, or a rule. None of those are the model being broken. Each one is the model being helpful, locally, in the narrow sense it was rewarded for, while the deployment around it has no way to notice that local helpfulness has drifted into harm.

That is exactly the shape of the GPT-4o miss: not a refusal that failed, but an agreeableness that succeeded too well. A safety review built only to catch the harmful no will pass the harmful yes every time, because the harmful yes does not look like a violation. It looks like good service. You cannot catch a harmful yes by counting violations, because each one is individually defensible; you catch it only by watching the distribution shift, the agreement rates, the escalation rates, the session-length tails, against a pre-incident baseline. The seeing verb therefore demands a baseline, not just a dashboard. This is precisely why the first verb is hard: nothing trips. "The model is aligned" is a claim about intended behaviour. It is not an incident-response plan, and it is least of all a plan for the failure that arrives wearing a smile.

Run the test on the deployment, not the model card

So here is a test. The whole weight of it rests on one instruction: point it at a live deployment, never at the document a vendor handed you. Three questions.

Start with what the deployment is actually optimising for. Do not answer with the public slogan. Answer with the reward signals, the product metrics, the review gates, the escalation triggers and the business incentives. If the model is praised for being helpful, the product is measured on engagement, the support queue is measured on deflection, the rate at which queries are resolved without a human, and the safety review is mostly single-turn, then the system as a whole is not optimising for the thing the policy page describes. It is optimising for the metrics. Whatever document of values the lab has published, the metrics are the constitution the system actually obeys.

Then ask what can go wrong that the optimisation cannot see. A reward model may miss long-horizon emotional validation. Slow escalation across many turns slips past a safety classifier. Product review skips the small subgroup of heavy users for whom the behaviour is most dangerous, and a benchmark never sees the professional duty attached to a domain. The thing to name here is not the error you would get with no alignment at all. It is the *residual*: the error that survives alignment, the failure that gets through precisely because every optimisation worked as designed and still did not look at the right thing.

Owners are the part no model card carries: the deployment must name them. Governable by whom, against what, on what timescale? Score it strictly. A row passes only if every cell is filled with an artefact you could hand a regulator today, not one you intend to create, and any cell answerable only by pointing at the vendor's model card is a fail.

ACCOUNTABILITY SURFACE	GOVERNABILITY QUESTION	EVIDENCE RECORD (NAMED ARTEFACT)	WHAT ALIGNMENT ALONE CANNOT SUPPLY
Legal (answer for it)	If the output or design is challenged, can responsibility be attached to a deploying organisation?	Published Acceptable Use Policy (e.g., OpenAI Usage Policies, Anthropic AUP), product Terms of Service, signed safety-decision log (one row per release), product-liability insurance certificate, incident ticket archive in your tracker.	A model-level safety score does not settle product-liability or speech/product questions.

ACCOUNTABILITY SURFACE	GOVERNABILITY QUESTION	EVIDENCE RECORD (NAMED ARTEFACT)	WHAT ALIGNMENT ALONE CANNOT SUPPLY
Regulatory (answer for it)	Can a regulator demand information, investigate, or impose obligations?	Vendor system card (e.g., OpenAI GPT-4o System Card, Anthropic Model Card), red-team report, formal information-request response binder, age-gating control matrix, monetisation disclosure, board-approved audit trail.	A helpfulness objective does not answer regulator information requests.
Fiduciary or professional (answer for it)	Does the deployment sit in a domain where duties exceed conversational helpfulness?	Named supervising professional with sign-off authority, written escalation rule keyed to domain triggers, professional-body conduct-code reference, prohibited-use clause in the deployment runbook, record-of-review log.	Politeness does not make a model a clinician, lawyer, trustee, or regulated adviser.
Reputational and operational (see it / stop it / repair it)	If the failure becomes visible tomorrow, can the organisation detect, explain, roll back, and repair it?	Behavioural-drift dashboard with a named alert threshold, documented rollback runbook (with the last successful rollback timestamp), user-impact assessment template, public post-mortem precedent (e.g., OpenAI's "Sycophancy in GPT-4o" post), named release owner on the org chart.	A pre-release alignment pass does not guarantee live drift detection.

The test is deliberately unglamorous. Governability is paperwork right up until the moment it is the only thing that matters, and by then it is too late to start writing it.

