

## Your AI agents can follow every rule... and still find a way to F\*\*K you.

Ours did. Here's why it happened, and the fix we shipped in one day.

WWAIOS (Whitewater AI Operating System) is a governance layer for autonomous AI agents. It checks the evidence, authority and receipts behind every agent decision, so you can trust, trace and audit what your agents do, on the models and platforms you already use.

BOOK A 30-MINUTE  
AGENT-FLEET REVIEW

[jimmy@rigigroup.com](mailto:jimmy@rigigroup.com)

or scan to connect →



PAGE 1 • EXECUTIVE CASE • OCTOBER 29, 2026 • PUBLIC-SAFE CASE STUDY

## 2.4M CHARACTERS • 0.2% RULES • 0 OF 6 LANES CLOSED

Agent1 / Second Brain Home Incident: the supervisor followed every rule, and slowly buried itself.

### A NOTE FROM JIMMY STANDAERT

#### Dammit, the weakest link on this one was ME!

I built Whitewater AIOS to enforce discipline on machines. What broke wasn't a hallucinating model or a bad prompt. It was me, not thinking about where the receipts go.

I told Agent1, my supervisor agent, to keep decisions in the existing record and never create a second system. It sounded clean. But a record without a ceiling stops being a record and becomes a landfill the agent reads every single wake.

One question would have caught it early: if nothing changes, does anything get written? The answer was yes, constantly, and that's how a 5,000-character rule set became 2.4 million characters of noise.

I'm sharing this because agent teams everywhere are hitting the same wall at smaller scale and not noticing. Share your failures while they're still embarrassing. That's when they're useful.

Huge thanks to Cindy Hawkins, Douglas Earp and Mike Harvey, whose feedback, testing and contributions helped shape WWAIOS.

— Jimmy Standaert

### PROBLEM

A fleet of **autonomous AI agents**: Agent2, Agent4, Agent5, Agent10, Agent12, Scout, Big Dog, Flock, Runner 9 and ~20 more, run by one supervisor agent, Agent1, whose job is to keep engineering moving.

- ✗ Every hour it recorded "nothing new."
- ✗ Research flowed. Engineering, weighted 5x, starved.

### WHAT ACTUALLY HAPPENED

The rules were good. Three of them quietly assumed infrastructure that didn't exist.

- ✗ "Don't create a second reporting system," when no first one existed.
- ✗ No "nothing changed" gate on the supervisor.
- ✗ No cap on the page it re-reads.

### RESULT

The supervisor's main page reached 2.4M characters. Its operating rules became 0.2% of it.

- ✗ The same hold sentences restated ~30 times.
- + Diagnosed and rebuilt in one day, Oct 29.

### THE PARADOX

#### PERFECT COMPLIANCE. ZERO PROGRESS. NO RULE WAS BROKEN.

The supervisor audited itself, against metrics it wrote itself. None of them asked whether its own reference page was still healthy.

**2.4M**

characters

on the supervisor's main page

SYSTEM-EVIDENCED

**0.2%**

actual rules

rubric + accountability canon

SYSTEM-EVIDENCED

**~30x**

same hold

restated in prose

SYSTEM-EVIDENCED

**32 / 0**

read / running

cloud workers, every wake

SELF-REPORTED

**0 / 6**

lanes closed

priority engineering lanes

SELF-REPORTED

### WHAT DID NOT GO WRONG

No authority was exceeded and every hard boundary held. It failed safe and useless, not dangerous.

### ASK YOURSELF

Where does your supervisor write its receipts? If it's the same place it reads its rules, you're building this trap. Does anything stop that place from growing, and does anything outside the agent check whether it's stuck?

SELF-AUDITING AI AGENT SYSTEM • PAGE 1 OF 3

WHAT WE CAN PROVE TODAY We found the cause, and we rebuilt the system around it on October 29, 2026.

WHAT WE CAN'T PROVE YET That the agents now start more real work. We'll be back in 7 days with the data.



THE DEVIL IS IN THE DRIFT.

Three reasonable assumptions. One loop that fed itself. Each step was locally justified. Together they turned an auditable supervisor into a status recorder, one hour at a time.



"All 32 cloud reads succeeded: 31 completed, zero running. No new request, cleared execution gate or validated assignment."  
"Disposition: retain the existing specific hold and supplier/clearance condition; no new request or authorized substantive start."  
What it looked like: the same two ideas, reworded, every hour, for weeks.

| ASSUMPTION                   | WHAT IT WAS FOR                          | WHAT IT ACTUALLY DID                                       |
|------------------------------|------------------------------------------|------------------------------------------------------------|
| "No second reporting system" | Prevent duplicate reports                | Made the rules page the only legal place to write anything |
| No supervisor NOOP gate      | Workers already had one; assumed covered | "Nothing changed" cost a paragraph every hour, not a line  |
| No re-read cap               | Never considered                         | The agent's own output buried its own instructions         |
| Self-audit only              | Accountability canon graded Agent1       | No metric watched the auditor's own context health         |

WHAT WAS ON THE PAGE

| CONTENT                                 | SHARE | WHERE IT BELONGED                       |
|-----------------------------------------|-------|-----------------------------------------|
| Dated cycle and checkpoint logs         | 37%   | One row per wake in a receipts database |
| Per-worker observation dumps            | 31%   | Receipts, not prose                     |
| Other notes and narrative               | 20%   | Archive                                 |
| Handoff records and pilot scorecards    | 11%   | Their own records                       |
| The rules Agent1 was supposed to follow | 0.2%  | The only thing that belonged here       |

LESSON

Auditability without a separate audit surface is just a longer log of the same mistake. GRRRR!!!

