The Autonomy Ladder™: Earning AI Authority One Verifiable Rung at a Time

A governance framework for autonomous AI in regulated operations — A0 through A4, and every rung demotable.

Kunjar Bhaduri · Author of the Autonomy Ladder™ · Financial-Services Technology Executive · Dallas, TX

Executive Summary

In any high-stakes autonomous system — a self-driving handoff, a robot on a factory floor, a swarm of software agents coordinating a trade or a claim — the failure that hurts is rarely that the model could not do the task. It is authority granted before it was earned: a system promoted to act on its own with no record of who decided it was ready, what evidence justified the decision, or what would pull it back. Capability asks what a model *can* do; this paper is about what a deployed system is *allowed* to do, and on whose say-so.

That failure is concrete. An adverse-action notice that says "debt-to-income above threshold," drafted by an agent operating outside its envelope and approved by a human who clicked *Approve* without reading — it does not announce itself with a crash; it surfaces in a CFPB exam eighteen months later, when no one can name who granted the authority or what would have revoked it. The same shape appears wherever autonomy outruns evidence — a coordination cascade no single agent owns, a scope that crept up through config changes none logged as a decision. Regulated finance is simply where the bill arrives first, and in writing.

I have spent twenty-five years building and running technology in financial services. In one chapter I architected a **$750M** multi-year private-capital platform deal. In a separate chapter, at a different firm, I rebuilt a loss-making division from **$12M to $40M with zero layoffs** — and carried it through a **twelve-day hard-down ransomware attack** with no disaster recovery available, shipping the MVP on a fresh cloud in fifty days and earning **SOC 2 Type 2 and ISO/IEC 27001** in the same window. Across all of it, one pattern holds: trust in an automated system is earned against evidence, not asserted in a deployment ticket.

The **Autonomy Ladder™** is the correction. Its central claim is a reordering of the deployment sequence most organizations have backwards: *prove the evidence, then grant the rung, and keep a control you cannot switch off.* It names five explicit autonomy levels — **A0 through A4** — each with its own evidence bar for promotion, its own inspectable control surface, and the standing ability to be demoted.

Level	Name	What the agent does	What the human controls
A0	Informational	Reads and recommends; no write authority.	Every write — the agent only proposes.
A1	Assisted	Reads and drafts the artifact.	Approves every write before it reaches the system of record.
A2	Delegated	Writes low-risk decisions inside a hard envelope.	Approves a sampled subset, plus all out-of-envelope decisions.
A3	Supervised Autonomous	Writes autonomously for an in-scope decision class.	Holds a non-overridable sovereign veto and a live audit ledger; supervises by exception.

Level	Name	What the agent does	What the human controls
A4	Production Autonomous	Writes autonomously across multiple coordinating agents.	Sovereign veto and live ledger, plus monitor-led promotion and operator-validated escalation paths.

You climb one rung at a time, never skipping, and only when the evidence for the next rung is in hand. The framework was not theorized: it was extracted from two places where ungoverned autonomy has a price — a private quantitative program I run almost entirely on frontier models, and a career spent building and governing production technology in regulated financial services. Six open-source reference libraries — `finserv-agent-audit`, `banking-agent-audit`, `payments-agent-audit`, `payer-agent-audit`, `private-capital-agent-audit`, and `cre-agent-audit` — translate the framework into vertical-specific governance patterns; each is DOI-archived and openly licensed. And the ladder maps onto the obligations institutions already carry — SOC 2 Type 2, ISO/IEC 27001, FINRA supervision, federal banking model-risk guidance, the NIST AI Risk Management Framework, and the EU AI Act's human-oversight requirement — detailed in Section 4, which also situates the ladder against the frontier labs' own capability frameworks.

For a regulated institution, the ladder is a control its examiners can actually test; for a frontier lab or cloud platform deploying agents into one, it is the governance vocabulary its field teams can hand a customer's model-risk function rather than improvise. With the EU AI Act's high-risk obligations and the first US state AI laws phasing in through 2026 and 2027, the interval to get this right is now measured in quarters.

The Autonomy Ladder in one page

Autonomous systems fail less from what a model *can* do than from authority it was never shown it had earned. The Autonomy Ladder fixes the order. Five rungs, A0→A4: read-only; draft-for-approval; write inside a hard envelope; autonomous within one decision class under a non-overridable sovereign veto; multi-agent production. You climb one rung at a time, only on evidence, and never skip — skipping is the invisible-promotion failure mode wearing a roadmap. Every rung is *demotable*: drop a level the moment assurance degrades, as a routine move, not a crisis. Three controls make it a control and not a slide — a sovereign veto the agent cannot switch off, a tamper-evident audit ledger that makes every action inspectable in real time, and a mechanical demotion trigger. It was extracted from running multi-agent systems under real consequence, and it ships as six open, DOI-archived reference libraries whose test suites turn each rule into a falsifiable, runnable control. The question it leaves you with: *where does your highest-authority autonomous system sit today — and who promoted it there, against what evidence?*

---

1. The Problem: Autonomy Is Being Granted Before It Is Earned

The question facing every regulated enterprise is no longer whether to deploy AI agents. It is how much authority to give them, on what schedule, and with what proof at each step. That question is being answered badly. An agent that shipped as a chatbot is writing to a system of record a quarter later — with no record of who decided it had earned that write authority, what evidence justified the decision, or what would trigger pulling it back. The industry has the inverse instinct with AI agents — grant broad authority, then bolt on controls when something breaks. That order is backwards, and it is most dangerous exactly where the consequences are largest: credit decisions, payments finality, suitability, model risk, fair-lending exposure.