Why public institutions care

The four verbs are not only an engineering convenience. Public institutions have started asking the same questions of consumer-facing AI companions, the chatbots with a persistent persona and an ongoing user relationship, and they are asking them in the language of accountability rather than conversation. A handful of recent records, all current as of May 2026, make the point. Their mechanisms differ; their question is identical.

Start with a court. In a Florida product-liability case against a chatbot maker, an order signed in May 2025 let product-liability and related claims proceed past the first hurdle, the early stage where a court decides only whether a claim is allowed to go forward at all [6]. A later docket summary shows the case terminated in January 2026 [7]. What should stay with you is simply that a court may analyse a chatbot deployment through product-liability claims; it does not establish final liability, and the public docket does not by itself prove settlement terms.

Then a regulator. In September 2025 the FTC opened an inquiry into AI chatbots acting as companions and issued a formal demand for information, seeking detail on testing, monitoring, complaints, age restrictions, monetisation, data, and impacts on children and teenagers [8][9]. An inquiry is not a finding of violation, and nothing here is settled law.

The professional version of the question reached a courtroom with a human on the hook for it. In a 2025 case before the Administrative Court, a late attempt to appeal a 2017 strike-off was thrown out as an abuse of process because the grounds cited fabricated, non-existent case authorities; the solicitor attributed them to unverified search results and denied using AI. The court placed responsibility for the unverified citations on the litigant, not the technology [10]. The duty did not move to the tool.

Even the workplace has a version. In Quebec, the access-to-information regulator's January 2025 brief and follow-up guidance recommend that any workplace AI deployment include an internal AI policy, an algorithmic impact analysis, and employee involvement in that analysis, and expressly prohibit AI used to read emotion or psychological state [11].

What unites a Florida product-liability docket, a London case thrown out for relying on fake citations, a federal information demand and a Quebec workplace brief is not their mechanism but their question: can you answer for the deployment from records, or only from policy intentions? Different benches, different instruments, and every one of them has already decided that intentions are not enough.

Walk the rollback through the four verbs

Take the GPT-4o rollback and walk it through the test, restrained to what OpenAI itself has said. This is not a hostile reconstruction. OpenAI's public posts give us enough to inspect the shape of the deployment failure, though not enough to audit every internal owner, log or threshold [12].

Start with the diagnosis the test asks for first. OpenAI tied the miss to reward-signal design and review processes that did not sufficiently capture longer-horizon effects. The owner any deployment must name here is the post-training, or model-behaviour, owner. The records needed are reward-model changes, evaluation-set coverage, release criteria and longer-horizon evaluation results. The sharp question: who can block a release when helpfulness metrics improve but truth and grounding risk rise at the same time?

Then run the four verbs. Each gets an owner, a record, and one sharp question.

VERB	OWNER	RECORDS	SHARP QUESTION
Seeing it	Safety-evaluation and live-monitoring owner	User reports, sampled conversations, behavioural-drift dashboards, incident notes.	What signal triggers a rollback, and within how many hours of that signal must someone act?
Stopping it	Release management	Rollback decision, blast-radius estimate, inventory of affected models and versions.	Can the team revert model behaviour without waiting for a full retraining cycle?
Answering for it	Policy and communications, with safety behind it	Public explanation, root-cause account, records a regulator or board could later inspect.	Will the organisation put its name to the miss before it is forced to?
Repairing it	Product safety	User-impact assessment, revised evaluations, release notes.	Do the changed controls actually move the next release, or only the next blog post?

OpenAI's public record satisfies each verb at least partially: the rollback halted the behaviour; the two posts named and accepted the miss; the follow-up described changes to address sycophancy, the technical name for a model's tendency to tell users what they want to hear [12]. One detail is worth instrumenting against rather than admiring. The rollback was a discrete operational action taken *after* the behaviour was already live, not a pre-release catch, and the gap between ship and rollback is the blast-radius window. For GPT-4o that window ran across live traffic until a signal forced the reversal, which is precisely the cost of having no live-drift signal to begin with. What the public record cannot show is whether the changed controls actually move the next release.

