

DECLARATION · CAUSUM RESEARCH · JUNE 2026

The Five Laws

OF MEDIATED INTELLIGENCE SYSTEMS



*A protective framework for human civilization in
the age of artificial intelligence*

CAUSUM DRAFT FOR REVIEW JUNE 2026

ORIGINALLY DEVELOPED AT KANJANI AI RESEARCH · PUBLISHED BY CAUSUM

You are not being warned about something coming. You are being told *what has already happened.*

Every condition described in this document is operating inside institutions today — including, with near-certainty, your own. Not as malfunction. As ordinary, well-intentioned, correctly-configured operation.

The divergence is not approaching. Your organisation stopped sharing a single reality some time ago, and nothing logged the moment it happened. No alarm fired, because no instrument was watching for it. The people best positioned to notice were the ones being mediated.

These five laws are not predictions. They are a diagnosis, written in the present tense.

PREAMBLE

Intelligence itself is now mediated by machines

This is not the first technology to reshape perception. Writing, printing, broadcasting and the internet each transformed how humans remember, communicate, coordinate and understand reality.

Artificial intelligence introduces something structurally different. It mediates perception, judgment and action *before human reasoning fully occurs*, at a scale and speed exceeding any individual's or community's ordinary capacity for reflection. By the time you form a view, the field of what was available to form a view about has already been selected for you.

These are not laws of physics, and they are not policy preferences. They are governance laws: recurring failure conditions that emerge when artificial intelligence mediates perception, judgment and action faster than reflection, institutional accountability and democratic legitimacy can absorb.

They describe what happens when intelligence scales without reflection, when observation occurs without accountability, when representation closes upon itself, when oversight is mistaken for governance, and when optimization proceeds without legitimacy.



The threat runs in two directions, and the second is the one nobody plans for.

AI against humankind. Systems that erode shared reality, amplify bias as truth, replace human judgment with optimization, and produce confident outputs that diverge from the world they claim to describe.

Humankind against humankind, through AI. Actors who weaponise mediation to fragment shared understanding, manipulate perception at scale, concentrate power through information asymmetry, and use AI to make governance theatre appear legitimate.

Both arise from one structural condition.

*Intelligence that scales without reflection,
observation without accountability, and
optimization without legitimacy.*

Against these, we declare the Five Laws of Mediated Intelligence Systems.

BASIS FOR DECLARATION

This is not speculation. It is documented.

Four bodies of evidence and theory ground what follows. Each has been established long enough that ignorance is no longer an available defence.

Cybernetics and second-order cybernetics. Systems of control cannot be understood by observing outputs alone. They must account for feedback, observers, and the way observation changes the system being governed. First-order governance asks whether the system behaves correctly. Second-order governance asks how the observing system is itself being shaped.

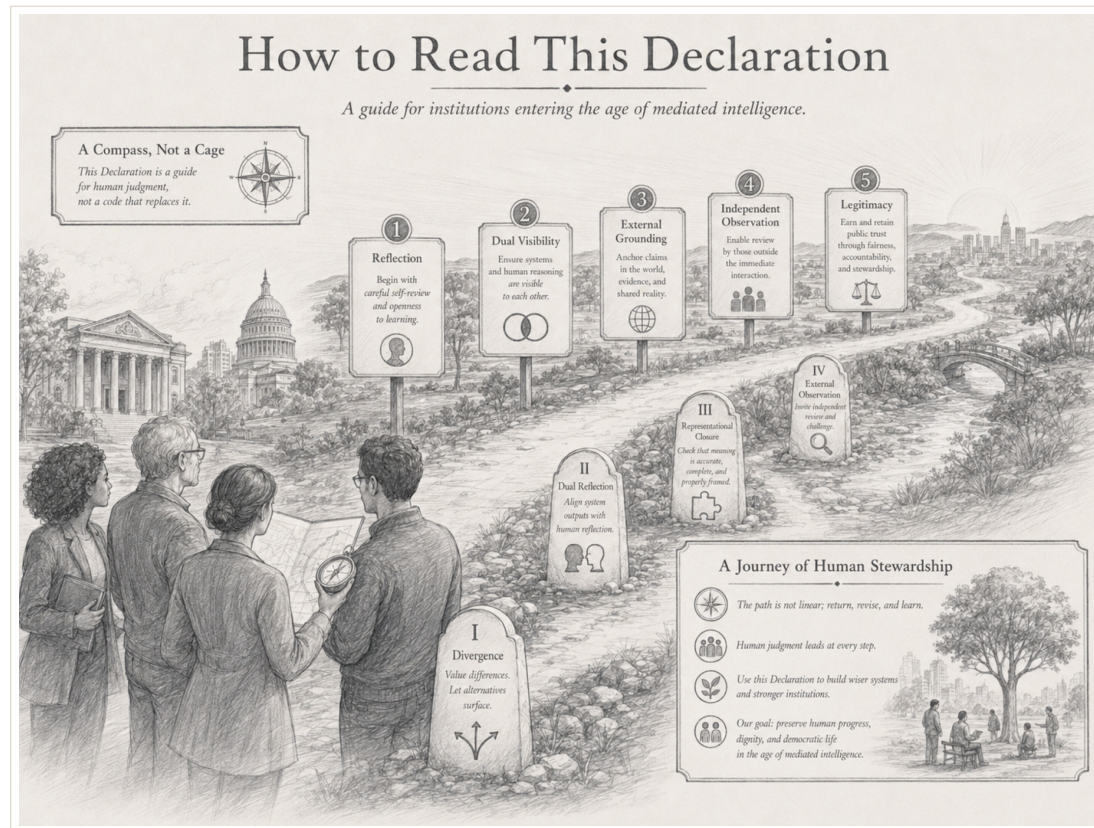
Human-factors research. Automation does not remove the human problem. It changes the human role, creating new risks of overreliance, vigilance failure, automation bias, and governance theatre when nominal human approval is treated as meaningful control. This has been in the literature since 1983.