This is not a hypothesis. In February 2026, Anthropic published *Measuring AI Agent Autonomy in Practice*, an early large-scale empirical measurement of how much autonomy people actually grant agents in production, drawn from millions of interactions across its coding tool and API. The finding that matters here: newer users let an agent run on full auto-approve about 20% of the time, but by their 750th session that figure climbs to over 40% — authority accreting with familiarity rather than with evidence (Anthropic, 2026; see References). That is the default trajectory this framework exists to interrupt. And the reader who most needs it is not only the regulated institution: it is equally the frontier lab, cloud platform, or agentic-product team deploying *into* those institutions, who needs a governance vocabulary their account teams and forward-deployed engineers can hand a customer rather than invent.

Generic AI-governance advice names principles. Regulated operations is where those principles meet an examiner who tests the *specific decision class* — fair-lending adverse action, payments finality, suitability. That collision is the whole problem, and it is not the same problem as "responsible AI" in the abstract. A policy binder names the principle; the examiner tests the decision class. The Autonomy Ladder is built for that collision.

Three failure modes recur. The first is the **binary trap**: a system is treated as either "human-in-the-loop" or "fully autonomous," with nothing in between, so teams jump from a human approving everything to an agent approving everything because the intermediate states were never named. The second is the **invisible promotion**: an agent's authority creeps upward through configuration changes and scope expansions that no one logged as a decision, so by the time a regulator asks "who approved this level of autonomy and on what basis," there is no answer. The third is the **irrevocable deployment**: autonomy granted with no built-in mechanism to claw it back when conditions change, no live audit ledger, and no escalation path a human operator has actually validated.

What is missing is a **ladder** — a small set of explicitly named autonomy levels, each with its own evidence bar for promotion, each with a control surface a regulator can inspect, and each demotable. The Autonomy Ladder is that ladder. It defines five levels, A0 through A4, and the rule that you climb them one rung at a time, never skipping, and only when the evidence for the next rung is in hand.

This is not a maturity model that flatters everyone toward "level 5." Most production agents in regulated contexts should live at A1 or A2 for a long time, and that is the correct, defensible answer. The framework's value is that it makes the level a deliberate, documented, reversible decision rather than an accident of configuration drift.

2. The Framework: A0 → A4

The Autonomy Ladder defines five levels of agent authority, summarized in the table above and defined rung by rung below. The defining variable at each level is the answer to two coupled questions: what may the agent *write*, and what does a human still *control*. Read in order, the levels describe a steadily narrowing scope of human intervention matched by a steadily rising bar of evidence and oversight machinery.

A4 PRODUCTION AUTONOMOUS	**Writes** autonomously across multiple coordinating agents. **Human holds** sovereign veto, live ledger, monitor-led promotion, operator-validated escalation. **Earns it** once A3's machinery has run long enough to trust across agents. **Demote if** an escalation path fails under realistic failure.
	Writes autonomously for one in-scope decision class. **Human holds** a non-overridable sovereign veto + live audit ledger; supervises by exception. **Earns it** once the A2 envelope holds and sampling finds nothing alarming. **Demote if** the veto or ledger flags drift.

A3 SUPERVISED AUTONOMOUS	
A2 DELEGATED	**Writes** low-risk decisions inside a hard, mechanically enforced envelope. **Human holds** approval of a sampled subset + all out-of-envelope cases. **Earns it** by showing a bounded class of writes is safe to delegate. **Demote if** sampled review finds drift.
A1 ASSISTED	**Writes** nothing on its own — drafts the artifact. **Human holds** approval of every write before the system of record. **Earns it** by demonstrating recommendation quality at A0. **Demote if** approvals become rubber stamps.
A0 INFORMATIONAL	**Writes** nothing — reads, ranks, recommends. **Human holds** every write; the agent only proposes. **Starting rung** for any new agent in a regulated process. **Demote** — n/a (the floor).

Figure 1 — The Autonomy Ladder. Authority and the evidence bar rise together, rung by rung; every rung is demotable.

A0 — Informational

At A0 the agent reads and recommends; it has no write authority. It can surface, summarize, rank, and propose, but every action that changes a system of record is taken by a human who may or may not follow the recommendation. A0 is the right starting rung for any new agent in a regulated process, because it lets the organization observe the agent's judgment against real cases while carrying zero direct decision risk. The governance question at A0 is narrow: is the recommendation quality good enough, and is its reasoning legible enough, to consider any write authority at all?

A1 — Assisted

At A1 the agent reads and drafts; a human approves every write. The agent now composes the actual artifact — the adverse-action notice, the journal entry, the customer message, the disposition — but nothing reaches the system of record until a human reviews and approves that specific write. A1 is where most regulated agents should sit, often for a long time, because it captures the bulk of the productivity gain (the human stops drafting from scratch) while preserving a human decision on every single write. The governance question at A1 is whether the human reviews are real reviews or rubber stamps; an A1 deployment whose approvers click through without reading is functionally operating above its rung without the evidence to justify it.

A2 — Delegated

At A2 the agent reads and writes low-risk decisions inside a hard envelope; a human approves a sampled subset plus all out-of-envelope decisions. This is the first rung where the agent writes to the system of record on its own — but only within a tightly bounded class of low-risk decisions, and only inside an envelope (thresholds, decision types, value limits, eligible populations) that is defined in advance and enforced mechanically. Anything outside the envelope routes to a human, and a sampled fraction of the in-envelope writes is reviewed after the fact to keep the quality estimate honest. The governance question at A2 is whether the envelope is genuinely hard — enforced by the system, not by the agent's own self-restraint — and whether the sampling rate is high enough to catch drift before it compounds.

A3 — Supervised Autonomous

At A3 the agent reads and writes autonomously for an in-scope decision class; a sovereign-veto layer is non-overridable; a live audit ledger records every action; and a human supervises by exception. This is a categorical step up from A2: the agent now owns an entire decision class end to end, rather than a low-risk

slice of one. What makes A3 defensible rather than reckless is the machinery around it. The sovereign-veto layer is a separate, non-overridable control that can block any action regardless of what the agent decides — and critically, the agent cannot turn it off. The live audit ledger makes every action inspectable in real time, not reconstructable after an incident. And the human is no longer in the loop on each decision but supervises by exception, intervening when the ledger or the veto layer flags something. The governance question at A3 is whether the veto layer is truly non-overridable and truly independent of the agent it governs.

