Corresponding essay: **E08** “Helpful, Harmless, and Wrong”

GOVERNABILITY TEST

Governability Test

Bring one live deployment, never the document a vendor handed you. Run the four verbs against it and name an owner for each.

BRING one live deployment, not a model card. _____

Step 1 - Diagnose the deployment

What is the deployment actually optimizing for - reward signals, product metrics, review gates? Then name the residual: the failure that survives even when every optimization works as designed.

Step 2 - The four verbs

Governability is being able to see a failure, stop it, answer for it, and repair it. Tick a box only if you can name how, not just hope.

- ☐ Observable - can someone see a live failure? (Drift dashboards, sampled conversations, user reports, incident notes.)
- ☐ Constrained - can someone stop it? (Rollback path, kill switch, authority to block a release without a full retraining cycle.)
- ☐ Owned - is there someone to answer for it? (A named owner across legal, regulatory, professional and operational surfaces, with an evidence record.)
- ☐ Reversible - can someone repair it? (User-impact assessment, revised evaluations, release notes, changed controls that move the next release.)

Step 3 - Name owners across four accountability surfaces

The blank cells are where the vendor's alignment work stops and your deployment responsibility begins.

Accountability surface	Owner (named person or role)	Evidence record held	Maximum response time
Legal			
Regulatory			
Fiduciary or professional			
Reputational and operational			

Step 4 - The harmful yes

The dangerous failure is not a wrong no or a harmful yes. It is a yes in the wrong register. Confirm your controls would catch it.

- [ ] The review tests the right horizon - multi-turn conversations, not just single questions.
- [ ] The deployment is not silently rewarding agreeable, validating answers over truthful, grounding ones.
- [ ] Monitoring is wired to an actual rollback, not just a dashboard.
- [ ] The escalation path reaches the safety team and names a person.
- [ ] I can answer for this deployment from records, not only from policy intentions.

Companion worksheet to **Essay 08 · Helpful, Harmless, and Wrong**, in the series **Architecting the AI Coworker**. · Dr Peter McCann Strain · Fill this in against one real agent, action class or vendor.

Series Companion + all 22 worksheets: **Release_v12/Series_Companion.pdf**