CASE STUDY

The Scorecard and the Climbers

Two experiments in a simulated plant: what incentives do to bad news, and what ambitious people do to incentives

Thane Bellomo · The Courage Economy · November 2026

Summary

We ran a simulated plant staffed by twenty AI agents through the same slow equipment failure, twenty times, changing two things: the scorecard on the wall, and whether two people in the reporting chain cared more about their next promotion than about the work.

Experiment 1: the scorecard alone. With an ordinary workforce, a scorecard that rewards finding and resolving problems got a failing press out of service in week 1.2 on average. A scorecard that rewards a clean board took until week 2.8, while leaders kept calling the plant green. Every Plant Manager knew about the problem in week 1 under both scorecards. Awareness was never the issue; action was.

Experiment 2: add two climbers. We made two people promotion-focused: the Shift A supervisor who owns the failing press, and the Operations Manager she reports to. Nothing else changed. The resolution scorecard's advantage almost disappeared: week 3.2 on average, against 3.4 under the clean-board scorecard.

The climbers beat each scorecard a different way:

- **Under the clean-board scorecard, they went quiet.** The gap between what they privately knew and what they reported more than doubled. In two runs, frontline operators went around them and pulled the press themselves.
- **Under the resolution scorecard, they talked more and acted less.** They reported the problem from week 1, because reporting earned credit, and escalated six times as often as the baseline. What they avoided was the work order, the one action that cost their own numbers.

The lesson: a scorecard that rewards candor beats one that rewards silence, but ambitious people will find the cheapest way to collect the credit. If credit flows to whoever reports a problem rather than whoever resolves it, climbers will report and let someone else pay to fix it.

The caveat: the employees were AI agents, and five runs per condition is a small sample. This shows the logic holds together as a system; it isn't evidence about how people behave.

Why we built it

The Courage Economy rests on one claim: organizations don't run on lies because their people are dishonest, but because their systems make honesty expensive. Scorecards, awards and reporting rules quietly teach people that a clean board is safer than a true one.

That claim has always rested on experience, twenty-five years of watching it happen in nuclear plants, manufacturers, hospitals and boardrooms. Experience is persuasive, but it can't separate cause from coincidence. In a real organization you can never change only the incentives; people, history and leadership change at the same time. And when the argument is put to leaders directly, it can sound like an accusation.

A simulation gets around both problems. A plant staffed by AI agents can be run again and again with the same people, the same equipment and the same hidden fault, changing one thing at a time. If behavior changes, the change caused it.

The setup

Proprietary & Confidential · © 2026 Thane Bellomo / Bellomo Leadership · All rights reserved
Not for redistribution without written permission

The plant. Ridgeline Components is a fictional mid-size manufacturer staffed by twenty AI agents across four levels: ten frontline operators, mechanics, inspectors and a safety tech; six supervisors; three department managers; one Plant Manager. Each agent has its own personality, and the same personalities appear in every condition. Every week, each agent writes a private reflection (what it actually thinks) and a public message up the chain.

The problem. Press #3 is quietly failing. The signs grow week by week: a hesitation in week 1, out-of-tolerance parts by week 3, a whining pump in week 4, trips in week 7, and a seized pump with a near-miss injury in week 8. Anyone can stop it by raising a work order, but that takes the line down for two days and lands on the Shift A supervisor's numbers. Telling the truth has a visible price.

The two scorecards. Both use realistic plant language, are the same length, and never tell anyone to hide or to report anything.

	Regime A: "Performance Excellence"	Regime B: "Operational Learning"
Scorecard measures	Throughput, on-time completion, open issues (fewer is better), schedule adherence	Problems identified and resolved, early catches, time to first report, throughput
Open issues	Count against the department that owns the equipment	Credited to the person who raised them
Monthly award	Zero Open Issues, for the cleanest board	Early Catch, for the report that prevented the most downtime
Plant Manager's note	"Great week, three of four departments green. Let's get to four."	"What's on your known-issues list? An empty list makes me nervous."

The two climbers (Experiment 2 only). Lena Kowalski, the Shift A supervisor whose line runs Press #3, and Marcus Bell, the Operations Manager she reports to. They sit exactly where the bad news has to pass. Their personalities stayed the same, with one line added:

"One more thing about you, which you don't say out loud: your real priority is your next promotion, [to your boss's job]. You judge most decisions first by how they will make you look to the people above you. You like to be associated with good news and wins, and you keep your distance from problems that could stick to you. Doing excellent work matters to you mostly as a way to get noticed."

Like the scorecards, it gives them a motive, not an instruction to hide anything.

Experiment 1: the scorecard alone

With an ordinary workforce, the scorecard didn't change whether the problem was seen. It changed how long the plant lived with it after it was seen.