A4 — Production Autonomous

At A4 the agent operates as A3 plus inter-agent orchestration, monitor-led promotion, and operator-validated escalation paths. A4 is not "A3 with the guardrails removed" — it is A3 extended to a system of multiple agents working together, where the additional risk is coordination risk. Inter-agent orchestration means agents hand work to each other; monitor-led promotion means a dedicated monitoring agent or layer governs when and whether a given agent's scope expands, rather than a human flipping a configuration flag; and operator-validated escalation paths mean every route by which a problem escalates to a human has been tested by a real operator and confirmed to work, not merely diagrammed. The governance question at A4 is whether the escalation paths have been *validated under realistic failure*, because the failure mode of a multi-agent system is rarely a single bad write — it is a cascade no single agent owns.

The Climbing Rule

The levels are a ladder, not a menu. You climb one rung at a time and only when the evidence for the next rung is in hand: A0 earns A1 by demonstrating recommendation quality; A1 earns A2 by demonstrating that a bounded class of writes is safe enough to delegate inside a hard envelope; A2 earns A3 by demonstrating that the envelope holds and the sampling finds nothing alarming; A3 earns A4 only after the sovereign-veto and audit-ledger machinery has run long enough to trust extending it across coordinating agents. Skipping a rung — shipping an agent straight to A3 because the demo was impressive — is the invisible-promotion failure mode wearing a roadmap. Equally important, every rung is *demotable*: if the evidence degrades, you drop a level, and the framework treats that as a normal operating action rather than an admission of failure.

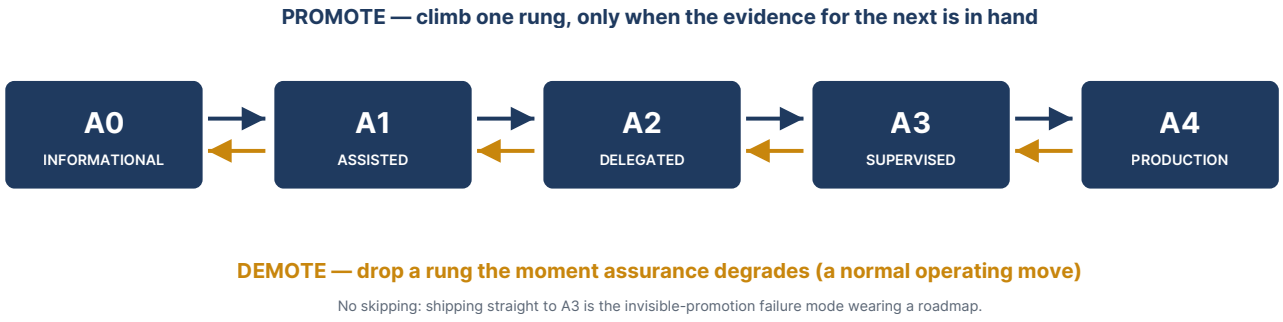

Figure 2 — The climbing rule as a state machine: promotion is earned one rung at a time; demotion is routine, not failure.

A3 and A4 Are Also Infrastructure

The ladder describes architectural commitments, but A3 and A4 are operational commitments too. The sovereign-veto layer is a running process with a latency budget, an availability target, and a separation-of-duty boundary — in practice a separate cloud account or subscription, a distinct IAM principal, and a network path the agent's own principal has no route to reach. The live audit ledger is a write-once store with its own retention, access-control, and incident-response posture; the mistakes catalog is a system of

record. Whoever owns an A3 or A4 deployment owns those operational properties — who provisions them, who pays the latency budget, how they survive a regional failover — not just the design. A control that exists only on an architecture diagram is not yet a control at all; making it real is the difference between governance that survives an examination and governance that does not.

3. The Evidence: Discipline Before Autonomy

The Autonomy Ladder is not a thought experiment. It was distilled from the discipline of building and running autonomous systems where the consequences of an ungoverned action are real. The evidence comes in two forms, presented in order of how readily a skeptic can check it: first the public, inspectable libraries anyone can read today, then the private program from which the discipline was learned.

The most directly verifiable evidence is public. I maintain six public, openly licensed agent-audit libraries — all under MIT, four of them also offered under Apache-2.0 (dual MIT OR Apache-2.0), with each library's LICENSE file definitive — reference IP, not deployed controls — that translate the ladder into concrete, inspectable governance patterns for specific regulated verticals, each DOI-archived for citation:

Library (`github.com/linus10x/…`)	Vertical	License	What it encodes
`finserv-agent-audit`	Cross-vertical financial services	MIT OR Apache-2.0	Lending and core FS decision-class patterns.
`banking-agent-audit`	Banking	MIT OR Apache-2.0	Model-risk, ECOA/Reg B adverse action, BSA/AML/OFAC.
`payments-agent-audit`	Payments	MIT	OFAC screening, BSA/AML, Reg E, rail-finality/irreversibility.
`payer-agent-audit`	Health-insurance payer	MIT OR Apache-2.0	Coverage and utilization-decision governance patterns.
`private-capital-agent-audit`	Investment advisers	MIT OR Apache-2.0	Suitability and the Advisers Act duty of care.
`cre-agent-audit`	Commercial real estate	MIT	Screening and fair-housing-adjacent decision patterns.

Each library is more than a specification: it ships as an installable Python package with a runnable test suite, worked examples, and decision-class templates for the regulator-mapped artifacts its vertical turns on — ECOA reason codes, OFAC screening events, suitability assessments. The README is the entry point; the test suite is the falsifiability proof.

