RISKOPILOT

RISKOPILOT BRIEFING · CRITICAL CONTROLS SERIES

The rung you never grant: how much authority to give an AI agent

The debate about agents is held in two words, autonomous or supervised. A site director needs five, and a signature beside each one.

BRUNO HOUNKPATI, CEO AND FORMER VP HSE

NOVEMBER 2026 · 6 MINUTES

A proposal reaches the executive committee. An AI agent will read the shift notes, the work orders and the roster, and tell the incoming supervisor what matters. The supplier calls it supervised autonomy. Someone asks whether that is safe.

The question cannot be answered as put. "Supervised" does not say what the software may touch. A better question is which rung of authority the agent holds, for which function, and who can take it away at two in the morning.

Why an agent cannot hold a duty

A critical control works when a competent person, or an engineered function, acts in time. Three tests separate a duty from a task.

Can the holder be there? A physical state is established by someone who can see the guard or the isolation. A record of the state is a different thing.

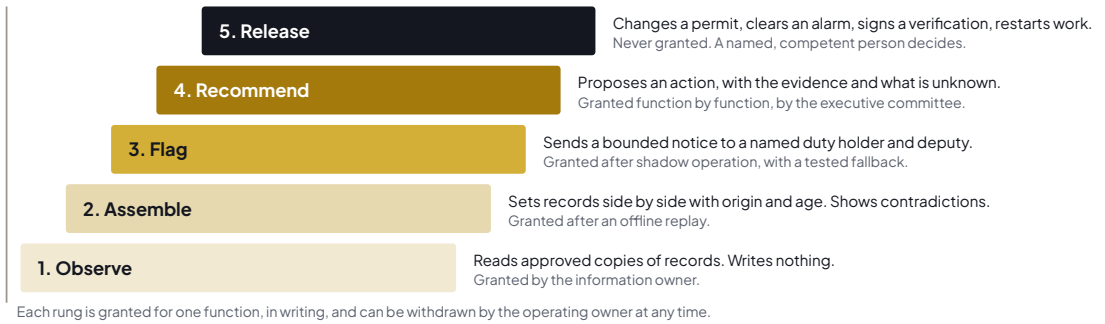
Is the holder competent in the sense the law means? Verification and release belong to people whose training and appointment can be shown to a regulator.

Can the holder answer for the result? After a serious event, an investigator asks who decided.

An agent fails all three. It can still be useful, because a fourth kind of work sits behind every control: processing information. An agent reads everything, every time, and notices when two records about one asset disagree.

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

EXHIBIT



Each rung is granted for one function, in writing, and can be withdrawn by the operating owner at any time.

Five rungs

We propose a ladder. Each rung is granted to one agent for one function, in writing.

1. **Observe.** The agent reads approved copies of records and writes nothing. The owner of each source grants it.
2. **Assemble.** It sets records side by side, with their origin and age, and shows contradictions. Granted after a replay on past handovers.
3. **Flag.** It sends a bounded notice to a named duty holder and a deputy. Granted after shadow operation, with a deadline and a manual fallback the night shift has used.
4. **Recommend.** It proposes an action, with the evidence and what is unknown. Granted by the executive committee, function by function.
5. **Release.** It changes a permit, clears an alarm, signs a verification or restarts work. Never granted.

What the law already says

The EU AI Act requires that a high-risk system can be overseen by people who understand its limits, stay aware of automation bias, can override its output and can stop it. The organisation deploying it must assign that oversight to people with the competence, training and authority to exercise it.

Whether a given agent is legally high-risk is a question for counsel. The four abilities are a sound test for any agent near a control. If the people on shift cannot do those four things, the design is not ready.

The application dates moved in July 2026. The high-risk requirements now apply from 2 December 2027 for the uses listed in the Act's third annex, and from 2 August 2028 for AI that is a safety component of a regulated product. A later date leaves the employer's existing duty to assess the risk where it was.

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

Why rung four is reserved

A recommendation asks a person to agree. Research on that arrangement is old and uncomfortable. A 2010 review found that people supervising automation miss what the aid does not flag and follow what it wrongly advises. Experts were affected as well as novices. Training and instructions did not prevent it.

Telling a supervisor to stay alert is therefore a weak control. The design has to make challenge easy and expected. Count how often people overrule the agent. A rate of zero over many months deserves a question.

Many functions never need rung four. A clear statement of what the records say, and where they disagree, is often enough for a competent person to decide.

The objection worth hearing

"Interlocks already act without a person. A closed top rung is dogma."

They do act, and they are deterministic, validated against a specification and governed by functional-safety standards. No equivalent route exists today for a language-model agent. A harmonised standard for such use would change our view. So would an independent field evaluation. Agents are already deployed in industrial maintenance. We found no independent measure of their effect on verification or release.

What to do on Monday

Ask which agents already read safety records in your operations, and who approved the access. For each one, ask which rung it holds, in writing. Then ask who on the night shift can switch it off, and whether they have ever done so.

> **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. Regulation (EU) 2024/1689 (AI Act), Articles 14 and 26, read on the European Commission's AI Act service desk on 19 November 2026. Paraphrased.
2. Regulation (EU) 2026/1744 of 8 July 2026, Official Journal of 24 July 2026. New application dates for high-risk requirements.
3. Parasuraman, R. and Manzey, D. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors* 52(3), 381 to 410. Abstract read. Evidence from laboratory and aviation tasks, before language-model agents.

The ladder is a Riskopilot proposal and has not been tested against injury outcomes. Nothing here is engineering or legal advice. The classification of an AI system under any law is for the duty holder and its advisers.

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