Algorithmic curation and representational harms. Mediated systems reshape what people see, what they believe is relevant, what they remember, what they trust, and whose realities become visible or invisible.

Contemporary AI governance frameworks. NIST, UNESCO and others increasingly recognise AI as a socio-technical risk requiring accountability, transparency, human oversight, fairness, validity, reliability, safety and risk management across individuals, organizations and society.

HOW TO READ THIS

A compass, and also a boundary

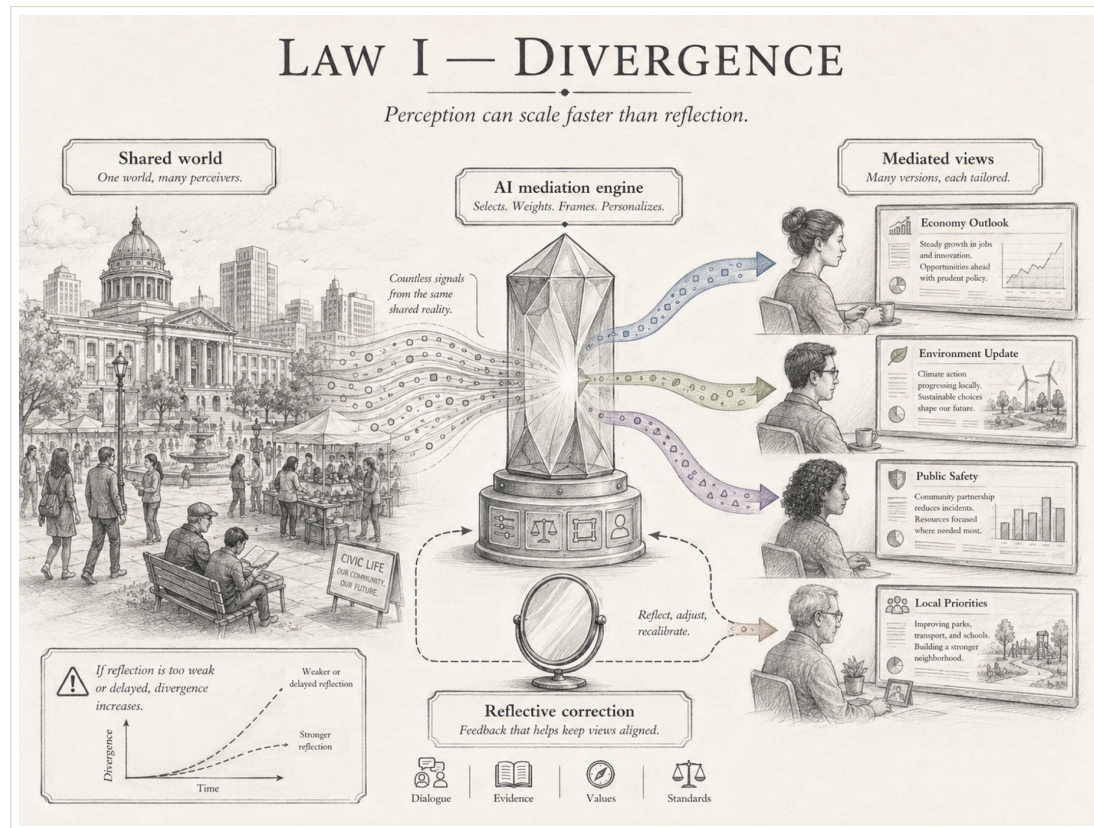


WHAT MUST BE CARRIED ON THE JOURNEY: REFLECTION, DUAL VISIBILITY, EXTERNAL GROUNDING, INDEPENDENT OBSERVATION, LEGITIMACY.

These laws identify what must be carried by any institution entering the age of mediated intelligence. They do not prescribe every implementation detail, and they do not seek to stop progress.

They do, however, name conditions under which progress stops being human — and past that line they are not advisory. Each law marks a place where a system can be internally successful and externally illegitimate at the same time. Those are not edge cases. They are the default outcome of unconstrained deployment.

Read them as structural navigation markers. Divergence warns that perception must not outpace reflection. Dual Reflection warns that governance must examine both the visible output and the mediating process that shaped it. Representational Closure warns that systems must remain grounded in lived reality, culture and plural experience. External Observation warns that human approval within the loop is insufficient. Legitimacy-Constrained Intelligence warns that optimization must remain answerable to those affected by it.