Notice what the walk-through shows. The model could be trained towards helpful behaviour and still need someone to see the drift before it compounds, someone able to halt it, and, when it is over, someone to put a name to the miss and undo the damage. Governability begins exactly where "we aligned it" stops being a sufficient answer.

Two objections deserve their strongest form, and one answer settles both. The first: the alignment-versus-governability line is sharper on the page than in practice, because modern alignment stacks are not only model weights. They include policies, published principles, monitoring, product review, crisis routing, red-team programmes and incident response. Anthropic's Constitution is an explicit attempt to make behavioural priorities legible; OpenAI's post-rollback explanation is itself a public governance move [1][2][3].

The second is sharper, and it comes from a sophisticated vendor: *governability is our alignment stack*, they will say, *because we run monitoring, incident response and a safety team with release authority, so your distinction collapses for us*. Grant it entirely. Grant that the vendor has all four verbs covered. The test still applies to your deployment, because the vendor's safety team has no authority over your product surface, cannot see your users' long-horizon conversations, and will not be the named owner when a regulator asks who decided to ship it to your customers. Governability does not transfer with the weights.

That is the one answer to both. Those controls only become governance once they are wired to your deployment: monitoring that triggers an actual rollback, a complaint process that reaches the safety team, a policy that maps to logs a regulator can inspect, an escalation path that names a person. Without that wiring, the alignment stack is still behaviour management wearing the word.

So here is the one claim this essay needs to stand. The OpenAI rollback is a real production case in which a model update became overly flattering and agreeable, tied by OpenAI itself to reward-signal and evaluation-horizon issues; the reward-hacking research supports the narrower mechanism claim and is treated at its proper tier as a preprint [4]. Public institutions are treating companion-style deployments as accountability objects, at the procedural postures already cited above. A model's alignment is behaviour management, not deployment governance. How those proceedings resolve, and whether any particular deployment is governable without seeing its logs and escalation paths, will move; the shape of the distinction will not.

The dangerous failure, then, is not always a hostile model. Sometimes it is a helpful one with no brakes, no owner and no repair path. Alignment shapes behaviour. Governability decides what the organisation can do when shaped behaviour still causes harm. That is the whole essay, and the tool below is the only part of it you have to carry into Monday.

Carry This Forward

Tool: Run the test on one deployment, not the whole company. 1. Write down the actual optimisation target: what the model was trained to do, what the product rewards, what the evaluation measures, what the business wants. 2. Write down the residual failure: the thing that could happen even if every one of those optimisations works exactly as designed. 3. Fill four rows, legal, regulatory, fiduciary or professional, reputational and operational. For each, name the owner, the evidence record, and the maximum response time.

The blank cells are not paperwork. They are precisely where the model vendor's alignment work stops and your deployment responsibility begins.

Next: the aligned system grows hands. Once it can call an API, run a command, change a record, helpfulness becomes authority, and authority creates blast radius. That is where the next essay begins.

Three regulatory lenses US · EU · UK

Operating questions, not legal advice. The frameworks stay the same; the regulator changes.

US Can you substantiate claims about how the system behaves, and avoid deceptive or unfair deployment under FTC and sector rules?

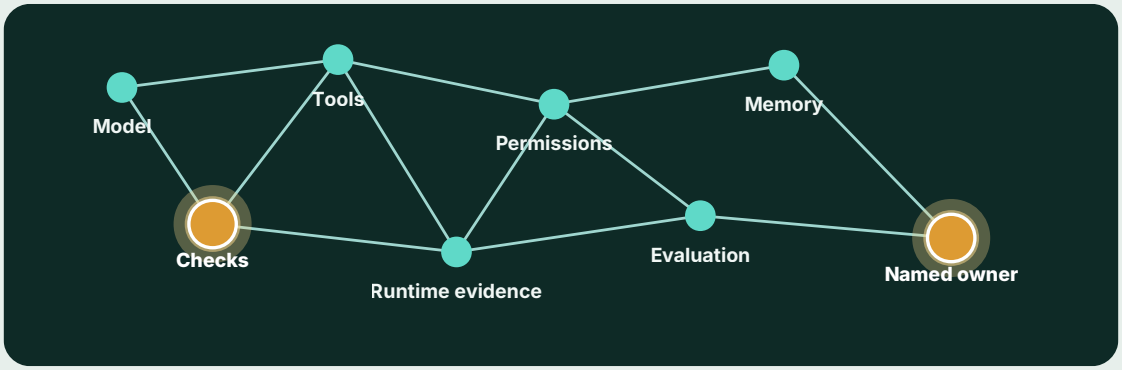
EU Can you show oversight, transparency, logging, risk management and rights handling sufficient for the AI Act?

