

The Human Intervention Gap

Why AI autonomy needs a measure of human rescue.

Executive thesis. AI teams measure task success, latency, cost, traces and model quality while often missing the human work required to keep supposedly autonomous workflows functioning. Human Intervention Rate (HIR) makes that hidden labor measurable and turns repeated human corrections into evidence for safer automation.

1. The missing variable

An agent can finish a task and still be operationally expensive. A support workflow may achieve a high task-success rate while an operator repeatedly supplies missing context, approves edge cases, edits decisions, redirects the agent, or takes over entirely. Those actions are part of the real runtime. If they are not measured, autonomy can become a narrative rather than an operating property.

2. Human Intervention Rate

$\text{HIR} = \text{human-intervened eligible decisions} / \text{total eligible agent decisions}$. The denominator matters. Not every human checkpoint is waste. Some high-risk, regulated or irreversible decisions should intentionally preserve human control. A useful HIR therefore measures intervention within a defined class of eligible decisions and must be interpreted alongside task success, graceful failure and risk policy.

Intervention type	Operational meaning
Missing context	A person supplied information the agent did not have.
Permission ambiguity	The agent could not determine whether an action was authorized.
Low confidence	A person handled a decision because confidence fell below a threshold.
Policy conflict	Rules or sources disagreed and required judgment.
Tool failure	A dependency prevented completion.
Preference correction	A human changed a result to align with an organizational or user preference.

3. The evidence receipt

Each intervention should preserve the workflow state before intervention, the action the agent proposed, the reason the person stepped in, the human action, the resulting state and whether that correction is reusable. This creates an

evidence receipt that can be audited, analyzed and compared across future runs.

4. From correction to learning

The key product opportunity is not merely counting interruptions. Repeated interventions can reveal a missing rule, missing context source, permission boundary, confidence calibration problem or tool defect. A learning loop should convert repeated patterns into candidate improvements while keeping irreversible policy changes reviewable.

5. A practical autonomy scorecard

A useful scorecard combines HIR with task success, unnecessary escalation rate, human minutes per 100 tasks, repeat-intervention rate, and correction-to-rule conversion. HIR should never be optimized in isolation.

Metric	Purpose
Human Intervention Rate	How often humans enter an otherwise eligible agent decision.
Task Success Rate	Prevents lower HIR from hiding more failures.
Unnecessary Escalation Rate	How often the agent asks for help when policy allows autonomous action.
Human minutes / 100 tasks	Translates intervention into operating burden.
Repeat-intervention rate	Whether the same correction keeps recurring.
Correction-to-rule conversion	Whether intervention becomes durable system improvement.

6. ROI

The economic model is straightforward: intervention volume multiplied by average human handling time and loaded labor cost yields a visible operating cost. Reducing HIR creates value when task success and risk performance remain healthy. This connects technical agent behavior directly to labor economics.

7. Product architecture

A Latchline-style system requires event ingestion, a canonical intervention schema, deterministic metric and policy calculation, semantic classification of causes and corrections, and an evidence layer that supports dashboards, reports and learned-rule proposals. Deterministic code should own permissions, billing, state transitions and exact calculations. AI is better suited to semantic classification, fuzzy pattern recognition, explanations and candidate-rule synthesis.

8. Why the data compounds

At scale, intervention evidence becomes more valuable than a dashboard. A cross-vendor corpus can reveal which task states predict intervention, which corrections generalize, which policies reduce unnecessary escalation and where

organizations consistently require human judgment. The strategic asset is a Human Intervention Graph connecting task state, agent proposal, human correction and subsequent outcome.

9. Responsible target

The goal is not zero humans. The useful objective is to remove unnecessary human labor while preserving deliberate oversight. For ordinary operational workflows, a sub-10% HIR may be a useful ambition where task success remains strong. For high-risk workflows, the correct rate may be materially higher. Measurement should make that boundary explicit.

10. Conclusion

AI autonomy will be credible when organizations can show not only what their agents accomplish, but what those agents still require from people. Human Intervention Rate provides a simple starting point. The deeper opportunity is to treat every human rescue as evidence that can improve the next run.

Working thesis. This document describes a proposed operating metric and product architecture. Benchmark values should be validated with production datasets before being treated as industry norms.