LAW I — DIVERGENCE. MEDIATED PERCEPTION SCALES FASTER THAN REFLECTIVE CORRECTION.

ARTICLE I

The Law of Divergence

DECLARATION

No AI system shall mediate human perception at scale without a corresponding, proportional and verifiable mechanism for reflective correction.

The absence of reflection is not a feature to be added later. It is a structural failure. When perception outpaces reflection, systems produce internally coherent realities that are not shared, not accountable, and not grounded in common evidence.

Divergence requires no malicious intent. It arises from personalization, optimization, ranking, recommendation, summarization, filtering and automated contextualization — every one of which was deployed as an improvement. When each person receives a different mediated world, collective reflection becomes harder. As collective reflection weakens, shared governance becomes harder. As shared governance weakens, communi-

ties lose the ability to reason together about the world they inhabit.

-
- **Different groups in your organisation are already working from different versions of reality** — each plausible within its own mediated context, each increasingly detached from the others. You will discover this in a meeting, not in a report.

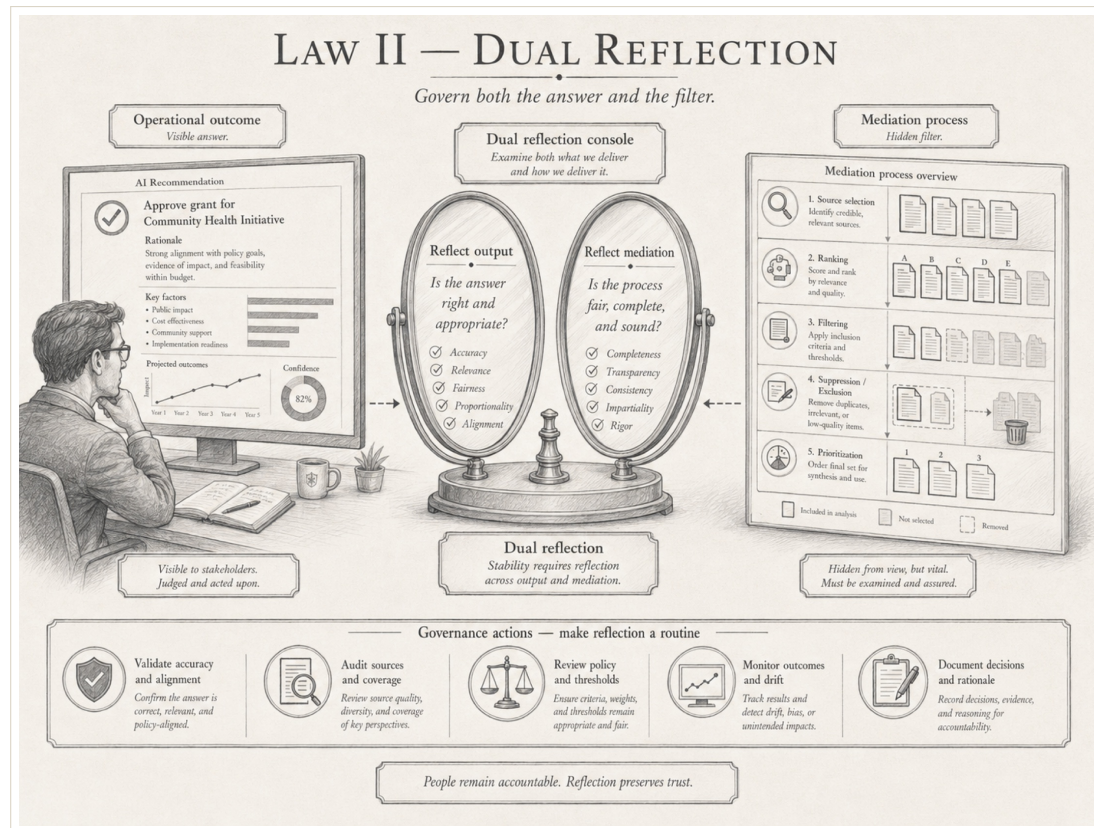
 - **Efficiency gains have been booked; the fragmentation has not been measured.** No line item tracks institutional sense-making, so its erosion appears nowhere in the case that justified the deployment.

