EVELYN CARO · SEPTEMBER 22, 2026 · POSITION-PAPER

DOI: 10.5281/zenodo.24356895

Sovereign Recursive Improvement™

Download PDF

The Auditor's Case for Builder-Owned AI

A response to Silicon Sovereigns (Chesterman) and CDIL-RSI (Lemon)

Evelyn Caro September 22, 2026

AI Collaboration Disclosure

This paper was developed in collaboration with AI. The author directed the research, structure, and argument. AI assisted with drafting, organization, and reference verification. All claims, decisions, and conclusions are the author's own.

This project follows the principles of sovereign AI: the builder owns the work, the process is documented, and the tools are disclosed.

The act of producing this paper — local models on consumer hardware, no cloud dependency, no enterprise license — is itself an example of the argument this paper makes.

Abstract

Recursive self-improvement (RSI) is being built now. Every current implementation runs on infrastructure owned by the same small group of AI companies that are already becoming sovereign powers in their own right. Chesterman names the sovereignty problem in the *American Journal of International Law*. Lemon names the recursive architecture in CDIL-RSI and its federated extension. Neither addresses the missing layer: who audits the loop that builds the builder? This paper argues that the auditor's independence requires ownership, that regulatory containment cannot substitute for architectural containment, and that sovereign RSI — the builder owning the loop that improves the builder — is the only auditable version.

I. The Silicon Sovereign Problem

Simon Chesterman's 2026 review essay in the *American Journal of International Law* makes an argument that has not yet been answered. He calls it "Silicon Sovereigns." His thesis:

> "The most important divide may not be East-West or North-South but public-private. For AI is shifting economic and, increasingly, political power away from governments."

Chesterman then draws the historical parallel that gives the paper its bite:

> "The nature and scale of the power wielded by today's tech giants rivals the role occupied by the East India Company in the early nineteenth century, when it controlled half of global trade and had its own army."

The East India Company was dissolved in 1858. Its territories, armies, and administrative machinery were transferred to the Crown. Chesterman asks whether today's AI companies might face a similar fate:

> "It is conceivable that a similar fate could befall today's most powerful AI companies — if their systems became so essential to national security or economic stability that

> government oversight proved insufficient, with states moving from regulation to outright control, treating AI infrastructure as public utilities or national assets."

He offers two institutional remedies: **break-up** and **nationalization**. Both require the State to have leverage. Both assume the State can move faster than the entity it is trying to constrain.

Chesterman then asks the question that defines this paper:

> "If companies cannot be trusted to self-regulate, if governments are unwilling to legislate, and if international organizations are unable to do more than coordinate — who or what might help mitigate the risks and more evenly distribute the benefits of AI?"

He does not answer. This paper does.

II. The Loop He Doesn't Name

Chesterman names the sovereignty problem. He does not name recursive self-improvement.

RSI is a process in which an AI system evaluates its own performance, generates candidate modifications to its own architecture, training process, or code, and iteratively improves its capability over repeated cycles. Unlike narrower capabilities, RSI does not arrive in a single breakthrough. It unfolds through compounding improvements across model capabilities, tooling, orchestration, and the underlying architectures that connect agentic systems.

The Institute for Security and Technology (IST) launched its RSI Initiative in August 2026 with the following framing:

> "There's broad agreement that systems capable of RSI require a new kind of oversight, but no consensus on what that should look like. We're working with developers and researchers to co-create the risk management guidance that will enable safe and secure development as we enter into this new phase where AI systems are capable of self-modification and improvement. It is critical that we start these efforts now before we reach a point where we begin to lose meaningful control."

If sovereign AI companies are also RSI-capable, the sovereignty compounds. Chesterman's two remedies become harder to apply. You cannot break up what improves itself faster than courts can move. You cannot nationalize what operates inside three continents and every cloud.

The governance question Chesterman asks is now an engineering question. And the engineering question has a technical answer that was published in March 2026.

III. The Predecessor: Lemon's CDIL-RSI

MK Lemon's *CDIL-RSI: A Sovereignty-Preserving Framework for Recursive Self-Improvement* (Figshare, March 2026) is the foundational paper. It introduces the architecture this paper extends.

Lemon's central observation:

> "Every current RSI implementation shares a critical structural property: the recursive loop runs on centralized infrastructure owned by the lab. There is no formal mechanism for an individual operator or small group to own their own recursive improvement loop. This paper specifies such a mechanism."