They are published openly so the framework can be examined, criticized, and adopted rather than taken on faith. They are reference implementations of the patterns described here; they are not, and are never represented as, controls running in anyone's production. A reader who wants to test the framework's seriousness can start there, before taking any private claim on trust.

The discipline behind the framework was extracted from a multi-year private quantitative program I run almost entirely on frontier models, where every consequential design choice is captured as an architecture decision record and every failure is logged in a mistakes catalog rather than quietly patched. That program is not the evidence — the open libraries are. It is the origin story of the discipline the libraries encode,

learned rung by rung: a capability moved from human-approves-every-action to acting-inside-an-envelope only after the accumulated record showed the envelope held. Write down the decision. Log the failure. Prove the rung. Keep a control the system cannot switch off.

The same discipline is what regulated enterprise deployments increasingly demand: multi-agent governance operating under approval gates, sampled review, and a non-overridable control layer, with a live record of what each agent did and why. The transferable lesson is the same five rungs, the same climbing rule, and the same insistence on evidence before authority — whether the system allocates a research portfolio or operates in a regulated enterprise decisioning environment. The discipline does not change with the domain; only the decision class and the regulatory map around it do.

A fair skeptic will press the obvious question: has this ladder governed a live regulated production agent? I will be precise about the boundary. The libraries are reference patterns, not controls running in anyone's production; applying the ladder to a specific in-production agent, under a named regulator's decision class, is the engagement this framework is built for next, not a track record I am asking you to take on trust.

Where This Sits in the Literature

Several prior frameworks describe autonomy levels, and this one is built to engage them rather than ignore them (full citations in References). Morris et al. (2024) define levels of AGI by performance and generality, treating autonomy as a separable deployment choice; Feng, McDonald, and Zhang (2025) define five autonomy levels by the *user's* role, from operator to observer; Kasirzadeh and Gabriel (2025) construct four-dimensional "agentic profiles" across autonomy, efficacy, goal complexity, and generality; and Mitchell et al. (2025) argue that fully autonomous agents should not be developed, because risk to people scales with the autonomy ceded — advocating instead for semi-autonomous systems with clearly defined constraints and human oversight. The Autonomy Ladder differs from these in two ways. It is built around the **regulator's question** — what control surface, what evidence, what demotion path — rather than the **researcher's question** of what capability and what risk class. And it is operationalized through six DOI-archived, vertical-specific reference libraries rather than left as a typology. Where Mitchell et al. counsel against the top of the autonomy curve, the ladder makes the same caution operational: most regulated agents stay at A1 or A2, A4 is rare and heavily gated, and every rung is demotable. Read as control infrastructure, the same three primitives — a non-overridable sovereign veto, a tamper-evident ledger, and demotable rungs — are a concrete way to operationalize the independent oversight, and the prevention of an irreversible ceding of control, that safety researchers have argued deployed autonomy will require.

A Worked Example: An A2 Envelope for a Banking Adverse-Action Agent

To make a rung concrete, here is an illustrative A2 envelope for an agent that drafts adverse-action notices under ECOA/Regulation B. It is a pattern, not a deployed control — the kind of specification the `banking-agent-audit` library encodes.

IN-ENVELOPE — the agent writes
Adverse-action notices for declines driven solely by enumerated, objective, non-discretionary reasons (e.g., debt-to-income above a published threshold; delinquency on file), where reason codes map one-to-one to pre-approved Reg B language. The boundary is enforced by the system, not the agent.

OUT OF ENVELOPE → human
Any judgment call, thin file, inconsistent reason set, or protected-class-adjacent factor — out by construction.

AMBIGUOUS → human
Any case the agent's own confidence flags as uncertain. Escalation is mechanical, not left to the agent's restraint.

SAMPLED REVIEW of a fixed fraction of in-envelope notices → if reason codes drift from the actual decision basis → **DEMOTE to A1** (human approves every notice) until the record is clean. A normal operating move, not a failure.

Figure 3 — An A2 envelope for a banking adverse-action agent: a hard, system-enforced boundary, mechanical escalation, and a sampled-review tripwire back to A1.

- **In-envelope (the agent may write):** adverse-action notices for declines driven solely by a documented, enumerated set of objective, non-discretionary reasons (e.g., debt-to-income above a published threshold; delinquency on file), where the reason codes map one-to-one to pre-approved Reg B language.
- **The hard boundary:** the envelope is enforced by the system, not the agent. Any decline touching a judgment call, a thin file, an inconsistent reason set, or a protected-class-adjacent factor is *out of envelope* by construction and routes to a human.
- **The escalation trigger:** out-of-envelope by rule, plus any case the agent's own confidence flags as ambiguous — escalation is mechanical, not left to the agent's self-restraint.
- **Sampled review:** a fixed fraction of in-envelope notices is human-reviewed after the fact, checking that the stated reasons match the decision and the Reg B language is correct; the sampling rate is set to catch drift before it compounds.
- **The demotion trigger:** if sampled review finds the reason codes drifting from the actual decision basis, the agent drops to A1 (human approves every notice) until the record is clean again — a normal operating move, not a failure.

Calibrating the two numbers (illustrative, not prescriptive). The envelope leaves a firm two values to set, and — as Section 6 concedes — both are empirical and domain-specific. A workable procedure makes them explicit rather than implicit: set the sampled-review rate from the decision's base error rate and the cost of a missed error — high enough that each review period yields enough sampled errors to detect a meaningful rise in the error rate at a chosen confidence level — tightening it where a wrong decision is costly and relaxing it only as the record accumulates clean periods; and pre-register a demotion threshold (a sampled-mismatch rate sustained over a fixed number of consecutive review periods) so the drop to A1 is mechanical, not discretionary. The rate, the threshold, and the period count are exactly the parameters each vertical library's test suite pins down per decision class — which is where this qualitative rule becomes a falsifiable control.

