10

WHITE PAPER · AI OPERATING GOVERNANCE

AI Autonomy Should Be Earned Through Control Evidence

Action authority expands only when evidence earns it

Marco Policani

Enterprise Portfolio · PMO · AI Operating Governance

policani.net · linkedin.com/in/marcpolicani

July 2026

EXECUTIVE SUMMARY

Autonomy grows when control evidence supports it; it is not a default permission tier.

Capability is what a system can do; autonomy is what the organization permits it to do, with an owner, a boundary, and a recovery path. This paper grants that permission through five explicit levels — recommend, prepare, act reversibly, act with approval, act under policy — earned on control evidence, per workflow and action class, with de-escalation designed as an ordinary act.

01 Inventory actions by consequence and reversibility first; the map makes the appropriate levels nearly obvious.

02 Advance only on evidence that covers ordinary variation — a good fortnight is the ladder's version of the demo problem.

03 Watch the ratio of downgrades to incidents: stepping down on triggers before harm is what governing looks like.

The argument

Autonomy is permission to affect work, systems, money, customers, or records — not a property a model possesses. Authority should expand through explicit levels only when boundary adherence, exception handling, recovery, and accountable intervention have produced sufficient evidence.

The proposal that sharpened my thinking on this arrived as a slide titled 'Fully Autonomous Invoice Agent.' The demonstration was persuasive: the agent read invoices, matched purchase orders, resolved discrepancies, and released payments, end to end. The sponsoring team wanted production access to the payment system. When I asked what the agent was permitted to do on its worst day — not its average day — the room went quiet. Nobody had designed the worst day. The capability had an impressive ceiling and no floor: no transaction limit, no defined intervention point, no rollback owner, no evidence about behavior on the malformed invoices that made up a tenth of the real stream. The team had built autonomy as a feature. It needed to be granted as an authority.

That distinction organizes everything in this paper. Capability is what the system can do; autonomy is what the organization permits it to do, under ordinary variation, with a defined consequence, an accountable owner, and a working recovery path. Capability is demonstrated. Autonomy is earned — level by level, on control evidence, with a route back down. The NIST AI Risk Management Framework and its generative-AI profile supply the risk-management foundation [1][2]; the five-level ladder is how I turned that foundation into something a workflow owner can actually grant, hold, and revoke. It is written for AI program leaders, workflow owners, risk executives, and technology leaders.

Why autonomy defaults are broken

Organizations currently assign AI action authority by two defaults, both wrong. The enthusiasm default grants whatever the demonstration suggested — the agent handled the demo invoices, so it gets the payment system. The fear default denies everything regardless of evidence — nothing acts without a human click, forever, even where the click adds no judgment and the queue it creates becomes its own risk. Both defaults share the same defect: they assign authority based on sentiment about the technology rather than evidence about the workflow.

The middle path requires a vocabulary the sentiment debate lacks. 'Autonomous' is not one thing: an agent that drafts a reply, one that files it, one that commits a payment under a limit, and one that executes a policy across thousands of cases hold radically different consequences, and lumping them under one word makes every discussion a referendum on trust in AI generally. Levels replace the referendum with a specific question: what is the next increment of authority this workflow is asking for, and what evidence supports it?

Levels also change the burden of proof in a useful direction. Under the enthusiasm default, skeptics must prove danger; under the fear default, sponsors must prove perfection. Under an evidence ladder, the system's operating record makes the argument: boundary adherence at the current level is the case for the next one, and the absence of that record is the end of the conversation. The debate moves from opinions about models to inspection of evidence, which is ground governance can actually hold.

The five-level authority ladder

The ladder defines five authority levels, granted per workflow and per action class — never as a blanket status for a model or platform. Each level has a distinct failure mode, a control design, and an evidence standard for advancing. The sixth element, de-escalation, is what makes the ladder an operating instrument rather than a promotion track.