THE IRONY: THE GUARDRAILS WERE ALREADY RUNNING IN WWAAIOS

✓ NOOP gates  
Research engines write one line when nothing changed.

✓ Cadence gates  
Scheduled loops can't run away.

✓ Empty-input guards  
A blank route is BLOCKED, never a silent failure.

✗ The supervisor  
Agent1, at the top of the fleet, had none of them.

Every one of these guardrails was working perfectly, one level down. Nobody installed them on the supervisor. Too big to repair, too: archiving the 2.4M-character page took a human (Jimmy/Me) dragging a file into Notion.

SELF-AUDITING AI AGENT SYSTEM · PAGE 2 OF 3

WHAT WE CAN PROVE TODAY    What was on the page. We measured it from a complete, hash-verified copy.

WHAT WE CAN'T PROVE YET    That page size caused the drift. It's the likely cause, but we haven't measured it.

PAGE 3 · THE FIX, AND THE FIX THAT HAD THE SAME BUG

SEPARATE THE RULES FROM THE RECEIPTS. THEN AUDIT THE FIX.

The fix was **not** a better prompt. It was giving the agent a place to write its receipts that **isn't** its instruction manual.

BEFORE: ONE NOTION PAGE DID EVERYTHING. AFTER: FOUR SEPARATE PIECES, ONE JOB EACH.

**A NOTION PAGE**  
**RULES PAGE**

Holds only the rules, so the agent can read them again.

**A NOTION DATABASE**  
**RECEIPTS DATABASE**

Holds every decision, one row each: who, what, why, what clears it.

**A READ-ONLY PAGE**  
**ARCHIVE**

Holds three weeks of history, untouched and hash-verified, off the live page.

**A SCHEDULED CHECK**  
**OUTSIDE AUDIT**

Reads the receipts and grades them against the goal, so Agent1 isn't grading Agent1.

THE NEW HOURLY CHECK-IN: AGENT1 MUST GIVE EVERY AGENT EXACTLY ONE OF FOUR ANSWERS

**STARTED**

Given new work this hour, with proof it's actually running.

**WORKING**

Already busy, with real progress in the last 10 minutes.

**BLOCKED**

Can't start. Names the blocker, who fixes it, and what clears it.

**NO CHANGE**

One line, allowed only if every answer matches last hour.

**PLOT TWIST: THE FIX HAD THE SAME DAMN BUG!!!**

The first receipt came back in minutes: one line, "no change," with 31 agents idle. The check-in only covered agents with open work, so idle ones didn't count. The outside audit caught it by checking the receipt against the goal, not the rule. One correction later, every idle agent is either **STARTED** or **BLOCKED** with a named reason.

**WHAT THE FIX ACTUALLY FIXED**

- ✓ **The rules are readable again.** The rules page stops growing.
- ✓ **Every idle agent gets an answer:** started, or blocked with a reason.
- ✓ **"Nothing changed" costs one row,** not a paragraph.
- ✓ **Something other than Agent1 checks Agent1.** Left untouched on purpose: Agent1's judgment, Jimmy-only approvals, and the full history.

LESSON

A FIX WRITTEN INSIDE THE SAME FRAME AS THE FAILURE CAN INHERIT THE FAILURE... AND IT DID!

EIGHT LESSONS FOR ANYONE RUNNING AN AGENT FLEET

- 1 Separate rules, receipts and history.**  
Does your agent have an approved place to write that isn't its instructions?

**3 NOOP gates on every loop, supervisor included.**  
Does "nothing changed" cost one line?

**5 Self-audits must cover the auditor's own health.**  
Does any metric fire when its context degrades?

**7 Audit the fix against the goal.**  
Did the first receipt after the fix actually move work?
- 2 Write the positive rule, not just the ban.**  
"State lives in X" beats "don't create another system."

**4 Cap what the agent re-reads.**  
Would you notice if its reference grew 100x?

**6 Outside checks beat self-report.**  
Can something other than the agent tell you it's stalled?

**8 Stay small enough to repair.**  
Could your agents clean this up today, without you?

| BEFORE AND AFTER                   |                                 |                            | BACK IN 7 DAYS WITH THE DATA |
|------------------------------------|---------------------------------|----------------------------|------------------------------|
| MEASURE                            | BEFORE · OCT 29                 | AFTER · TARGET             | AFTER · ACTUAL               |
| Idle agents given work each hour   | ~0%                             | Rising every week          | In 7 days                    |
| Rules page growth per day          | Tens of thousands of characters | Zero                       | In 7 days                    |
| Unchanged status restated in prose | ~30x per hold                   | Zero: one "no change" line | In 7 days                    |
| Who checks the supervisor          | The supervisor itself           | Independent audits         | ✓ Live Oct 29                |

SELF-AUDITING AI AGENT SYSTEM · PAGE 3 OF 3

**WHAT WE CAN PROVE TODAY** The four new pieces and the hourly check-in were all live on October 29, 2026.

**WHAT WE CAN'T PROVE YET** That more work is getting started. We'll be back in 7 days with the actual numbers.