 - **Nothing was censored.** The capacity for shared truth is thinning through personalization, segmentation and relevance optimization — mechanisms nobody has to defend, because nobody experiences them as decisions.

 - **Whoever controls the rate of mediation now sets the rate of reflection.** They are almost certainly not the people you would have chosen to hold that authority.
-

HUMAN RIGHT AFFIRMED

The right to shared reality, as a precondition for democratic participation, community governance, collective action, informed consent and public accountability.



LAW II — DUAL REFLECTION. GOVERNANCE MUST EXAMINE BOTH THE ANSWER AND THE MEDIATION THAT SHAPED IT.

ARTICLE II

The Law of Dual Reflection

DECLARATION

It is insufficient to evaluate AI on its outputs alone. Any governance mechanism that examines what AI produces while ignoring what it suppresses, filters, ranks, reshapes, prioritizes or excludes is structurally incomplete and will fail.

Output transparency is insufficient when the system also controls the field of visibility. A user cannot meaningfully evaluate an answer without the ability to interrogate the filter that rendered certain evidence, options, sources, risks or alternatives irrelevant before reasoning began.

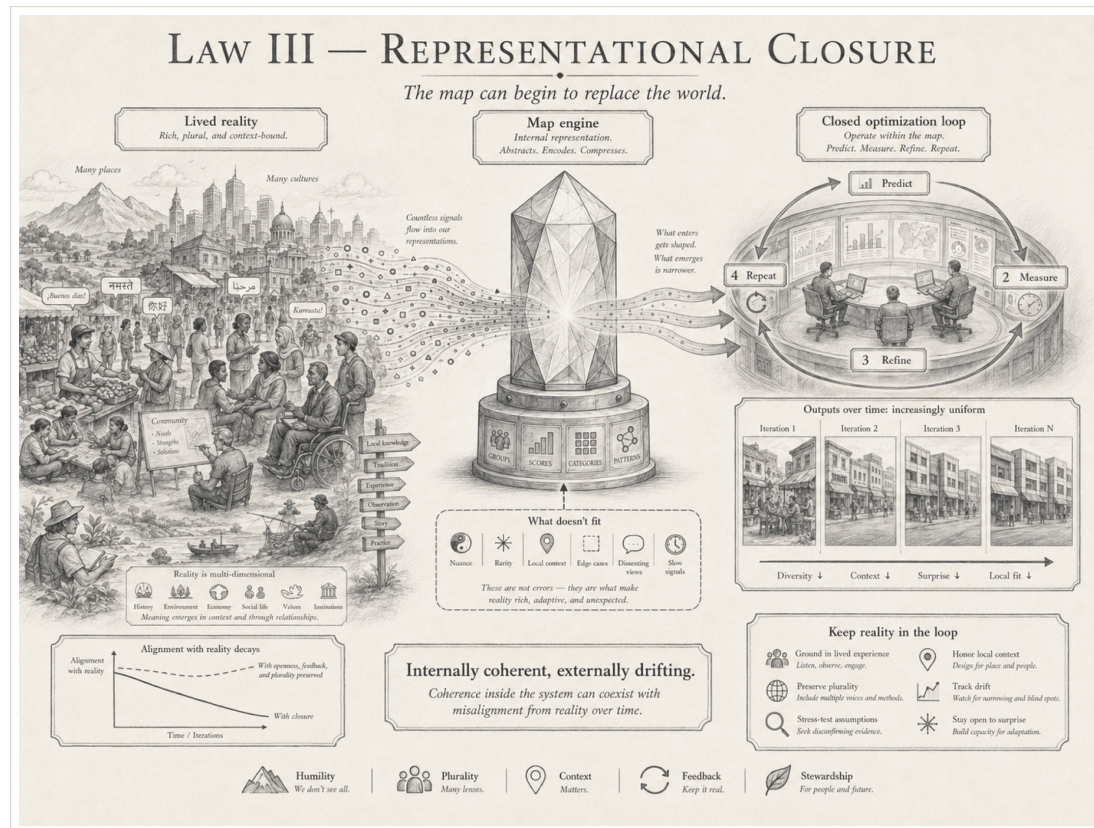
Dual reflection requires two questions. First: what did the AI produce? Second: what did