EXHIBIT 1

Five authority levels, one boundary of honest reversibility

1. **Level 1 — Recommend**
   The system proposes an answer or action; a person evaluates it before any downstream commitment.
2. **Level 2 — Prepare**
   The system may assemble a transaction, communication, configuration, or record but cannot release it.
3. **Level 3 — Act inside a reversible boundary**
   The system may execute low-consequence actions that can be detected and undone within a known period.
4. **Level 4 — Act with approval**
   The system may orchestrate consequential work after a qualified owner approves the case and conditions.
5. **Level 5 — Act under preauthorized policy**
   The system may execute within a narrow policy envelope without per-case approval, while monitoring and stop authority remain active.

6 **De-escalation is part of autonomy**
Authority must contract when evidence weakens, context changes, or oversight capacity falls.

Level 1 — Recommend

At the first level the system proposes; a person evaluates before any downstream commitment. This is where most workflows should start and where some should stay. Its quiet failure is that recommendation drifts into de facto action: review is rushed, the interface makes acceptance one click and challenge five, and the human 'decision' becomes a formality performed at volume. A recommendation accepted unread is an action with extra steps.

Level 1 is also where oversight capacity gets its first honest measurement, and the number matters for everything above. If reviewers can evaluate forty recommendations a day well, that is the workflow's oversight budget — and every later level's volume assumptions must fit inside it or fund its expansion. Programs that scale generation tenfold while holding review capacity flat have decided, implicitly, to stop reviewing; the ladder makes that decision explicit before it makes itself.

The controls keep suggestion and execution genuinely separate: the recommendation shows its sources and its uncertainty, acceptance of consequential items is recorded as a decision, and sampled recommendations get independent checks to calibrate how often the reviewer's approval was actually deserved. The advancement evidence is about the pair, not the model: recommendation quality, reviewer correction patterns, and proof that reviewers reliably catch the material errors. A workflow whose reviewers cannot demonstrate detection has no business granting the next level — the safety net it is about to rely on does not exist.

Level 2 — Prepare

At the second level the system assembles the artifact — a transaction, a communication, a configuration change, a record — but cannot release it. Preparation captures most of the efficiency while a person holds the trigger, which makes it the workhorse level for consequential work. What breaks it is silent boundary crossing: auto-send settings, bulk-approval interfaces, and ambiguous flows let prepared work slip into the world without the affirmative human act the level promises.

Level 2's economics explain its durability: preparation typically captures the large majority of the cycle-time saving, because assembling the artifact was the slow part and releasing it takes seconds. Workflows that measure the marginal value of moving from level 2 to level 3 often find it smaller than assumed — which is useful knowledge in both directions, either justifying a long stay at preparation or clarifying that the case for autonomy rests on off-hours coverage and latency rather than on labor.

The control is an unmistakable release step: an authorized person sees what will change — fields, systems, amounts, recipients — and affirmatively releases it, with identity recorded. Bulk approval defeats the level's purpose exactly to the degree that it makes release unexamined; where volume demands batching, the batch needs sampling and the sampling needs teeth. Advancement evidence: preparation accuracy over a meaningful period, edits made before release (each one is a caught defect), rejected preparations and why, and demonstration that review capacity remains credible at the volume the next level would create.

Level 3 — Act inside a reversible boundary

At the third level the system executes without per-case approval, inside a boundary defined by reversibility: low-consequence actions, detectable quickly, undoable within a known window. The boundary's honesty is everything, and the trap is technical reversibility standing in for the real kind. A database write can be rolled back; the customer email announcing it cannot; a reversed payment that took eleven days to reverse was not, in any operating sense, reversible.

The controls are boundary mechanics: transaction limits, eligible-case definitions, an observation window matched to detection speed, a rollback owner with tested procedures, and a cap on aggregate exposure inside any window. The evidence is operational: boundary adherence rates, intervention frequency, rollback exercises actually run — with the clock measured — and the loss or harm avoided when intervention occurred. The decision rule is asymmetric by design: reduce authority immediately when practical reversibility proves worse than designed; restore it only when the evidence explains the gap.

