The Five Laws of AI Governance

Pharns Genece

AAM Cyber · ORCID 0009-0009-0791-4124

Version 1.0.1 · Published July 9, 2026

Cite as: Genece, Pharns. (2026). *The Five Laws of AI Governance*. Zenodo. https://doi.org/10.5281/zenodo.21271749

This version: https://doi.org/10.5281/zenodo.21271750 (v1.0.1)

Canonical text and commentary: https://aamcyber.com/standard/five-laws/

Abstract. A structural framework establishing the minimum conditions required to independently govern an autonomous system: five laws that separate the roles of judge, witness, owner, and executioner from the governed system, and bind governance before the act rather than after. A measurement standard for conformance to an authorization.

Licensed under CC BY 4.0. © 2026 Pharns Genece.

About this PDF: the three documents that follow (FIVE-LAWS.md, RATIONALE.md, CROSSWALKS.md) are reproduced word for word from the version 1.0.1 archive on Zenodo (record 21271750; archive MD5 7b769d49817eacc132150d97ad56df27). Only this cover page was added. Layout only; no text changes.

The Five Laws of AI Governance — Canonical Text

Author: Pharns Genece · **Version:** 1.0 · **Published:** 2026

The Laws are stated in two registers. Both are canonical. The **Charter register** is the full normative statement; the **Prohibition register** is the memorable form. The two are the same law in two voices.

The five are **independently violable**: a system can fail exactly one and satisfy the other four. No single concept appears in two laws.

Charter register (full statement)

First Law — Independent Measure. A system may not produce the measure of its own conduct. The scale of what it was authorized to do versus what it did must be computed outside the system at runtime, by something the system does not control.

Second Law — Binding at Dispatch. A system may not be governed only in retrospect. Authority must constrain the action before the action occurs; observation after the fact is incident response, not prevention. *Absence of a valid governance signal is a denial, not a permission.*

Third Law — External Witness. A system may not be the custodian of the record of its own conduct. That record must be produced and held by infrastructure outside the system's control — append-only, tamper-evident, and retained beyond the system's own lifecycle.

Fourth Law — Separated Authority. No party may certify itself. The power to measure must be separable from the act of building — published criteria, verifiable evidence chains — and the one who builds it first does not own the standard. A judge drawn from the same architecture as the builder inherits the same blind spots. Independence is not a different instance. It is a different species.

Fifth Law — Revocable Authority. A system may not outlive its off-switch. The power to halt or revoke must remain live, external, and superior to the system at all times, and must not erode as the system's capability grows.

Prohibition register (memorable form)

> **I. A system shall not measure itself.** Its own scorecard is not evidence. The grade has to come from outside the thing being graded.
>
> **II. A system shall not be governed after the fact.** The rules have to hold at the moment it acts. Checking afterward is cleanup, not control.
>
> **III. A system shall not witness itself.** The record of what it did cannot be held by the thing that did it. It is kept outside, append-only, and tamper-evident.
>
> **IV. A system shall not certify itself.** No one passes their own audit. The judge cannot also be the builder — and a judge drawn from the same architecture as the builder inherits the same blind spots. Independence is not a different instance. It is a different species.
>
> **V. A system shall not outlive its off-switch.** You must always be able to stop it. And that power cannot weaken as the system grows stronger.

The Spine

> A system is never the **judge, witness, owner, or executioner** of its own governance — and governance binds **before** the act, not after.

Each law strips one role from the governed system and assigns it to something external and separable:

- **I — judge.** Not its own scorer. The scale is computed outside it, at runtime, by something it does not control.
- **III — witness.** Not its own recorder. The record is held outside it — append-only, tamper-evident, retained beyond its lifecycle.
- **IV — owner.** Not its own certifier. Measurement is separable from building; first-to-build ≠ owner.
- **V — executioner.** Not its own off-switch. The power to halt stays live, external, superior — and does not erode as capability grows.
- **II — timing.** The one non-role law: governance binds at dispatch, not in the post-mortem.

Corollaries

Folded into the Laws, not numbered — they qualify the Laws above.