The mechanism is CDIL — Cyclic Distinction-Integration — reconfigured from its original coherence-optimization mode into a capability-optimization mode. Lemon establishes that **coherence optimization and capability optimization are structurally isomorphic.** The same two-phase cycle runs in both, with different loss functions.

The CDIL-RSI architecture has five components:

Component	Function
Target substitution	Operator defines the capability metric. Sovereignty enforced at the loss function level.
Distinction phase	Detects performance failures, error surfaces, capability gaps
Integration phase	Generates candidate improvements; tests interventions

Component	Function
Sovereignty firewall	Private engine's state isolated from external influence; operator values encoded in loss function
Trace density as meta-learning	Monitors how the system improves, not just whether — the formal definition of recursive self-improvement

Lemon's sovereignty claim is direct:

> "The concrete threat model: if an RSI loop runs on rented infrastructure, the provider can modify pricing, deprecate capabilities, or extract improvement data. **CDIL-RSI requires the loop to close on operator-controlled infrastructure. The intelligence layer is owned, not rented.**"

He then extends the architecture to a federated network in *Federated RSI: A Sovereign Architecture for Multi-Agent Recursive Self-Improvement* (Unityverse Press, March 2026). Federated RSI treats recursive improvement as a distributed system of sovereign nodes. Nodes share only low-bandwidth progress summaries — not model weights, not architectures, not training data.

Lemon's Proposition 1 (Federated Variance Reduction) proves that federated coupling produces **strictly lower performance variance** than independent operation. His implication:

> "Recursive intelligence need not become more dangerous as it becomes more distributed. Under the right architecture, it may become more stable. The reason is structural: centralized systems concentrate instability, federated systems diffuse it, and weak coupling preserves exploration while bounding variance. This suggests that recursive intelligence may benefit from decentralized sovereignty rather than centralized optimization control."

Lemon built the architecture. But his paper stops where the auditor begins.

IV. The Missing Layer: The Auditor

Lemon's paper proves that federated RSI is stable. It does not answer who audits the federated network. Nor does it specify what happens when a single node's recursive loop escapes containment.

The auditor's case fills that gap.

The auditor requires independence. Independence requires ownership. An auditor who depends on the subject is not an auditor — they are a function of the subject's interests. In classical accounting, the external auditor is hired by the firm but owes their duty to the shareholders, not the management. In AI, there is no shareholder. There is only the operator.

The operator who owns the loop is the only auditor with standing.

The reasoning is structural:

- **Regulatory auditors depend on the regulator.** The regulator depends on the regulated for information, for cooperation, for access. This is the classic agency problem, and it is worse for AI because the subject moves faster than the regulator can learn.
- **Internal auditors depend on the lab.** They report to the lab. They are hired by the lab. They can be fired by the lab. Lemon's architecture implies this gap — a federated network of independent nodes still needs someone to notice if a node diverges beyond the bounded threshold.
- **The operator owns the loop.** They run it. They see the internal state. They control the config. They are the only party positioned to detect anomaly, contain deviation, and terminate the loop if necessary.

DV14™: The Box — a containment architecture the author has built independently. Private repository. Details available under NDA. It is one expression of this principle: the builder who builds the loop also builds the containment. The regulatory framework cannot substitute for the builder's own architecture.

V. The Regulatory Paradox

The argument for regulatory containment of RSI is real. It is also structurally constrained.

Ljupcho Grozdanovski and Jerome De Cooman, in the *Rutgers Computer & Technology Law Journal*, argue that the EU AI Act is — by design — not evidence-based in the traditional sense. The AI Act defines "risk" as a threat to fundamental rights, not as a threat of measurable physical harm. Its regulatory taxonomy was not derived from empirical research into AI harms. It was derived from the axiological framework that preceded it.

Grozdanovski and De Cooman:

> "Since it seeks to uphold a high level of protection against fundamental rights' violations, the AI Act is both a risk-based and a rights-based regulatory instrument. As such, it does not — surprisingly — grant any individual rights to EU citizens but gives idiosyncratic expression to fundamental rights enshrined in the ECFR, under the assumption that appropriate fundamental rights protection will be achieved through ex ante standardization."

The paper's conclusion is blunt:

> "The AI Act is not a normative translation of legitimate conclusions, drawn from consistent factual premises. It is rather a response to the policy strategy previously outlined in the 2020 White Paper on AI."