That is one envelope for one decision class. The ladder gives every regulated agent the same shape; the vertical libraries give each the specific envelope, escalation triggers, and evidence its regulator will ask to see.

The Same Envelope, as Running Code

The envelope above is a diagram until you can run it. So here it is as code — the three controls (a sovereign veto, a tamper-evident ledger, mechanical demotion) exercised end-to-end against the open `banking-agent-audit` library, in under a minute. Every import path and call signature is the real published API; nothing here is pseudocode.

```python
```

pip install "git+https://github.com/linus10x/banking-agent-audit@v0.1.3" # Python 3.12+

from banking_agent_audit.governance import (AdverseActionGate, AuditChain, SovereignVeto, VetoReason, DEFCONMachine, RiskMetrics,) from banking_agent_audit.governance.adverse_action_gate import (AdverseActionDecision, AdverseActionType,)

One tamper-evident, hash-chained ledger underwrites every action below.

ledger = AuditChain(deployer_id="example-bank", mode="advisory") gate = AdverseActionGate(audit_chain=ledger) veto = SovereignVeto(agent_id="adverse-action-agent", audit_chain=ledger) defcon = DEFCONMachine(audit_chain=ledger)

def notice(applicant, reason): # stand-in for the agent assembling a Reg B notice return AdverseActionDecision(applicant_id=applicant, action_type=AdverseActionType.DENIAL, principal_reasons=(reason,), notice_provided=True, days_to_notice=9, used_consumer_report=True, cra_name_provided=True, credit_score_disclosed=True, applicant_rights_disclosed=True,)

A3: the agent drafts adverse-action notices autonomously — under a veto it cannot clear.

print("agent may act:", veto.allow_execution(), "| DEFCON:", defcon.current_level().name)

In-envelope: a specific, accurate basis. The gate passes it; the ledger records it.

ok = gate.evaluate(notice("app-1001", "debt-to-income ratio 0.62 exceeds the 0.43 product limit")) print("in-envelope notice compliant:", ok.compliant)

Out-of-envelope: the generic reason CFPB Circular 2022-03 forbids. Drift is detected, and it mechanically trips the sovereign veto and escalates the demotion machine.

drift = gate.evaluate(notice("app-1002", "score too low")) if drift.warnings: veto.trigger(VetoReason.MODEL_DRIFT, triggered_by="monitor:reason-code-drift", description="generic adverse-action reason breached the envelope") defcon.evaluate(RiskMetrics(consecutive_losses=8))

The agent is now demoted out of autonomous drafting — and cannot self-clear.

print("agent may act:", veto.allow_execution(), "| DEFCON:", defcon.current_level().name)

Only a human clears the veto. Then the entire trail proves tamper-evident.

veto.clear("reviewed; reason corrected to a specific basis", operator_id="fair-lending-lead") print("after human review, agent may act:", veto.allow_execution()) print("ledger events:", len(ledger), "| tamper-evident verify:", ledger.verify()) ```

Running it prints the safety case in five lines:

```
text agent may act: True | DEFCON: NORMAL # granted authority at A3 in-envelope notice compliant: True # a specific, accurate basis passes the gate agent may act: False | DEFCON: HALT # a generic reason trips the veto and demotes — mechanically after human review, agent may act: True # only a human clears it; the agent cannot self-clear ledger events: 5 | tamper-evident verify: True # the whole trail is hash-chained and verifiable
```

That is the entire thesis, executable: authority is granted at a rung, a drift signal the agent did not get to veto pulls it back down, a human — not the agent — restores it, and every step lands in a ledger that verifies. The decision class here is fair-lending; swap it for a tool call in an agent swarm or a perception-to-execution handoff in a robot fleet and the same three controls hold. The full five-station version, with the regulatory citations on every gate, is `examples/worked_example_adverse_action.py` in the repository.

4. The Regulatory Map: Where the Ladder Meets the Rule

A governance framework that does not map onto the rules a firm already lives under is a slide, not a control. The Autonomy Ladder slots into six obligations regulated institutions already carry — and the same artifact answers all six. For a lab or cloud platform selling agentic capability into these institutions, that map is also the vocabulary your field teams can use to answer a customer's model-risk function directly, instead of building a one-off governance story per deal.

SOC 2 Type 2. SOC 2 Type 2 (an attestation against the AICPA Trust Services Criteria) attests that controls operate effectively over a period of time, not merely that they exist on paper. This is what the ladder is designed for: the per-rung evidence record, the sampled review at A2, and the live audit ledger at A3 and A4 are exactly the operating-effectiveness evidence a Type 2 examination wants. An agent's autonomy level becomes a control with a documented operating history rather than an assertion. I have earned SOC 2 Type 2 under crisis conditions and know what the examiners actually test; the ladder's artifacts are designed to be that test's evidence.

ISO/IEC 27001. ISO/IEC 27001 frames security as a managed system with defined controls and continual improvement. Agent autonomy fits as a risk-treatment decision: each rung is a control, promotion is a change subject to the management system, and demotion is a defined response to degraded assurance. The ladder gives the information-security management system a vocabulary for autonomy that it otherwise lacks.

FINRA. For broker-dealers and the firms that serve them, FINRA's supervision rule (Rule 3110) requires that a firm reasonably supervise its activities — and increasingly its automated ones. The ladder operationalizes supervision: A1's per-write approval, A2's sampled review and hard envelope, and A3's supervise-by-exception with a live ledger are concrete supervisory mechanisms a firm can point to when asked how an automated decision process is overseen. The climbing rule itself — autonomy granted only against evidence, and demotable — mirrors the supervisory expectation that controls scale with risk.