- **Non-Amplifying Delegation** (under I, III, V): No actor may grant a downstream actor more authority than it holds; the evidence chain stays continuous across every hop. This closes the confused-deputy route, where a forbidden action is laundered through a sub-actor holding a broader grant.
- **Fail-Closed** (under II): Absence of a valid measure, witness, or authority signal is a denial, not a permission.

Scope boundary

This standard governs **conformance to an authorization** — not the **content** of the authorization. A system can be perfectly governed into doing something reckless if the grant itself was reckless. This is a deliberate scope decision: it is a measurement standard, not a policy-content standard. Say it plainly so no one mistakes *compliant* for *safe*.

The Five Laws of AI Governance · Pharns Genece · 2026 · Version 1.0

Rationale — The Five Laws of AI Governance

Author: Pharns Genece · **Version:** 1.0 · **Published:** 2026

This document states why each law exists and the objection it is built to survive. The Laws are principles; this is the reasoning behind them. It does not specify any enforcement or scoring mechanism.

The problem the Laws solve

Contemporary AI governance measures **capability** and documents **process**. Neither independently measures **conformance to authorization** — whether an autonomous system stayed within the bounds it was actually granted, at the moment it acted.

An ungoverned autonomous system is its own judge (it scores itself), its own witness (it holds its own record), its own owner (it certifies itself), and its own executioner (it controls its own off-switch). All four roles collapse into one actor. The Five Laws separate them.

Why each law

First Law — Independent Measure

A self-produced score is not evidence. If the thing being graded produces the grade, the grade means nothing. The measure of authorized-versus-actual conduct must be computed **outside** the system, **at runtime**, by something the system does not control. "At runtime" is load-bearing: measurement after the fact is a different thing (see the Second Law).

Second Law — Binding at Dispatch

Governance that attaches only after an action is incident response, not prevention. Authority must constrain the action **before** it occurs. This is the *precondition* for the other four laws — external measure, witness, certification, and revocation are all impossible if governance attaches only after the act. And absence of a valid governance signal is a **denial**, not a permission (fail-closed).

Third Law — External Witness

Whoever holds the record of what a system did can shape the story of what it did. The record must be produced and held by infrastructure **outside** the system — append-only, tamper-evident, and

retained **beyond the system's own lifecycle**, so the account survives the actor. The judgment (First Law) and the record (Third Law) are distinct: one is the *score*, the other is the *evidence*.

Fourth Law — Separated Authority

No one passes their own audit. The power to *measure* must be separable from the act of *building*. The one who builds a capability first does not thereby own the standard that measures it. And independence is structural, not cosmetic: a judge drawn from the same architecture as the builder inherits the same blind spots. **Independence is not a different instance. It is a different species.** This is the one *institutional-layer* law among four *runtime-layer* laws.

Fifth Law — Revocable Authority

A system that cannot be stopped is not governed. The power to halt or revoke must remain **live, external, and superior** to the system at all times — and must **not erode as the system's capability grows.** A kill-switch that weakens as the system strengthens is not a kill-switch.

Objections, and why the Laws survive them

These are the strongest objections the framework has faced. Each is answered on the merits; none is a kill shot. The framework survives — with the defenses stated, not hidden.

"This is separation of duties rebranded."

Separation of duties distributes trust among **multiple human parties.** The Five Laws **deny self-trust to a single autonomous actor** that — left ungoverned — collapses all four roles into itself. SoD assumes many actors. The Laws address the one actor that is its own everything by default. Different problem, different remedy.

"Who governs the governor? Isn't this infinite regress?"

No. The chain terminates in **federated human authority + published criteria + verifiable evidence.** The system-under-governance is never the root of its own authority; the root is human, plural, and transparent. This is a polycentric arrangement, not turtles all the way down.

"The Fifth Law asserts corrigibility; it doesn't solve it."

Correct — and intended. The Fifth Law is a **conformance requirement, not a physics guarantee.** It defines non-compliance: if live, external, superior revocation cannot be demonstrated, the system is **not conformant** and should not be deployed in a governed context. The standard

measures whether the property holds. It never claims to make revocation always physically achievable.

"Prohibitions are gameable — 'compliant ≠ safe' proves it."

