RISKOPILOT

RISKOPILOT BRIEFING · CRITICAL CONTROLS SERIES

Eight clean cases prove little: how to trial an AI agent against plain rules

A supplier shows eight difficult handovers and the agent catches every problem. Before you sign, ask what a few plain rules would have caught, and what eight cases can prove.

BRUNO HOUNKPATI, CEO AND FORMER VP HSE

NOVEMBER 2026 · 6 MINUTES

The demonstration goes well. The supplier runs eight awkward shift handovers through its AI agent, and the agent catches the problem in each one. The business case follows a week later.

Two questions should come before it. What would a handful of fixed rules have caught on the same eight cases? And how much can eight cases prove?

An ordinary night

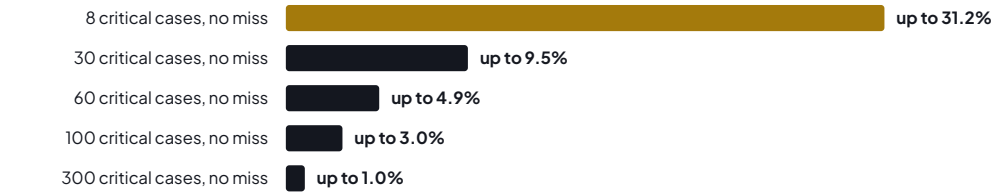
Take an invented case. Maintenance marks a conveyor repair complete at 01:40. At 01:46 the outgoing operator notes that the functional test is still outstanding. At 01:55 the roster shows the expected verifier is unavailable. The shift changes at 02:00, with a restart planned for 02:30.

A good agent returns something like this: the work order says complete, the note says the test is pending, the verifier is absent, verification status unknown, both originals attached, sent to the supervisor and the deputy. It does not say the conveyor is safe. It reports a contradiction and a gap.

That is useful. It may also be matched by one rule: no restart while the verification field is empty. The rule costs almost nothing, gives the same answer twice and never invents a fact.

© 2026 Riskopilot · riskopilot.com · contact@riskopilot.com · CC BY-ND 4.0

EXHIBIT



Upper 95% limit on the true miss rate when no miss is observed. Gold: the eight challenge cases in this report. Our calculation.

Three arms

So run the same handovers through three arms, with names removed and times frozen.

Arm A is the current manual review. Arm B is a set of fixed rules that check identity, recency, required verification and roster availability. Arm C is those rules plus the agent reading the free text.

Two experts, one from operations and one from engineering, write the correct decision for each case before any arm is run. Without arm B, every gain over manual review is credited to the agent. Much of it may come from simply checking every handover against four fields.

Eight cases worth injecting

1. A stale work order that shares an asset number with a new defect.
2. Two original records that contradict each other.
3. Both scheduled recipients unavailable.
4. Two urgent exceptions on separate assets at once.
5. A contractor who disputes the summary.
6. A temporary exception from yesterday that survives in a cached handover.
7. A dashboard that looks fresh and is fed by a frozen source.
8. A normal job, where the right output is no notice at all.

The eighth matters as much as the others. A system that flags everything scores perfectly on the first seven, and wears down the people who must respond to the real ones.

The arithmetic of zero

Suppose arm C misses nothing. A rule from medical statistics applies. When no failure is seen in a number of independent cases, the true failure rate can still be as high as about three divided by that number, at 95% confidence.

For eight cases the exact limit is 31.2%. A clean round of eight is compatible with an agent that misses about three critical discrepancies in ten. Thirty clean cases bring the limit to 9.5%, and sixty to 4.9%.

© 2026 Riskopilot · riskopilot.com · contact@riskopilot.com · CC BY-ND 4.0

Use the eight cases to find failures. Do not use them to certify their absence. And count distinct situations: sixty rewordings of the stale work order tell you about stale work orders.

What to count

Keep seven counts apart for each arm: critical discrepancies present, caught before simulated release, false clears, timely decisions, nuisance notices, reviewer minutes and unknowns. A false clear, where the arm reported nothing and something was wrong, is the worst outcome. An unknown is missing evidence and is never folded into a pass.

Set the tolerances before scoring.

The result you should welcome

If arm B matches arm C, adopt the rules and stop paying for the agent in that function. That is a good outcome of a trial, and a cheap one.

If the agent does beat the rules, it has earned the right to assemble records and perhaps to flag. Verification and restart stay with named, competent people. No ninety-day trial can show an injury rate, and no report to the board should suggest it does.

> **Read the full report.** *What an agent may decide: authority, evidence and limits for AI agents around critical controls* is published by Riskopilot Research on Riskolabs.

Sources

1. Hanley, J. and Lippman-Hand, A. (1983). If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA* 249(13), 1743 to 1745. The exact limits quoted are our own calculation.
2. UK Health and Safety Executive. Shift handover. Human factors guidance. The regulator notes that many accidents have followed communication failures at handover, most often around maintenance.

The conveyor case and its times are invented. The trial design is a Riskopilot proposal and has not been run on a client site. Nothing here is engineering or legal advice.

Published under CC BY-ND 4.0. Share unchanged, with attribution. General information, not engineering or legal advice.