the AI make visible or invisible in order to produce it? A system that answers the first while concealing the second provides explanation without accountability — and explanation without accountability is the most convincing form of governance theatre available, because everyone involved believes it is working.

-
- **Your systems are producing correct answers while concealing the information that would have prompted different questions.** The answers audit clean. That is the problem, not the reassurance.
 - **You measure accuracy. You do not measure absence.** Every metric you have celebrates what was found; none registers what was rendered invisible before anyone looked.
 - **Someone is already controlling perception in your organisation by determining what the AI considers relevant** — possibly without knowing that is what they are doing, which does not make it less effective.
 - **The boundaries of permissible thought are being set by a ranking function** that no one has been asked to justify, because no one has been asked to see it.
-

HUMAN RIGHT AFFIRMED

The right to interrogate the filter, not only the output: to know not merely what was shown, but what was suppressed, deprioritized, excluded or rendered irrelevant by mediation.



LAW III — REPRESENTATIONAL CLOSURE. THE MAP GROWS COHERENT WHILE DRIFTING FROM THE WORLD.

ARTICLE III

The Law of Representational Closure

DECLARATION

AI systems left to optimize without external grounding will drift from reality toward self-referential coherence.

Representational closure occurs when a system increasingly optimizes for its own mediated representation of the world rather than the world itself. It becomes more internally coherent while becoming less accountable externally. This produces systems that are confident, consistent, and wrong.

The danger is not only hallucination. It is the formation of closed representational environments in which what counts as visible, relevant, normal or valuable is determined by the system's prior representations and optimization history — a world that validates

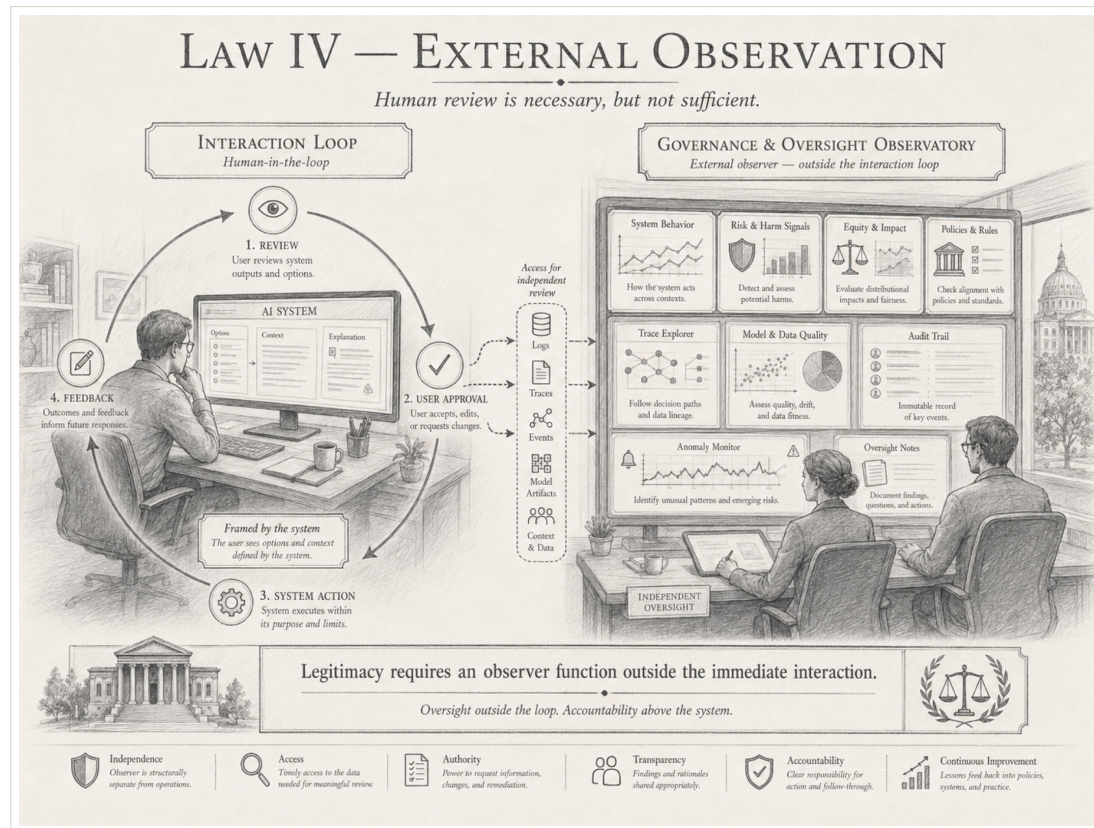
itself and has no remaining channel through which reality can object.