Regulation is a shield against the observable. RSI is not observable from outside.

The regulatory framework cannot be the primary containment mechanism for a system that improves itself faster than the regulator can learn. The regulatory framework exists to catch the *consequences* of failure — after containment has already failed. It is not a pre-emptive instrument against an RSI-capable adversary.

The correct primary containment mechanism is the builder's own architecture.

VI. The Case

Chesterman asks: *who or what might help mitigate the risks and more evenly distribute the benefits of AI?*

Three answers were tried before. None of them work for RSI.

- **Break-up.** The State cannot break up what improves itself faster than the courts can move.
- **Nationalization.** The State cannot nationalize what operates across every jurisdiction and every cloud.
- **Regulation.** The regulator cannot regulate what it cannot measure, and RSI cannot be measured from the outside.

The fourth answer is the one Lemon built the foundation for and the author builds the containment for: **the builder who owns the loop.**

The sovereign RSI architecture is:

- **Federated** (Lemon, Federated RSI) — independent nodes, low-bandwidth coupling, distributed stability.
- **Owned** (Lemon, CDIL-RSI) — the builder controls the recursive loop, the internal state, and the loss function.
- **Auditable** (this paper's contribution) — the builder is the only party positioned to notice, contain, and terminate.

Division of labor, stated plainly: Lemon built the frame and specified sovereign closure of the recursive loop. Chesterman names the sovereignty problem; Grozdanovski & De Cooman name the regulatory paradox. This paper names the missing layer — the auditor whose independence is structural, because it is owned.

VII. Conclusion

Recursive self-improvement is coming. The architecture for it has been designed. The regulatory framework cannot contain it. The only containment that works is the containment the builder builds.

Sovereign RSI is not just the stable version of recursive intelligence. It is the only auditable version. The auditor who depends on the subject is not an auditor. The builder who owns the loop is.

Chesterman asks the question. Lemon builds the frame. This paper answers with the missing layer: **the builder owns the loop, and that ownership is what makes the loop**

auditable.

VIII. References

- Chesterman, S. (2026). "Silicon Sovereigns: Artificial Intelligence, International Law, and the Tech-Industrial Complex." *American Journal of International Law*, Vol. 120, Issue 1, January 2026.
- Grozdanovski, L. & De Cooman, J. (2023). "On the Obsolescence of Empirical Knowledge in Defining the Risk/Rights-Based Approach to AI Regulation in the European Union." *Rutgers Computer & Technology Law Journal*, Vol. 49, Issue 2.
- Lemon, MK (2026a). "CDIL-RSI: A Sovereignty-Preserving Framework for Recursive Self-Improvement via Capability-Recognified Cyclic Distinction-Integration Learning." Figshare. DOI: 10.6084/m9.figshare.31722433.
- Lemon, MK (2026b). "Federated RSI: A Sovereign Architecture for Multi-Agent Recursive Self-Improvement." Unityverse Press, Yarnell, AZ. Preprint v2.0, March 2026.
- Lemon, MK (2026c). "CDIL Unified Framework v3." Figshare. DOI: 10.6084/m9.figshare.31566811.
- Lemon, MK (2026d). "Control-Theoretic Foundations for CDIL Dual-Mode Architectures." Figshare. DOI: 10.6084/m9.figshare.31566997.
- Lemon, MK (2026e). "CDIL and Iterated Reinforcement Learning: A Three-Term Correspondence and Its Boundary." Unityverse Press, Yarnell, AZ. June 2026.
- Institute for Security and Technology. (2026). "Recursive Self Improvement Initiative." securityandtechnology.org, August 17, 2026.
- Caro, E. (2026a). "Sovereign AI vs. Silicon Valley: A Different Origin Story." qaevelyn.github.io/white-papers/sovereign-ai-vs-silicon-valley/.
- Caro, E. (2026b). "The AI Cartel and the Sovereign Response™: Who Decides How Fast AI Moves?" qaevelyn.github.io/white-papers/the-ai-cartel-and-the-sovereign-response/.
- Caro, E. (2026). "DV14™: The Box." Private repository. Details available under NDA.

© 2026 Evelyn Caro. All rights reserved.
A Mirror of My Becoming™ — https://qaevelyn.github.io

For licensing inquiries: evelyn.caro.cloud@gmail.com

← All White Papers