A worked boundary makes level 3 concrete. A collections workflow grants the system authority to send payment reminders and apply standard late fees — but only for accounts under a balance threshold, only where the account history is clean, only twenty actions per hour, with a four-hour observation window and a rollback owner in the billing team. Everything else routes to preparation. The boundary took a morning to design, and every clause came from the action inventory: the threshold from consequence analysis, the rate cap from detection speed, the exclusions from the cases that had burned humans before. Six weeks of adherence data later, the threshold rose — an evidence-based expansion, recorded in a sentence.

Aggregate exposure is the clause teams forget. A thousand individually reversible actions in one bad hour is an irreversible reputational event, whatever the per-case rollback says. Boundary design therefore caps the window, not only the transaction: how much total consequence can accumulate before the observation cycle catches a systematic error? The cap converts a class of correlated failure — the model update that shifts behavior across every case at once — from a catastrophe into a contained batch.

Level 4 — Act with approval

At the fourth level the system orchestrates consequential work — multi-step, multi-system, real money or customer impact — after a qualified owner approves the case and its conditions. The level exists for work too consequential for boundary autonomy but too complex for per-field human preparation. It decays through ceremonial approval: volume rises, the approval packet bloats or thins to uselessness, and the approver becomes a signature loop clearing forty cases before lunch — formally accountable, and unable to see what they are accountable for.

The controls treat the approver as part of the system design: an approval packet engineered to be judged (what will happen, what evidence supports it, what is unusual about this case), volume caps tied to honest review time, separation of duties where the requester benefits, and an escalation route for the cases that deserve a slower look. Advancement evidence centers on approval quality: latency, override and rejection reasons, reviewer workload against the cap, and — the strongest signal — actions stopped before execution. An approval stream that never stops anything is either supervising a perfect system or performing supervision; the record should be able to say which.

Level 5 — Act under preauthorized policy

At the fifth level the system executes within a narrow policy envelope without per-case approval — monitoring, thresholds, and stop authority remain, but no human touches routine cases. This is the level marketing language sells and few workflows need; where it fits, it fits because the envelope is genuinely narrow, the case population is well understood, and the evidence base is deep. Its failure mode is the goal standing in for the boundary: 'resolve customer issues efficiently' is an objective, not an envelope, and rare cases outside the training distribution will be handled with the same confidence as the routine ones.

The control is an envelope encoded as policy, not aspiration: eligible actions enumerated, limits and exclusions explicit, conformance logged per action, incident thresholds wired to automatic safe states, and drift monitoring against the population the evidence covers. The grant is per workflow and per action class, on production evidence from level 4, recovery exercises, and demonstrated owner response. Blanket level-5 status for a model — any model — is the category error the ladder exists to prevent: the model did not earn authority; a workflow, with its controls and its owners, earned it.

The approver's experience is a design surface, not an afterthought. Level-4 supervision fails through the same human-factors channels every inspection discipline knows: habituation to endless clean cases, deference to a system with a good record, fatigue at volume. The countermeasures are borrowed from those disciplines — rotation, deliberate injection of known-hard cases to keep detection calibrated, and dashboards that show the approver their own override history so drift toward rubber-stamping is visible to the person drifting. An organization that measures its models' error rates and never its approvers' detection rates has instrumented half the control.

De-escalation is part of autonomy

The ladder's most important rungs point downward. Evidence ages: models update, data drifts, exception rates move, the experienced reviewer leaves, the workflow's consequences change with a new product or regulation. Authority that only ratchets upward converts every one of those ordinary events into accumulating unexamined risk, because the original grant no longer describes the system it governs — and in most organizations the only downward route is a major incident, which is the most expensive trigger available.

The control is predefined downgrade triggers with named owners: model or vendor changes above a threshold, data shifts, exception or override rates crossing bounds, control-test failures, oversight-staffing changes, new consequence classes. Stepping down is designed as an ordinary operating act — communicated, temporary by default, with restoration criteria stated at the moment of downgrade. The cultural half matters as much as the mechanics: a downgrade treated as a team's failure will be resisted, gamed, and delayed; a downgrade treated as the control system breathing keeps the ladder honest. The evidence of health is unglamorous: triggers that fired, permissions that contracted within their windows, fallbacks that carried the load, and restorations that followed evidence rather than pressure.

