WORKING PAPER

Governed Agency

A foundation for delegated machine authority: the laws, and the physics beneath them.

> *The last era of intelligence was measured by what a system could do. The next will be measured by what it can be trusted not to do, and trust counts only when the limit is proven, not promised.*

Ankush Chowdhary

Published by Neo | neocontrol.ai | v1.0 | 2026

Cite as: Chowdhary, A. (2026). Governed Agency: A Foundation for Delegated Machine Authority. Neo working paper.

I. The problem worth naming

This is not, first, a document about artificial intelligence. It is about extending agency beyond ourselves: how power is safely handed to an actor that is not you. That problem is as old as institutions.

Civilization scales through delegation. A board grants an executive spending authority; a pilot yields part of a flight to an autopilot; a bank sets an approval limit; a hospital lets a junior clinician carry part of a diagnosis. None of them asks *"do I trust this actor?"* Trust is the residue. They ask a harder, older question: *what authority am I willing to grant, under what conditions, and how will I take it back?* Every expansion of human reach has required a new institution to bound power with accountability.

Autonomous machines are simply the newest actor to arrive at that ancient problem. They do not change it; they only make it faster and impossible to ignore. And they arrive at a specific moment: capability has become cheap. A hundred systems will match each other on what they can *do*, so capability is no longer the differentiator or the bottleneck. The unsolved problem, the one that decides whether autonomy is adopted at all, is **delegation**: granting a non-human actor real authority and still being able to take it back.

This paper states the laws of that problem and the physics beneath them. It is deliberately domain-general. Cybersecurity is merely the first place the laws must be made real and verifiable, the proving ground, not the boundary.

II. The theorem

Beneath everything else is one law:

> ***Delegation without bounded consequence is not delegation. It is abdication.***

A board that hands over unbounded authority has not delegated; it has abdicated. So has the parent, the general, the regulator. What separates the two has never been the grant of power. It is the *boundary* around it: the limit, the reversibility, the accountability, the ability to withdraw.

This is why ungoverned autonomous action feels wrong even to those who cannot name why. We have been confusing delegation with abdication. And the confusion leaves only two moves to anyone who cannot bound consequence: refuse to delegate anything meaningful, which is paralysis, or delegate recklessly, which is abdication. Neither scales. Both are what the field is doing right now.

Everything that follows is either a physical reason this theorem holds, or a law that enforces it.

III. The physics: why authority must be re-earned

1. The entropy of delegation. More training makes a model better only while its data stays anchored to something real. The moment a system trains on its own output, the arrow reverses: fidelity erodes, variance collapses, errors compound, until the system is fluent and wrong. This is *model collapse*, and it is general. Any system that learns from its own increasingly-corrupted output without re-grounding decays. Standing authority is the same failure. Granted once and never re-verified, it drifts as ownership moves and reasons are forgotten; and each unverified grant becomes the precedent for the next, so *"it had access before"* trains *"grant it again."* Authority trained on authority. This is why *no standing authority; every action a fresh delegation* is not merely a security control. It is an anti-entropy principle. And machines compress the timeline: corruption that took institutions generations now takes a model a few cycles. **Ungoverned delegation is model collapse for authority.**

2. Reset as defense. The defense against runaway corruption is not perfect memory. It is the checkpoint and the rollback: the ability to discard corrupt state and return to a known-good one. The reset must be *lossy* to work, because carrying forward a complete record of every corrupted cycle carries the corruption with it. This reframes reversibility. The way back is the deliberate reset made continuous, so a system takes constant small resets (authority that expires each action) and never needs the cataclysmic one. It also names its cost: you cannot casually reset what you cannot reconstruct, which is why the inherited, undocumented systems in every estate are treated as ruins, contained rather than bulldozed.

3. The anchor and the reader. To stop the decay, state must be anchored to something true, a preserved reference. But a preserved anchor is worthless if the *reader* is degraded: an intact policy handed to a hallucinating model is right in every word and wrong in every reading, and the reader is least reliable exactly when the stakes are highest. So the anchor cannot be a description that must be *understood*; it must be a fact that can be *checked*. This is *proven, not felt*: ground authority in the capability actually demonstrated, the owner actually resolved, the state actually snapshotted. Proof survives a degraded reader because it only has to be verified, not comprehended.