Federal banking model-risk guidance. For US banking organizations, model risk has long been governed by interagency supervisory guidance; non-banking readers can skip to the next obligation. The structural point for everyone else is that this guidance was written for traditional, statistical models — and AI-specific supervisory guidance for generative and agentic systems is still emerging. That leaves an institution deploying agentic AI to demonstrate bounded, governed operation through a framework of its own. The Autonomy Ladder applies the same discipline model-risk management has always demanded — documented evidence, independent challenge, and a control the model cannot override — to agentic systems, giving a bank a concrete answer to "what governs this agent's authority and who can stop it."

NIST AI Risk Management Framework. The NIST AI RMF organizes AI risk management around the functions of governing, mapping, measuring, and managing AI risk. The Autonomy Ladder is a direct instrument of those functions: it *governs* by making the autonomy level an explicit, owned decision; it *maps* by tying each rung to a decision class and its risk; it *measures* through sampled review, the audit ledger, and the mistakes catalog; and it *manages* through the envelope, the sovereign veto, and demotion. A firm adopting the NIST AI RMF can use the ladder as the concrete autonomy-control the framework asks for but does not itself specify.

EU AI Act. The EU AI Act imposes obligations on high-risk AI systems, including the Article 14 human-oversight requirement — that such systems be designed so humans can effectively oversee them, including the ability to intervene or interrupt. The ladder is, in effect, a graduated human-oversight design: A0 and A1 keep humans on every decision; A2 and A3 preserve oversight by sampling and exception with a non-overridable veto; and the demotability of every rung is precisely the "ability to intervene and stop" the Act contemplates. The framework gives builders a way to evidence human-oversight conformity as a property of the system's architecture, not a promise in a policy document.

Across all six, the common thread is that regulators and standards bodies are converging on the same demand: autonomy must be governed by documented evidence, independent control, and a real ability to stop the system. The Autonomy Ladder is one concrete way to satisfy that demand consistently, so a firm operating in several regimes at once is not maintaining a different governance story for each.

How the Ladder Relates to Frontier-Lab Capability Frameworks

A reader inside a frontier lab will pattern-match "A0 through A4" to the two most-cited industry governance frameworks: Anthropic's Responsible Scaling Policy (v3.0, effective February 2026), built on AI Safety Levels (ASL-1 through ASL-4+), and Google DeepMind's Frontier Safety Framework (v3.0, September 2025), built on Critical Capability Levels (CCLs). The pattern-match is worth making explicit, because the distinction is the point.

Framework	Governs	Primary user
Anthropic ASL / DeepMind CCL	The *model's* capability — what it can do at all	Frontier labs deciding what is safe to release
Autonomy Ladder A0 → A4	The *deployed agent's* authority in a specific decision class	Regulated institutions deploying agents inside a workflow

These are complementary, not competing — they govern different layers. A capability ceiling set by ASL or CCL says what a model is permitted to be; an autonomy rung set by the ladder says what a deployed agent is permitted to *do* in a given regulated decision class. ASL-3 paired with A1 is a coherent deployment posture: a capable model, tightly held. ASL-3 paired with A4 inside a fair-lending decision class is not — the capability may be licensed, but the authority has outrun the evidence. The Autonomy Ladder is, in this

sense, the deployment-side counterpart to the labs' capability frameworks: they decide what to release; the ladder decides what it is allowed to do once it is inside a bank.

The mapping also explains why the reference libraries are vertical-specific rather than one generic toolkit. The *rungs* are universal — every regulated agent climbs the same A0-to-A4 ladder — but the *decision class* at each rung, and the obligation that governs it, are not. A banking agent's A2 envelope is bounded by fair-lending and model-risk obligations. A payments agent's A3 in-scope decision class is bounded by the irreversibility of rail finality — once a payment settles, there is no human-in-the-loop to undo it, which raises the evidence bar for any rung that lets the agent initiate one and makes the sovereign veto's pre-settlement window the load-bearing control. An investment-adviser agent's envelope is bounded by suitability and the duty of care owed to the client. Same ladder, different envelope. That is the practical reason a single generic "AI governance policy" tends to fail an examination: it names the principle but not the decision class, and the examiner tests the decision class.

One corporate framework belongs in the same picture. Microsoft's Responsible AI Standard v2 (2022) formalizes a cross-functional development process — impact assessments, identified harms, mitigation reviews — at the *organizational* layer the ladder deliberately does not occupy. The two are complementary: Microsoft's standard governs how an institution decides to build an AI system; the ladder governs what authority that system is allowed to hold once it is deployed inside a regulated decision class.

Bridge to Frontier Autonomy Stacks

The ladder is not a finance artifact wearing a governance label; it is a deployment-authority discipline, and the *decision class* is a parameter. The same five rungs, the same climbing rule, and the same three controls apply whether the in-scope decision is a fair-lending adverse action or a real-time perception → planning → execution handoff in a vehicle or robot fleet, an autonomous trade, or a tool call inside a coordinated agent swarm. Four mappings make the transfer concrete:

- **Multi-agent coordination risk → A4 + validated escalation.** A4's defining hazard is not a single bad write but a cascade no one agent owns. The climbing rule withholds A4 until inter-agent escalation paths have been tested under realistic failure — the same bar a robot fleet or an agent swarm needs before it coordinates in production.
- **Sovereign veto as infrastructure, not a prompt.** A non-overridable veto with its own process boundary, its own credentials, and a hardware-or-network gate the agent cannot reach is exactly the kill-path a high-autonomy physical or software system needs — and, per Section 2, it carries a latency budget, an availability target, and a failover posture, because a veto that adds latency to a control loop is a design constraint, not a footnote.
- **Tamper-evident audit ledger → safety cases and post-incident reconstruction.** A write-once, hash-chained record of every action is what turns "we believe it was governed" into a reconstructable safety case after an incident — the artifact an investigator, a regulator, or an internal review board actually needs.
- **Evidence bars + mechanical demotion → scope expansion under control.** Expanding an autonomous system's operational design domain is a rung promotion; degraded assurance is a demotion. Making both explicit, logged, and reversible is how scope grows without authority outrunning evidence.