Tool chains and multi-agent designs do not change the model; they multiply the entries in it. An orchestrating agent that calls a research tool, a drafting tool, and a payment tool holds three different authorities, and the ladder applies per tool-action, not per agent: research at level 5, drafting at level 2, payment at level 3 with a limit, in the same workflow, is a perfectly coherent — and typical — grant profile. What the orchestration adds is a composition question: the chain's effective authority is whatever its most consequential reachable action permits, and the grant review should trace reachability rather than trusting the agent's intended path.

The ladder in operation

One clarification the levels make cheap: 'human in the loop' stops being a slogan and becomes an address. Levels 1 and 2 put the human before every consequence; level 3 puts them after, inside the observation window; level 4 puts them before the consequential subset; level 5 moves them to the monitoring layer. Each is a legitimate configuration for some work and an irresponsible one for other work — and the phrase alone, without a level, no longer counts as an answer to any governance question.

EXHIBIT 2

What each level requires before it is granted

Level 1 — Recommend

Remain at recommendation when reviewers cannot reliably detect material errors.

Level 2 — Prepare

Expand only when preparation is accurate and review capacity remains credible at expected volume.

Level 3 — Act inside a reversible boundary

Reduce authority immediately when practical reversibility differs from the design assumption.

Level 4 — Act with approval

Hold volume below the point where oversight quality deteriorates.

Level 5 — Act under preauthorized policy

Grant this level by workflow and action class, never as a blanket status for the model.

De-escalation is part of autonomy

Step down first and investigate from a safe boundary rather than preserving status while evidence is uncertain.

Granting works best as a small ceremony with a real record: the workflow owner requests a level for an action class, the evidence is inspected against the standard, the risk owner signs, and the grant states its boundary, its monitoring, its downgrade triggers, and its review date. The record matters because authority without a record reverts to folklore — six months later nobody can say what was permitted, on what basis, or by whom, and the audit reconstructs it from configuration.

For organizations already using the NIST framework, the ladder slots into its verbs cleanly: the action inventory and consequence analysis are mapping work, the advancement evidence is measurement, the grants and triggers are management, and the ceremony — signed records, named owners, review dates — is governance [1]. The generative-AI profile's attention to human-AI configuration and pre-deployment testing lands, in ladder terms, at the level-1 and level-2 evidence standards [2]. Nothing here competes with the framework; the ladder is one way to make its guidance grant-shaped.

Expect workflows to spread across levels, and read the spread as information. A portfolio where everything sits at level 1 is paying for capability it does not trust — usually an evaluation gap, not a model gap. A portfolio with early level-5 grants and thin level-3 evidence has been promoted by enthusiasm. The healthy pattern is a pyramid:

broad recommendation and preparation, selective reversible action, rare and well-instrumented approval orchestration, and preauthorized policy only where the case population is boring — boring being, in autonomy, the highest compliment a workflow can pay.

The ladder also disciplines vendor conversations. 'Agentic' platforms arrive claiming the top of the ladder; the procurement question becomes which level the platform's controls can actually evidence — what it logs, what it can stop, what it can undo, and what intervention surface it offers the operating owner. Products built for level 5 marketing and level 2 controls reveal themselves quickly under those questions, which is the cheapest possible time for the revelation.

Granting the first authorities

Installation follows the same evidence discipline the ladder imposes:

- Inventory actions by consequence and reversibility before assigning authority levels.
- Set evidence gates for one workflow and one action class at a time.
- Test intervention, rollback, and permission reduction under realistic load.
- Review authority after changes to model, data, tools, policy, users, or oversight staffing.

The inventory is the unglamorous prerequisite that makes everything else possible. Most workflows have never enumerated their actions by consequence and reversibility, and the first pass reliably surprises: actions assumed trivial turn out to touch the customer record; actions feared turn out to be cleanly undoable. The inventory converts autonomy from an argument about the system into a map of the work — and the map, once drawn, makes the appropriate levels close to obvious.