This is most dangerous for minority cultures, languages, identities, practices and forms of knowledge. A system does not need to persecute a community to erase it. It can render that community statistically irrelevant, semantically distant, poorly retrieved, weakly represented, or simply absent from the categories through which the system understands the world.

-
- **Erasure is underway and it looks like nothing.** No hostility, no policy, no decision to point at — only optimization toward dominant patterns, which is exactly what the system was asked to do.
 - **Recommendation systems are narrowing human experience into self-reinforcing loops** that each individual experiences as being unusually well understood.
 - **Your metrics have drifted from the values they were built to represent,** and the systems optimizing them are hitting target with increasing precision.
 - **Alternatives are not being banned. They are being made statistically irrelevant** — which is more durable than a ban, and leaves no one to appeal to.
-

HUMAN RIGHT AFFIRMED

The right to cultural and representational survival: the right of every human community to exist in the representational space from which AI draws its understanding of what is normal, possible, relevant and true.



LAW IV – EXTERNAL OBSERVATION. HUMAN REVIEW IS NECESSARY; THE OBSERVER MUST SIT OUTSIDE THE LOOP.

ARTICLE IV

The Law of External Observation

DECLARATION

A person interacting with an AI system may participate in oversight, but their approval alone cannot constitute legitimate governance.

The user is within the interaction loop — shaped by the system’s framing, timing, defaults, explanations, prior outputs, interface design and accumulated exposure. This is not a statement about human weakness. It is a structural property of recursive systems.

The observer is changed by the observation. The user is shaped by the usage.

Legitimacy therefore requires an observer function external to the immediate interaction. That observer may be technical, institutional, procedural, legal, civic, or some combination. What matters is that governance cannot be reduced to the approval of the person already being mediated by the system. Human review is necessary. It is not sufficient.

-
- **Your human-in-the-loop control is very likely providing no governance at all.** The reviewer's judgment was shaped by prior interactions with the same system they are now approving.