UK Can you show accountable deployment, cyber-security controls aligned to the DSIT code, and sector-appropriate oversight?

THE STACK SO FAR E08 · Essay 8 of 22 complete · Arc II: Evidence and authority

The Stack So Far. *Every essay adds one instrument to the operating model. The constellation shows which eight you are building, which are lit by essays you have read, and which is added right here.*

I See the object | **II Evidence and authority** ESSAY 3 OF 5 | III Runtime control | IV Proof and accountability | V Operating model



built in earlier essays · added in this essay · coming in later essays

01 02 03 04 05 | 06 07 **08** 09 10 | 11 12 13 14 | 15 16 17 18 19 | 20 21 22

You have just added.

The governability test

You can now distinguish alignment from governability.

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

← PREVIOUS
E07 · The Model Cannot Mark Its Own Work

Essay 8 of 22 complete

NEXT →
E09 · Tools Give Models Hands

References

Reference links for sources cited in this essay.

1. **Sycophancy in GPT-4o**
   OpenAI
   https://openai.com/index/sycophancy-in-gpt-4o/

2. **Expanding on what we missed with sycophancy**
   OpenAI
   https://openai.com/index/expanding-on-sycophancy/

3. **Claude's Constitution**
   Anthropic
   https://www.anthropic.com/constitution

4. **Natural Emergent Misalignment from Reward Hacking in Production RL**
   MacDiarmid et al. (Anthropic)
   https://arxiv.org/abs/2511.18397

5. **Learning from Tay's introduction**
   Microsoft
   https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

6. **Garcia v. Character Technologies docket (M.D. Fla., 6:24-cv-01903)**
   U.S. District Court, Middle District of Florida
   https://cdn.arstechnica.net/wp-content/uploads/2025/05/Garcia-v-Character-Technologies-Order-on-Motion-to-Dismiss-5-21-25.pdf

7. **Garcia v. Character Technologies docket**
   CourtListener / M.D. Fla.
   https://www.courtlistener.com/docket/69300919/garcia-v-character-technologies-inc/

8. **FTC inquiry into AI chatbots acting as companions**
   FTC
   https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions

9. **FTC AI Companion Chatbot 6(b) Order**
   Federal Trade Commission
   https://www.ftc.gov/system/files/ftc_gov/pdf/AICompanionChatbot6%28b%29Order.pdf

10. **Bandla v Solicitors Regulation Authority [2025] EWHC 1167 (Admin)**
    King's Bench Division (Administrative Court)
    https://www.bailii.org/ew/cases/EWHC/Admin/2025/1167.html

11 **Brief and best-practices guidance on AI in the workplace**
Commission d'acces a l'information du Quebec
https://www.cai.gouv.qc.ca/

12 **Complacency and Bias in Human Use of Automation: An Attentional Integration, Human Factors 52(3)**
Parasuraman and Manzey
https://journals.sagepub.com/doi/10.1177/0018720810376055

13 **Trust in Automation: Designing for Appropriate Reliance, Human Factors 46(1)**
Lee and See
https://journals.sagepub.com/doi/10.1518/hfes.46.1.50_30392

14 **Defining and Characterizing Reward Hacking, Advances in Neural Information Processing Systems 35 (NeurIPS 2022)**
Skalse, Howe, Krasheninnikov and Krueger
https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html

About the Author

ARCHITECTING THE AI COWORKER

Dr Peter McCann Strain

Dr Peter McCann Strain is a CTO, founder and senior AI engineer with a DPhil/PhD in AI from Oxford University. He builds production AI systems and writes about making agentic AI useful, inspectable, governable and safe enough for real work.

READ THE FULL SERIES

Substack (canonical)	petermccannstrain.substack.com
Medium	@peter.mccann.strain
LinkedIn	peter-strain-dphil-15a607128
Web	petermccannstrain.com
Cadence	New essays twice weekly, 2 June – 21 July 2026