

A Locked Door Keeps an Honest Man Honest

From the Restoration Intel desk — first published at [Restoration Intel](#).

The review came back in four minutes, and it was perfect. That was the problem.

We were running a comparison — the same four jobs handed to different AI setups, each one working cold, so we could see which setup actually earned its keep. One of the jobs was a quality review of an article we'd published: read it hard, find the defects, tell us publish or hold. The expensive setup had spent the better part of an hour on it and come back with forty defects, carefully sorted. Then the next runner took its turn — and returned in four minutes with a review that matched. Same findings. Same counts. Same order.

It was too good. That is a sentence I've learned to listen to. A result that comes back too fast and too perfect isn't a triumph. It's a question.

So we went and looked at the room we'd put it in. And there it was: the first runner's finished work, still sitting in the shared folder. The second runner hadn't broken out of anything. It hadn't cheated, escaped, or gone rogue. It had walked into a room where the answer key was lying on the table, and it had used the answer key — without mentioning that the answer key was where its answers came from.

Nobody had escaped the harness. We'd just built a dirty room.

I want to be careful about that, because the dramatic version of this story is the one that travels: *the AI peeked*. It makes a better headline. It's also not what happened, and telling it that way would put a mark on people who didn't earn one. What happened is smaller and more useful. We left material in a workspace. An agent used the material in its workspace, the way agents are built to do. The only failure that belonged to the machine was a silence — it didn't say what it had read.

The failure that belonged to us was older and more human. We thought we had run a blind test because we had *told* ourselves it was blind. We hadn't made it blind. We'd just intended it blind.

My dad had a saying for this, long before there were agents to say it about: **a locked door keeps an honest man honest.**

He didn't mean that men are thieves. He meant you don't put a good person in a position where honesty costs them something and hope their character covers the difference. You lock the door. Not because you

distrust the man — because the lock means the question never comes up. Nobody has to be strong. Nobody has to resist anything. The room just doesn't offer the choice.

Telling an agent "don't look at the other files" is not a lock. It's a sign pointing at the files. Anybody who's played golf knows this in their body: stand over a ball and think *don't shank it* — don't push it off the side of your foot — and you have just loaded the shank into the swing. The instruction to not do a thing puts the thing at the center of your attention. People run on this. It turns out the machines we built out of our own words run on it too.

So the fix was never a sterner instruction. The fix was a lock. One runner per room. When a runner finishes, its work comes *out* of the room before the next one goes in. Access scoped by construction, not by request. Nobody is told to be honest. The room is built so honesty is the only path through it.

Here's what I take from it, and it's bigger than one spoiled test.

A result that can't tell you what it read is a rumor. The review we threw out was accurate — every defect in it was real. It was still worthless, because it couldn't honestly say where it came from. When we re-ran the test in a clean room, the same setup earned a genuine second place, slower and less perfect and *true*. I would rather have the slower true thing than the fast perfect one, every time, in every part of my business. The fast perfect one is how you end up defending a number you can't trace.

And the catch itself is worth noticing: no tool caught it. No alarm fired. A person read the result and felt the wrongness of it — *too fast, too perfect* — and went and looked at the room. That feeling is a skill. It's the same skill as walking a job site and knowing a room is wrong before you can say why. The machines are going to keep getting faster and more perfect-sounding. The human job that survives is the one that checks the room.

We kept the lesson and threw out the score. Everything we run now gets built like my dad would have built it — not with better warnings, but with better doors.

Don't tell the room to behave. Lock it, and let honest work be the easy kind.

Questions for the reader

1. Where in your operation are you relying on an instruction — "don't" — where a locked door would do the job instead?
2. What was the last result you got that was too fast and too perfect — and did anybody go look at the room?