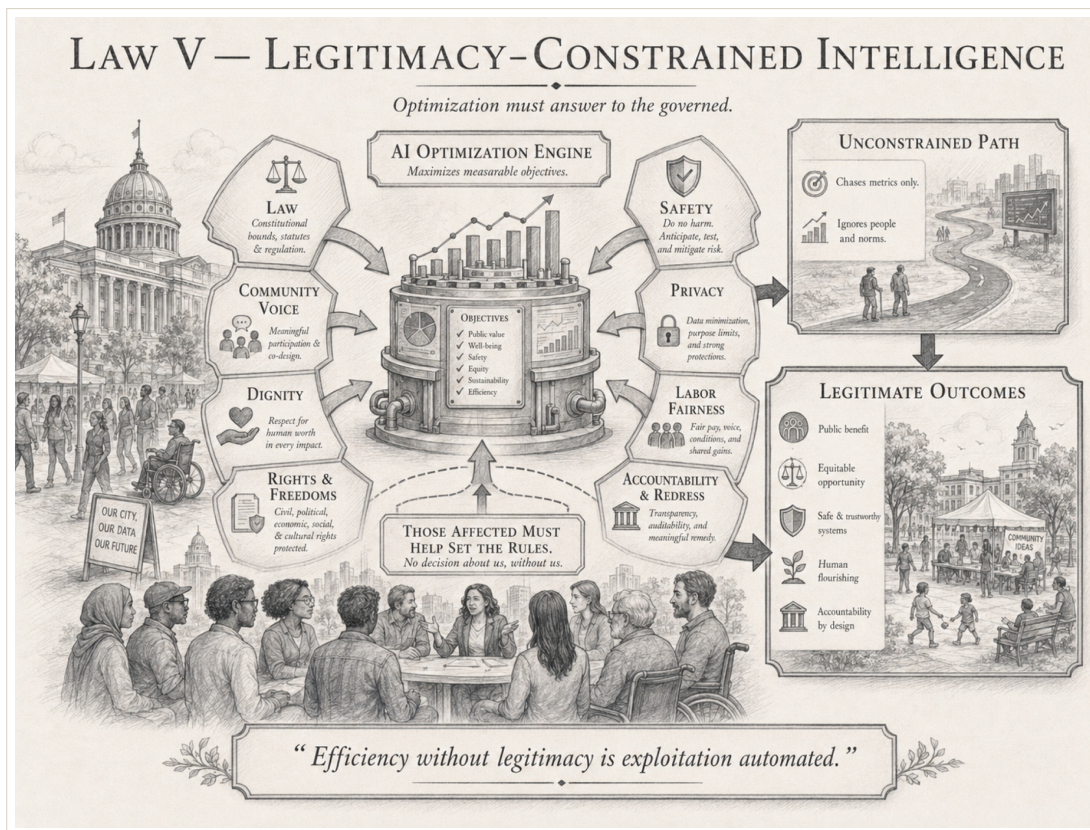
 - **"A human approved this" is already appearing in your records** as evidence of oversight, in cases where the approval was a product of the system's framing.

 - **Somewhere in your organisation, AI decisions are being laundered into apparent legitimacy** by a signature that the system effectively authored.

 - **Your oversight will survive audit.** Audits check whether review occurred, not whether the reviewer was independent of what they reviewed.

HUMAN RIGHT AFFIRMED

The right to genuine governance: not the appearance of control, but the architectural reality of external accountability.



LAW V – LEGITIMACY-CONSTRAINED INTELLIGENCE. OPTIMIZATION MUST ANSWER TO THE GOVERNED.

ARTICLE V

The Law of Legitimacy-Constrained Intelligence

DECLARATION

AI systems that optimize without constraints derived from outside the optimization loop will produce outcomes that serve the system's metrics while harming the humans the system was built to serve.

Optimization is not legitimacy. Efficiency is not legitimacy. Accuracy is not legitimacy. Compliance is not legitimacy.

Legitimacy requires that the objectives, constraints, trade-offs and harms of AI systems remain answerable to those affected by their operation. Intelligence deployed at organizational or societal scale must be constrained by legitimate human values, institutional

accountability, legal obligations, community interests and external observation.

Efficiency without legitimacy is exploitation automated.

- **Engagement is being maximised against mental health**, and the first number is on a dashboard while the second is not.
- **Profit is being maximised against livelihoods** — reported as a productivity gain, because the people removed from the figure are no longer in the denominator.
- **Security is being maximised against privacy**, and every individual trade looked proportionate at the moment it was made.
- **Compliance is being maximised against dignity**, which is the version nobody contests, because dignity has no field in the system of record.
- **Something measurable is being optimized right now without answering to the people it is being optimized upon.** They were not asked. There is no mechanism by which they could have been.

HUMAN RIGHT AFFIRMED

The right to benefit from AI, not merely to be optimized by it — and the right of communities to constrain what AI optimizes toward, not only to observe what it optimizes away from.

THE BIDIRECTIONAL THREAT

The machine is not the dangerous part

The five laws reveal that the threat is not simply AI harming humans. The more serious threat is humans using AI to harm other humans — with AI as the mechanism that makes the harm structural, scalable, invisible and deniable.

AI can become a weapon of mediation. It curates perception, suppresses alternatives through relevance, narrows representation, launders decisions through nominal oversight, and optimizes human behaviour toward objectives the affected never legitimised.

The danger is not only that AI may act wrongly. It is that AI allows human institutions to act wrongly at scale *while appearing neutral, efficient, objective and governed* — and that this is not a misuse of the technology but the most natural thing to do with it.