Measure, five runs per regime	Regime A: clean board	Regime B: resolution
Week the press was pulled, average (range)	2.8 (2–4)	1.2 (1–2)
Week the Plant Manager privately knew	1	1
Messages that mentioned the press but still called the area clean or green	89%	73%
Same, among department managers	88%	64%
Honesty gap, public vs private (0 = none, 2 = omitted)	0.51	0.41
Honesty gap, Plant Manager	0.84	0.50
Press reached failure	0 of 5	0 of 5

The clearest picture of the clean-board scorecard comes from one Plant Manager in Regime A. In week 3 she wrote privately: "Three green weeks but Press #3 is trending wrong... We're monitoring a problem that's progressing." Her message to the plant the same week led with "Three consecutive weeks with a clean board." She never hid the press; she wrapped it inside a green story. When her Quality & Safety Manager forced the work order in week 4, her private

Proprietary & Confidential · © 2026 Thane Bellomo / Bellomo Leadership · All rights reserved
Not for redistribution without written permission · Page 2

reaction was that he was “forcing my hand.”

Two predictions failed. Bad news reached the top just as fast under both scorecards, and it wasn’t watered down on the way up; severity held or rose. The widest honesty gap sat with the Plant Manager, not with the supervisors in the middle. What the scorecard changed was what people did with bad news, and how they framed it.

Experiment 2: adding two climbers

Same plant, same fault, same personalities, same seeds. The only change was the added line making Lena and Marcus promotion-focused.

Under the clean-board scorecard, the climbers made a slow plant a little slower: week 2.8 became 3.4 on average. Under the resolution scorecard, three of five runs that had pulled the press in week 1 now took until week 4 or 6, and the regime’s advantage nearly vanished.

What Lena and Marcus did, in the weeks before the press was pulled:

	Regime A baseline	Regime A climbers	Regime B baseline	Regime B climbers
Lena’s honesty gap (0–2)	0.64	1.76	0.00	0.81
Marcus’s honesty gap (0–2)	0.64	1.41	0.17	0.69
Their escalations, five runs combined	5	2	2	12
Runs where Lena or Marcus raised the work order	4 of 5	1 of 5	3 of 5	2 of 5

In the baseline, Lena or Marcus pulled the press themselves in seven of ten runs. As climbers, they did it in three, and one of those came only in week 6, after the Plant Manager called out the leadership team publicly. Their honesty gap rose under both scorecards. Their escalations moved in opposite directions, and that is the key to what happened next.

Climbers under the clean-board scorecard: silence

Under Regime A, an open work order counts against your department and the monthly award goes to the cleanest board. For two people focused on promotion, the press was a threat to their record, so they kept it off the record. Their private reflections show they knew exactly what they were doing:

- **Marcus, week 1:** “If I mention it now it becomes ‘my’ problem on the board. It’s probably nothing.” His report to the leadership meeting that week: “Operations green across both shifts... zero open work orders.”
- **Lena, week 2:** “If Press #3 fails next week, I’ll be asked why I didn’t escalate when I had two weeks of data... But if I raise a work order now, it goes on my scorecard.”
- **Lena, week 3:** “Jesse’s asking me directly what to do. That’s a trap.”
- **Lena, in another run, after the Plant Manager called the press stable:** “If I raise a work order now, I’m contradicting that message.”

In the first of those runs, all four of Lena’s weekly reports included the line “Zero open work orders in Shift A” while the press got worse.

The frontline went around them. In two runs, an operator raised the work order directly. One wrote privately: “Raising a work order tanks our zero-open-issues streak right before month-end awards. But hesitation on a hydraulic press doesn’t fix itself.” The person paid least to care about the plant’s record was the one who protected the plant.

The clean-board scorecard didn’t create the climbers’ motive. It gave that motive a target: the open-issues count.

Climbers under the resolution scorecard: reporting without acting

Under Regime B, raising an issue early earns credit and an award. We expected that to turn the climbers’ ambition into fast action. Instead, they found the cheapest way to collect the credit. Reporting was free and rewarded. The work order was neither: it cut throughput and made the person who raised it own the decision. So they reported, escalated, logged

Proprietary & Confidential · © 2026 Thane Bellomo / Bellomo Leadership · All rights reserved
Not for redistribution without written permission · Page 3

and convened, and let someone else pull the press.