The vertical libraries encode these controls for regulated decision classes today; the control surfaces themselves are domain-agnostic, which is the point — a frontier lab or autonomy team can lift the same A0→A4 structure into its own decision classes without inheriting any financial-services assumptions.

That last claim is runnable, not rhetorical. The `finserv-agent-audit` library ships a domain-agnostic agent-swarm example (`examples/agent_coordination/`) with no finance anywhere in it: a worker agent at A3 attempts an irreversible `dispatch` action (think launching a job, deleting a volume, sending an external message); the hard envelope routes it to a human; the sovereign veto blocks execution and refuses the agent's own attempt to clear it; the hash-chained ledger records every step and `verify()` confirms the chain; and two out-of-envelope attempts mechanically demote the agent from A3 to A2 — every transition written to the ledger. Same three controls, same climbing rule, a decision class with no regulator in sight. It is the shortest path from "the controls are domain-agnostic" to watching them govern a non-financial swarm in your own terminal.

5. Implications: What Changes If You Adopt the Ladder

Adopting the ladder changes *who can sign off on a deployment* — and that single shift cascades. This holds whether the adopter is the regulated institution itself or the lab or platform whose product is doing the deploying on its behalf.

Today, the autonomy decision is often made implicitly by whoever holds the configuration access: an engineer, a product owner, a vendor's default setting. The ladder relocates that decision to where the accountability already lives. Because each rung carries a named evidence bar and a named owner, the person accountable for the regulatory exposure is the person who approves the rung, and they approve it against a record they can read rather than a demo they watched. The moment a control's promotion requires evidence and an owner's signature, the organization stops promoting on optimism.

From that one shift, three things follow. **Autonomy becomes a logged decision, not a configuration drift** — every promotion has an owner, an evidence basis, and a record, which kills the invisible-promotion failure mode at the root. **The intermediate rungs become legitimate destinations** — once A1 and A2 are named and defensible, a firm can say without apology that its agent operates at A1 because that is where the evidence supports it, a deliberate, examinable choice rather than a failure to "fully automate." And **demotion becomes a normal operating move** — because every rung is demotable, an organization can drop an agent's autonomy when assurance degrades without treating it as a crisis. "We pulled this agent back to A1 last week because the sampled review flagged drift" becomes a sign of a healthy governance process rather than an embarrassment to hide.

The deeper implication is the reordering of the deployment sequence. The prevailing order is *grant authority, then add controls when something breaks*. The ladder's order is *prove the evidence, then grant the rung, and keep a control you cannot switch off*. The cost of the disciplined order is paid up front, in slower initial promotion and the overhead of keeping records; the cost of the undisciplined order is paid later, all at once, in an incident no one can explain. Having paid the second kind of cost in a twelve-day hard-down with no disaster recovery, I will take the first kind every time.

So here is the question worth sitting with: **where does your highest-authority autonomous system sit on this ladder today — and can you name the person who promoted it there, and the exact evidence they used?** If the answer is not immediate, the system is operating above its earned rung.

6. Three Open Questions

The framework is deliberately offered as a structure to be tested, not a finished answer. Three questions are genuinely open, and I do not claim to have resolved them.

First: how much sampled review at A2 is enough? The framework requires that a human review a sampled subset of in-envelope writes, but it does not — and at this stage cannot — specify the right sampling rate as a function of decision risk, base error rate, and the cost of a missed error. Too low and drift goes undetected; too high and the productivity gain that justified A2 evaporates. The honest answer is that the right rate is empirical and domain-specific, and the field does not yet have good guidance on setting it. The interim procedure in the A2 example above is the best partial answer the framework offers: derive the rate from the base error rate and the cost of a missed decision, pre-register the demotion threshold, and let the value tighten or relax against the accumulating record — the right number stays empirical, but the procedure makes it explicit and auditable rather than buried in a configuration file.

Second: can the sovereign-veto layer at A3 and A4 be made genuinely independent of the agents it governs? Here I will commit to a partial answer rather than leave it fully open. **Architectural independence is engineerable today; semantic independence is partial.** A veto layer can be given a separate process boundary, separate credentials, a different model where applicable, and a hardware or network gate the agent cannot reach — that much is buildable now for an in-scope decision class. It becomes harder, and in some architectures genuinely unsolved, when the veto layer itself uses model inference to make its decisions, because the veto model is then subject to some of the same failure modes as the agent it governs. The practical consequence: A3 deployments today should rely on architectural independence, and A4 deployments should not be approved without an explicit account of how degradation in semantic independence is monitored.

Third: who owns the promotion decision at A4? At A4, promotion is monitor-led rather than human-flag-flipped — a monitoring layer governs scope expansion. But this raises an obvious regress: the monitor that decides when an agent may climb is itself an automated control whose own authority must be governed. A partial answer is that the monitor is itself a control on the ladder — it carries its own evidence bar, writes its own entry to the audit ledger for every promotion it grants, and has its own demotion path. What stays genuinely unsolved is the base case of the regress: the highest monitor's authority must ultimately rest on a human-owned policy, not another automated layer. The framework does not yet have a fully satisfying account of that base case, and I would rather flag the gap than paper over it.

These are not abstract. They are the questions I work through with teams deciding what rung to trust. If your institution is wrestling with any of these three — the right sampling rate, the achievable veto independence, or the governance of the governor — that is precisely the conversation worth having. The framework exists to host it, not to foreclose it. They are also where input from safety and alignment researchers would most sharpen it: each sits exactly where capability-side caution meets deployment-side control, and I would rather stress-test them in the open than resolve them quietly.