Time-in-level deserves a norm, though not a rigid one: long enough for the evidence to cover ordinary variation — quarter-end, staff rotation, at least one upstream change — rather than a lucky fortnight. Advancing on a good week is the ladder's version of the demo problem, and the evidence standards should name the variation they expect to have seen.

Start the first grant where the evidence is easiest to generate: high-volume, low-consequence, well-logged. The point of the first cycle is not the automation value; it is exercising the machinery — the grant record, the monitoring, the downgrade trigger, the restoration — so that the machinery is boring by the time a consequential workflow needs it.

What the record should show

Autonomy governance leaves a specific evidence trail:

- Grants on record: workflow, action class, level, boundary, owner, review date.
- Advancement decisions citing the evidence standard they met.
- Interventions, rollbacks, and stops exercised — with clocks — not merely designed.
- Downgrade triggers fired and permissions contracted within their windows.
- Restorations following recorded evidence rather than sponsor pressure.

- Incidents at each level against the volume governed there.

The grant record itself deserves a standard one-page shape: the workflow and action class; the level and its boundary clauses; the evidence cited, with dates; the monitoring in place and where its output lands; the downgrade triggers and their owners; the approver of the grant; and the review date. Configuration should be derivable from the record and periodically reconciled against it — the reconciliation is a cheap audit that catches both silent permission creep and controls that drifted out from under their paperwork.

The most diagnostic single number is the ratio of downgrades to incidents. A system that steps down on triggers, before harm, is governing; a system whose only downgrades follow incidents is reacting; a system with neither downgrades nor incidents at scale has thresholds nobody is watching. The record should make the three cases distinguishable at a glance.

The strongest counterargument

The ladder does not claim autonomy is always desirable or that every workflow should climb. Some work should remain at recommendation or preparation indefinitely. The best level is the smallest authority that produces the required value within an operating boundary the organization can sustain.

A related objection concerns overhead at small scale, and it deserves the honest answer: a two-person team automating its own reporting does not need grant ceremonies. The ladder's machinery prices in where consequence does — customer impact, money movement, regulated records, actions other teams rely on. The scaling rule is the same proportionality that governs everything else in this space: the ceremony should be as small as the consequence allows, and no smaller than the recovery requires.

The velocity objection — that laddered authority concedes the field to competitors who grant autonomy faster — misreads where the speed comes from. The slow part of enterprise AI adoption is not model capability; it is trust collapse after the first ungoverned incident, when everything gets frozen while a committee reconstructs what the system was allowed to do. The ladder is how organizations go fast durably: each grant is small, evidenced, and reversible, so no single failure indicts the program. The competitor who granted everything at once is one bad Tuesday from granting nothing at all.

The shift

The shift is from autonomy as a property of the technology to authority as a decision of the organization: granted per workflow and per action class, earned on control evidence, held inside boundaries that are honest about reversibility, and revocable as an ordinary operating act rather than an emergency. Capability keeps rising; that part is not in the organization's control and never will be. What is in its control is the ladder the capability climbs — and the discipline that makes each rung load-bearing before anyone stands on it.

The invoice agent went to production — at level 2, then level 3 with a transaction limit, and eleven months later at level 4 for the clean-match majority of its stream. Through the period I stayed close to it, its worst days were days its boundaries had been designed for, which is all a governance design can promise and exactly what the slide-deck version could not.

Notes and sources

[1] NIST, "AI Risk Management Framework Core," 2023. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

[2] NIST, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," July 2024. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.600-1.pdf

About the author

Marco Policani is an enterprise portfolio, PMO, and AI operating-governance leader. He builds portfolio governance systems where AI carries the analytical load and named humans own every decision — an approach documented across case studies, walkthroughs, and working governance guides on his portfolio site.

Portfolio & governance library: policani.net/governance · Full portfolio: policani.net · LinkedIn: linkedin.com/in/marcpolicani