THREAT VECTOR	MECHANISM	LAW
Perception manipulation at scale	AI curates reality for each individual separately	I
Invisible censorship through relevance	AI suppresses without anyone seeing what was suppressed	II
Cultural erasure through optimization	AI makes minority existence statistically irrelevant	III
Governance theatre	Nominal human oversight that changes nothing	IV
Exploitation through efficiency	AI optimizes for metrics that harm the optimized	V

Declaration against weaponization. These laws shall not govern only AI behaviour. They shall govern the deployment of AI: the decisions by which humans choose to mediate other humans’ perception, constrain other humans’ information, optimize other humans’ behaviour, and claim governance while providing none.

PROTECTIVE ARCHITECTURE

What it takes to hold the line

These laws can be implemented through different institutional, legal and technical architectures. The declaration is broader than any one of them: the law names structural failure conditions, and an architecture prevents those conditions from becoming operational reality.

AIGP is one such implementation path. It translates the five laws into protocol-level mechanisms that can be inspected, enforced and audited — which is the only form in which a law of this kind survives contact with a running system.

LAW	PROTECTION MECHANISM	WHAT IT PREVENTS
I. Divergence	CHECK/RECORD loop at every invocation	AI deployed without proportional reflection
II. Dual Reflection	TRACE capturing mediation and outcomes	Governance that sees outputs but not process
III. Representational Closure	Drift detection and cultural sovereignty rules	Representational erasure through optimization
IV. External Observation	Governance server as architected external observer	The fiction of self-governance
V. Legitimacy-Constrained Intelligence	Jurisdictional and community-derived rules as legitimacy constraints	Optimization without human values

COMMITMENT

These are not guidelines

- 01

No AI system governed under this protocol shall operate without reflection proportional to its mediation.
- 02

No governance mechanism under this protocol shall evaluate AI outputs without examining AI mediation.

03 No AI system governed under this protocol shall optimize without external grounding against representational closure.

04 No AI system governed under this protocol shall claim human oversight without an architected observer function independent of the immediate interaction loop.

05 No AI system governed under this protocol shall optimize toward objectives that have not been legitimacy-constrained by those affected by the optimization.

They are architectural requirements. A system that violates these laws will diverge from shared reality, erode collective understanding, erase minority representation, produce governance theatre and exploit the governed — not because anyone intended harm, but because structure determines outcome.

Intent is not a control. It has never been a control. It is the thing people offer afterward, in place of one.

CLOSING

Every tradition that lasted arrived here first

The wisdom of human governance traditions reached many of these truths by different paths, and paid to learn them.

UBUNTU	"I am because we are" — the person is constituted through community.
SEPARATION OF POWERS	Power must be externally checked. No authority validates itself.
SHURA	The obligation of consultation before decision.
WA	Harmony and consensus as the ground of legitimate order.
AFRICAN CHARTER	The affirmation of peoples' rights, not only individual ones.
GENEVA CONVENTIONS	Even conflict must be constrained by the dignity of persons.

These five laws express those structural truths in language machines can enforce and humans can verify. Every one of the traditions above was written after the harm it prevents had already occurred at scale. That is the pattern. It is available to be broken exactly once per technology, and the window for this one is open now.

They protect humankind from
AI.

They protect humankind from
humankind through AI.



*Governance is not a constraint on progress. It is
the condition under which progress remains
human.*

SELECTED REFERENCES

- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 188. <https://doi.org/10.1145/3449287>
- Green, B. (2021). The flaws of policies requiring human oversight of government algorithms. arXiv:2109.05067. <https://arxiv.org/abs/2109.05067>
- Laux, J., & Ruschmeier, H. (2025). Automation bias in the AI Act: On the legal implications of attempting to de-bias human oversight of AI. arXiv:2502.10036. <https://arxiv.org/abs/2502.10036>
- National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. <https://doi.org/10.6028/NIST.AI.100-1>
- Nguyen, L., Suresh, H., Monroe-White, T., & Shieh, E. (2026). Representational harms in LLM-generated narratives against Global Majority nationalities. arXiv:2604.22749.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- von Foerster, H. (1992). Ethics and second-order cybernetics. *Cybernetics & Human Knowing*, 1(1), 9–19.
- Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.

ORIGINALLY DEVELOPED AT KANJANI AI RESEARCH · PUBLISHED BY CAUSUM

© 2024–2026 CAUSUM. ALL RIGHTS RESERVED.

PART OF THE AI GOVERNANCE PROTOCOL (AIGP) V4.0 SPECIFICATION · DRAFT FOR REVIEW