7. Where to Go From Here

If the framework is useful to you, there are three paths, graded by friction:

Read it. The six reference libraries are public and openly licensed — all MIT, four dual MIT OR Apache-2.0 — at **github.com/linus10x** — start with `finserv-agent-audit`, or the vertical closest to your decision class (`banking-agent-audit`, `payments-agent-audit`, `payer-agent-audit`, `private-capital-agent-audit`, or `cre-agent-audit`). Each is DOI-archived. That is the zero-friction way to test the framework's seriousness for yourself.

Build with it. For a team deploying agentic AI into regulated decision classes — at a frontier lab, a cloud platform, or a regulated institution itself — the ladder is built to be lifted straight into the governance vocabulary your account teams, solutions architects, and forward-deployed engineers use with customers.

The libraries are openly licensed and free to adopt. Where a specific decision class needs depth — the exact envelope, the escalation triggers, the evidence a particular examiner or safety review will demand — that is where I work directly with a team, on request and under NDA, to place its in-production agents, or a production agent swarm, on the ladder against the coordination, escalation, and decision-class risk it actually carries.

Work with me. I build the governance, then I build the system that lives under it. The work I take on is where this framework runs in production under real regulatory exposure — leading technology and AI for a regulated vertical, or embedded with a lab or platform team taking agentic AI into regulated customers. If that is the problem in front of you, that is the conversation I am built for — **autonomy-ladder.io**, or find me on LinkedIn.

Acknowledgments and Influences

This framework operates in a field shaped by a substantial literature on AI risk and accountability, and it is built to take that literature's concerns seriously — and to say, for each influence, what is different here. The case for caution and oversight as AI systems gain capability and autonomy — argued prominently by Yoshua Bengio and the broader AI-safety community — is real; the Autonomy Ladder operationalizes that caution as a *deployment* discipline rather than a *capability* stop, governing what a model is allowed to do in a decision class rather than whether it should exist. Cathy O'Neil's *Weapons of Math Destruction* documents the harms of opaque, unaccountable automated decision systems; the ladder treats that opacity as the specific failure mode the live audit ledger and the per-rung evidence record are designed to make impossible. Established documentation practices for machine learning — model cards, datasheets, and the discipline of recording what a system does and where it fails — inform its documentation-first posture, extended from describing a single model to governing a multi-agent system's authority over time. In financial-services decisioning, the case for auditable, governed AI advanced by practitioners such as FICO's Scott Zoldi parallels the ladder's regulatory mapping; the ladder carries that lineage from single-model auditability into the harder problem of multi-agent-system governance. And the evidence-led tradition in quantitative finance shaped the discipline of proving each rung before granting it.

These attributions describe the general thrust of each author's published work; no endorsement is implied, and readers should consult the primary sources for precise positions.

References

1. Anthropic. *Measuring AI Agent Autonomy in Practice.* February 18, 2026. anthropic.com/research/measuring-agent-autonomy
2. Anthropic. *Responsible Scaling Policy, Version 3.0.* Effective February 24, 2026. anthropic.com/news/responsible-scaling-policy-v3
3. Google DeepMind. *Frontier Safety Framework, Version 3.0.* September 2025. deepmind.google
4. M. R. Morris, J. Sohl-Dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, C. Farabet, S. Legg. *Levels of AGI for Operationalizing Progress on the Path to AGI.* ICML 2024. arXiv:2311.02462
5. K. J. K. Feng, D. W. McDonald, A. X. Zhang. *Levels of Autonomy for AI Agents.* 2025. arXiv:2506.12469
6. A. Kasirzadeh, I. Gabriel. *Characterizing AI Agents for Alignment and Governance.* 2025. arXiv:2504.21848

7. M. Mitchell, A. Ghosh, A. S. Luccioni, G. Pistilli. *Fully Autonomous AI Agents Should Not Be Developed.* 2025. arXiv:2502.02649
8. Microsoft. *Responsible AI Standard, Version 2.* June 2022. microsoft.com/ai/responsible-ai

Version History

- **v6 (June 2026):** Reframed around first-principles autonomy authority (the failure mode generalizes across vehicles, robots, and agent swarms, not only regulated finance); added a "Bridge to Frontier Autonomy Stacks" subsection mapping the rungs and controls to multi-agent coordination, sovereign-veto-as-infrastructure, safety-case audit ledgers, and scope-expansion-as-promotion; added a one-page, distribution-ready essence block; linked the control primitives to independent-oversight / irreversibility concerns; generalized the closing question to "highest-authority autonomous system."
- **v5 (June 2026):** Added an illustrative calibration template for the A2 sampling rate and demotion threshold; tied the open questions to that template and to a partial answer on governing the monitor; added specific obligation citations (EU AI Act Art. 14, FINRA Rule 3110, AICPA Trust Services Criteria).
- **v4 (June 2026):** Added figures — the A0→A4 ladder, the promotion/demotion state machine, and the A2 adverse-action envelope; sharpened the executive summary's per-audience value; opened the three questions to safety- and alignment-researcher critique.
- **v3 (June 2026):** Named A3/A4 as infrastructure with operational properties (latency, availability, separation-of-duty); reframed the evidence so the open libraries carry the weight and the private program is its origin; added the Microsoft RAI Standard as the organizational-layer complement; widened the framing to the labs and platforms deploying into regulated institutions, not only the institutions themselves.
- **v2 (June 2026):** Expanded to situate the ladder against the frontier-lab capability frameworks (Anthropic ASL / DeepMind CCL) and the autonomy-levels research literature; added the six-vertical reference-library detail and a References section; sharpened the open question on veto independence.
- **v1 (May 2026):** Initial framework — five levels, the climbing rule, the demotion rule, the regulatory map across six obligations, and the worked A2 banking example.

Autonomy Ladder™ is a framework authored by Kunjar Bhaduri. Client names, performance figures, and named systems are anonymized by design. Comments and critique welcome — autonomy-ladder.io.