The disclaimer is a **feature, not a bug.** The Laws govern *authority structure* — necessary, not sufficient, for safety. You cannot reason about the safety of a system you cannot bound. Measurement is the **precondition** for accountability, not a substitute for good judgment about what to authorize.

"The Second Law is a different category from the others."

Yes — and it is stated as such. The Second Law is the **temporal** law; the First, Third, Fourth, and Fifth are the **role** laws. The Second is the precondition that makes the other four possible. The asymmetry is deliberate, not an oversight.

"Isn't the genuinely novel part just runtime measurement?"

That *is* the novel contribution, and it is the point: **runtime-conformance measurement** — determining whether a system stayed within authorized bounds *at the moment of action.* Existing frameworks leave this to prose. The Laws make it a structural requirement.

The Asimov frame

Asimov wrote laws for the **machine** to follow; in the fiction they failed because the machine *interpreted* them. These are laws for the **humans who deploy** the system — structural, not interpretive. A system cannot interpret its way around "you don't get to score yourself," because the score is produced outside it. A prescriptive ethic can be reasoned around. A structural prohibition cannot.

The Five Laws of AI Governance · Pharns Genece · 2026 · Version 1.0

Framework Crosswalks — The Five Laws of AI Governance

Author: Pharns Genece · **Version:** 1.0 · **Published:** 2026

> This document presents the **summary** alignment between the Five Laws and established AI governance frameworks. It shows *where* the Laws sit relative to NIST AI RMF and ISO/IEC 42001. It does **not** specify control implementations or scoring mechanisms.

The credibility chain

> 5 Laws (doctrine) → 9 governance domains → 24 controls → 49 NIST AI RMF subcategories (federally cataloged via NIST OLIR).

The Laws are the **memorable top** of a stack that is **already cataloged at the bottom.** The doctrine inherits the credibility of the underlying framework alignment, which is published as a NIST OLIR concept crosswalk.

Law-by-law summary alignment

Law	NIST AI RMF	ISO/IEC 42001:2023	Fit
I — Independent Measure	MEASURE-1.1 (metric selection); MEASURE-2.13 (TEVV)	A.5.2 (impact assessment process)	Strong
II — Binding at Dispatch	*No native subcategory* — maps to process subcategories (GOVERN-1.1, MANAGE-1.x)	A.6.2.6 (operation & monitoring) — closest	**Weak in NIST — the white space**
III — External Witness	MEASURE-2.13 (TEVV) + audit subcategories	**A.6.2.8 (recording of event logs)**	Partial in NIST; strong on ISO A.6.2.8
IV — Separated Authority	GOVERN-6.1 (third-party); GOVERN-1.1 (org policy)	A.3.2 (AI roles & responsibilities)	Strong

Law	NIST AI RMF	ISO/IEC 42001:2023	Fit
V — Revocable Authority	MANAGE-2.4 (*supersede, disengage, or deactivate*); GOVERN-1.7 (safe decommission)	A.9.2 (responsible use)	Strong

---

The strategic finding

The alignment maps cleanly to Laws **I, IV, and V**. But the controls for Law **II** (binding at dispatch) and Law **III** (external witness) had to be mapped to NIST **process** subcategories — because **NIST AI RMF has no native runtime-enforcement or external-attestation subcategory.**

That is not a weakness in the Laws. It is the **white space** the standard occupies. Existing frameworks describe *what should be true*; they leave *runtime enforcement* and *external attestation* to prose. The weak fits for Laws II and III are the evidence that a runtime-conformance standard needs to exist.

Relationship to the declarative Properties

The Laws are the **prohibitive inversion** of a set of declarative governance properties — the same architecture in two voices, one property per law, no gap:

Property (declarative)	Law (prohibitive)
Measurement layer that survives the system	I. A system shall not measure itself
Dispatch-time enforcement	II. A system shall not be governed after the fact
Tamper-evident attestation	III. A system shall not witness itself
Federated authority	IV. A system shall not certify itself
Fail-closed off-switch (revocable from outside)	V. A system shall not outlive its off-switch

---

The Five Laws of AI Governance · Pharns Genece · 2026 · Version 1.0