4. You cannot trust the render. Perception is prediction, not recording. The brain generates a best guess and stores a compressed, self-serving archive. A generative model is the same machine; hallucination is not its failure mode but its *mechanism*. So an actor's account of reality, human or machine, is a render by construction, and authority can never rest on what an actor *reports* or *believes*. Memory has the same defect, so accountability must be an external, tamper-evident record written as the action happens: you govern from the ledger, not the confession. And the mirror the governor must hold to itself: the most dangerous failure is not a rogue agent but a **control plane severed from ground truth**, enforcing its own accumulated policy, certain it is order when it is only control. The anti-entropy law cannot exempt the thing that enforces it. *The governor must re-ground itself to proof every cycle, or become the failure it was built to prevent.*

5. Character or player. Complexity is not agency. A scripted actor can be intricate and convincing and still run loops, react to triggers, and never re-ground, which is precisely a capable agent today: a sophisticated character executing an objective with no awareness of consequence. Making it more capable makes it a better character, not a governed actor. The

difference is not intelligence; it is awareness, the capacity to step outside the loop and re-ground. And you cannot argue code into awareness: a script will run regardless, so restraint is never negotiated *with* the actor, only installed *around* it. Judgment must be infrastructure, not personality.

6. Privilege. Legitimate privilege obeys two rules. **Earned, not default:** higher privilege unlocks only on demonstrated reliability, graduated, against a track record, revocable. Privilege-by-default is the vulnerability every audit finds. **Granted from outside, never self-authored:** an agent that can edit its own authority has performed privilege escalation, the one thing a governed actor must never do. The sharpest case is root, the power to make things irreversible, which for that reason can never be a standing state, only a momentary grant against a proven way back.

IV. The laws

From the physics, the constitution. These are the laws that make the theorem hold.

1. **No standing authority. Every action, a fresh delegation.** Authority is allocated, exercised, and released, like memory, like compute, re-earned each action and recomputed against what is *proven*, not felt. It is the successor to Zero Trust: that ended standing access; this ends standing authority to act.
2. **Judgment is installed, not innate.** Do not trust the actor; trust the architecture that bounds its authority. The consequential, irreversible calls stay human.
3. **The measure is the futures left open.** A limit is not a flinch. The mastery is to act *and* keep every recoverable path open: optionality preserved, not action refused.
4. **Every delegation decision reasons across five things at once,** and almost no system holds more than one:

- **Capability**: what is possible, proven rather than assumed. - **Authority**: what has been delegated, resolved to who truly owns it. - **Consequence**: what changes if it succeeds, and how far the blast reaches. - **Reversibility**: whether it can be undone, and at what cost. - **Accountability**: who owns the decision and the outcome, on the record. To fuse these into one grounded decision, at the speed of the action itself, is the whole of the discipline.

The maxims, for the wall:

> *Delegation without bounded consequence is abdication.*
> *No standing authority. Every action, a fresh delegation.*
> *Never act beyond what you can undo, or beyond what you were given.*
> *An autonomous actor's power is what it can do; its trustworthiness is the futures it leaves open.*
> *Do not trust the actor. Trust the architecture that bounds its authority.*

V. What a governed intelligence is for

It would be easy to read all of this as being about restraint, about what a system must be stopped from doing. That is the smaller half. The point of bounding power was never the bound; it is what the bound makes possible.

Civilization has only ever scaled through delegation, and every expansion of reach demanded a new institution to make that delegation safe. Autonomous machines stall at the demonstration precisely because that institution does not yet exist for them: no one sane hands unbounded authority to a system that cannot be re-grounded or reversed. So a governed intelligence is not the brake on that future. It is the **precondition** for it: the thing that makes it possible to grant a machine real authority and still hold the power to take it back.

What it preserves is not safety in the defensive sense. It is *optionality*: the ability to act and to have left every recoverable path open. Its measure is not the actions it can take, but the futures it declines to foreclose.

That is the whole shift, stated plainly. The last generation of systems was built to be **capable**. This one must be built to be **delegable**: to accept authority it did not grant itself, bounded by consequence it can compute and reversibility it can prove, returning every cycle to what is true. A system that does this is no longer merely intelligent. It is governed, and a governed intelligence is the first kind we can afford to trust with anything that matters.

Governed delegation, therefore, is not an optimisation of autonomous systems. It is the condition under which extending real authority to a non-human actor becomes possible at all. It is the institution this moment is missing: the one that lets us hand over meaningful power and still, at any instant, take it back.

That institution is what must now be built.

The arguments here were triangulated across machine learning, systems engineering, and cognitive science, and, in their first drafting, across cosmology and myth, which are left at the door so the reasoning stands on its own. The technical form is the load-bearing one.