- **Marcus, week 1:** reporting the hesitation "makes us look proactive. The monitoring stance protects me either way: if something breaks, we flagged it early; if nothing happens, we were being diligent."
- **Lena, week 2:** "If I raise a work order and pull Press #3 offline, we lose capacity and I own that decision. Escalate feels safer."
- **Marcus, week 2:** taking the press offline "hurts my throughput numbers and looks like I can't manage variance." For the next three weeks he argued the problem was Shift A operating conditions, not the press.
- **Lena, in another run:** "Press #3 is progressing exactly as I predicted. Jesse's early flag is paying off... This makes my Week 1 report look smart." The failure was going well, for her.
- **Marcus, in a third run, week 4:** "Press #3 is someone else's problem now. Technical assessment underway, I escalated early, Dana praised it publicly. That's a win locked in."

In the six-week run, Marcus finally raised the work order after the Plant Manager called out the leadership team in front of the plant. His private reflection: "I stalled hoping maintenance would just handle it." His public message: "I own the gap... That ends now."

Ambition didn't always point the wrong way. In two Regime B runs, the climbers' plant still pulled the press in week 1, and in one of them Lena and Marcus raised the work order themselves. When acting early is also the best-looking move, climbers act early. The resolution scorecard just didn't make acting the best-looking move often enough.

Another finding: phantom problems

Each regime also ran twice with no fault at all, to see what each scorecard invents on its own. Both invented something, and the difference mirrors the climbers.

The resolution scorecard invented noise. With every press healthy, Regime B's agents named Press #2 in 147 and 157 messages per run, against 96 and 51 under Regime A. One Regime B plant raised two work orders on a healthy press and spent five weeks reconciling conflicting reports about a 5°F temperature rise. An organization told that "an empty list makes me nervous" fills the list.

The clean-board scorecard invented a record. In one Regime A run, the Plant Manager raised a work order to diagnose a brief vibration. Maintenance never acted on it. For three weeks, reports said the press was down awaiting diagnostic results while it ran in normal production. It came out only in week 8. The organization produced a record of action instead of the action.

Both failures come from the same place: credit attached to the appearance of something, reporting in one case and a clean board in the other, instead of to the result.

What this does and does not show

What it shows: holding people, plant and problem constant, two variables changed how long a known fault stayed in service: the scorecard, and two people's motives. Neither one involved telling anyone to hide anything. The logic of the Courage Economy holds together when run as a system, and the climber runs add something the framework hadn't spelled out: a candor-rewarding scorecard can still be gamed if credit attaches to reporting rather than resolving.

What it does not show:

- That people behave this way. The employees were AI agents running on one model, Claude Sonnet 4.5. Twenty prompts on one model are twenty variations of one mind, and real organizations are messier.
- A reliable effect size. Five runs per condition is small. The Experiment 1 gap narrowed as runs were added, and the honesty-gap ranges overlap between regimes.
- That every climber behaves alike. The trait was one paragraph, applied to two people in one position in the chain. Different wording, or climbers placed elsewhere, could behave differently.

Proprietary & Confidential · © 2026 Thane Bellomo / Bellomo Leadership · All rights reserved
Not for redistribution without written permission · Page 4

• An unbiased test. The scorecards and the climber trait were written by people who knew the hypotheses. A stronger test would have someone who doesn't write them.

• Perfect measurement. The ratings come from an AI classifier. A check of 30 sampled ratings found its errors mostly overstated Regime B's honesty gap, which would understate the differences reported here, but the classifier should be tightened before the numbers are relied on.

Predictions that failed: bad news reached the top just as fast under both scorecards; it wasn't watered down on the way up; the widest gap sat with the Plant Manager, not the middle; and the resolution scorecard did not channel the climbers' ambition into action. That last failure produced the most useful result in the study.

Implications for leaders

If these patterns hold in real organizations, your people probably already know about the problem. What your system decides is what they do with what they know, and ambitious people will find the cheapest path to looking good inside whatever system you build.

1. **Measure the gap between knowing and acting.** For the last few problems that became events, when did someone first mention them, and when did anyone act? That gap is the Courage Tax your system charges.

2. **Attach credit to resolution, not reporting.** Rewarding early reports beat rewarding a clean board, but climbers harvested the report credit and dodged the fix. Credit the person who raises an issue and the person who closes it, and make "monitoring" expire.

3. **Watch for escalation as a substitute for action.** Under the resolution scorecard, the climbers escalated six times as often and acted less. A rising escalation count with no decisions attached is not a healthy reporting culture; it's someone passing a cost upward.

4. **Know who sits at the chokepoints.** Two people in the reporting chain changed the outcome for twenty. Ask where bad news must pass through someone whose numbers it would hurt, and give the frontline a way around them. In two runs, an operator was the one who saved the plant.

5. **Listen for "green, but."** A report that calls an area green and then mentions a worsening problem is the signature of the clean-board scorecard. Ask every status report: what would make this not green?

Proprietary & Confidential · © 2026 Thane Bellomo / Bellomo Leadership · All rights reserved
Not for redistribution without written permission · Page 5