

A zero override rate is not a quiet system.

Put a human in the loop and one question decides whether the control is real: **is that human changing anything?** Override rate is the share of AI outputs a reviewer changes, rejects or flags before it ships — the rare oversight number that cannot be rubber-stamped. **A triggered diagnostic, never a target.** Give it a target and the thing it detects disappears.

THE READING

WHAT IT MOST LIKELY MEANS — AND WHAT TO DO

VERDICT

Near zero (~0–2%)

Quiet — and quiet reads as health

Rubber-stamping or automation bias — **a fluent wrong answer looks identical to a fluent right one.** An analyst whose escalation rate is zero rarely means a quiet network. **Lever: sample the overrides** — is the model right, or is nobody looking?

INVESTIGATE BEFORE CELEBRATING

Low–moderate (~15–30%)

Reviewers are catching and changing outputs

The feedback loop is plausibly real — **plausibly**, not proven. The cheapest way to hit a number is a trivial edit on every output. **Lever: log the type and severity of each override**, not the count, and track it per reviewer.

PLAUSIBLY REAL

Sustained high (>50%)

The human is doing the work twice

Not an oversight problem — the output is poor enough that review has become production. **Lever: re-scope the deployment, not the review step.** Read it against the capability tier first; a new high-stakes workflow and a mature narrow one differ.

RE-SCOPE, DON'T RE-REVIEW

Zero, with a named owner

Who has never exercised the authority

The ghost owner — no band detects it. The name stops the asking-around: there is an answer to “who decides this?”, so nothing escalates and nothing arrives to be overridden. **Zero, and correct. Lever: one decision the owner exercised last quarter.**

THIS SHOULD STOP A BOARD MEETING

WHAT SUBSTANTIATES THE NUMBER — AND IT IS NOT A BETTER LOG

You compute this from your own logs, about your own reviewers, and report it yourself — the same self-issued evidence a vendor hands you in an assurance pack and expects you to take on trust. **An override is substantiated by the decision it changed, not by the record that it happened.** Every genuine one has a consequence outside the log: an output that shipped differently, a customer who got the other answer, a release that moved. Ask for that.

THREE THINGS TO SETTLE BEFORE THE NUMBER GOES IN THE PACK

01

Is the reviewing seat human? Delegate review to another agent and this metric does not apply to that step — it measures agreement between two systems, not oversight. Govern the envelope instead. Either way, **the row says which seat it is.**

02

Name one decision the column owner exercised last quarter. Not a meeting attended — a hold placed, a budget refused, an escalation raised past the sponsor. If the row is empty, the override rate cannot be read at all, whatever it says.

03

Take ten overrides claimed this quarter. Can each be tied to a change in a system whose owner has no stake in this number? Most of ten and the metric has a substrate. Fewer, and it is self-attesting — **worth saying to a board** before a general counsel says it for you.

A sign-off can be performed in two seconds.
An override cannot be faked into existence.

Source: Fredrik Lindstrom, *Cybersecurity Metrics Toolkit* — AI Oversight Override Rate metric card, the anchor signal of the governance column in *Columns Not Layers* (11 August 2026). The metric and its framing are the author's. **The interpretation bands above are heuristics, not benchmarks.** No benchmark data exists for this metric; they are signal ranges to investigate, and presenting them as certified thresholds would be the overclaim the toolkit exists to correct. **Human-review scope:** Singapore IMDA's *Model AI Governance Framework for Agentic AI* (v1.5, 20 May 2026) requires human approval for irreversible, high-stakes or atypical actions, and states the preference for structural controls over prompt-layer ones. **Oversight effectiveness:** EU AI Act Article 14 requires that oversight can actually be exercised rather than nominally assigned, and Article 26(2) puts the competence, training, authority and support to exercise it on the deployer; the Article 4 AI-literacy duty has been in force since 2 February 2025 — whoever reviews must be competent to disagree, or the rate measures compliance with a rota. **This is not a date-driven metric** and nothing here turns on a compliance deadline. **Maps to:** NIST AI RMF MEASURE and MANAGE; ISO/IEC 42001 human-oversight provisions and Clause 9. Row groupings and verdict labels above are this sheet's reading of the cited source, not the source's own framing.

FOLLOW FREDRIK LINDSTROM FOR MORE

Fredrik Lindstrom

AI and cybersecurity governance for boards, CISOs and senior leaders. 25+ years in enterprise security.

A practitioner to guide you.

ARTICLES & BOARD ADVISORY

fredrikindstrom.info

Long-form governance analysis and toolkits

LINKEDIN

in/fredrikindstrom

Daily commentary and article launches

YOUTUBE

@Promise-and-Risk-AI

AI education without the hype or fear

COMPANION SITE

promiseandrisk.ai

Episode notes, sources and explainers

NEWSLETTER

The Governance Memo

Monthly, board-facing. Subscribe at fredrikindstrom.info