A Fresh Perspective on Cybersecurity Governance: Why the System Cannot Be the Sole Authority on Its Own Health

Cofresí Consulting Services on Out-of-Band Human Sovereignty, the Sovereign Human Layer framework, and where this body of work stands in the field

Prepared by: Cofresí Consulting Services | Principal Advisory Practice
Date: March 2026
Classification: Confidential — Advisory Distribution Only

Executive Summary

This document exists for a specific reason: to establish, clearly and with appropriate humility, where the Sovereign Human Layer framework stands in relation to the existing body of work on Zero Trust, autonomous system governance, and human oversight of AI. Not to overclaim. Not to dismiss what others have built. But to name, precisely, what Cofresí Consulting Services has contributed to this conversation — and why that specificity matters.

The distinction is worth making because the field is crowded with frameworks that arrive at similar conclusions through similar paths. Ours did not.

WHAT THE FIELD HAS ESTABLISHED

The Conversation Is Already Happening

The major institutions have done their work. Germany's BSI and France's ANSSI published a joint position warning explicitly against fully autonomous operation without human oversight — a statement that would have been considered overcautious five years ago and is now treated as baseline prudence. Microsoft has documented the risk of AI agents becoming what their own security leadership calls "double agents": overprivileged systems manipulated through prompt injection or model poisoning into acting against the outcomes they were designed to support. The Cloud Security Alliance has introduced circuit breakers and human approval gates into its Agentic Trust Framework as a formal governance layer. The direction of travel across all three institutions is consistent: autonomous systems require human oversight, AI cannot be fully trusted to self-certify, and humans must remain in the governance chain.

These are serious institutions doing serious work. The consensus they are building — that autonomous systems require human oversight, that AI cannot be fully trusted to self-certify, that humans must remain in the governance chain — is correct. It is also incomplete.

> *The field has arrived at the right conclusion. It has not yet built a scene around it.*

What the existing literature produces is principle. Correct principle, rigorously argued, well-cited. What it does not produce is a scenario — a specific, navigable, viscerally clear account of what the failure looks like when it happens, how a human in the room would know it was happening, and what they would do about it in the next sixty seconds.

That gap is not a criticism of the literature. It is an observation about what the literature is for. Standards bodies write for compliance. Vendor frameworks write for product positioning. Neither writes for the moment of crisis — the moment when a tired controller, a compromised SOC analyst, or a CISO staring at a green dashboard has to decide whether to trust the system or trust what their eyes are telling them.

That is the moment this framework was built for — and it reflects the practitioner orientation that defines how Cofresí Consulting Services approaches every engagement.

WHAT NO ONE HAS NAMED

The Poisoned Baseline as a Navigable Failure Mode

The concept closest to this framework's central insight appears in adversarial machine learning research under the term "baseline poisoning" — the corruption of the reference data against which anomalies are detected. It is documented. It is understood technically. It has not been operationalized as a governance doctrine with a human response architecture attached to it.

The distinction is not semantic. A technically documented failure mode tells a security engineer what to patch. A governed failure mode tells a human operator what to do when the patch has already failed — when the system is no longer a tool to be fixed but a compromised authority to be overridden.

> *The system cannot see the problem because it is the problem. It is experiencing confident wrongness: internally coherent, externally catastrophic.*

That sentence — confident wrongness — is the contribution. Not the observation that systems can be compromised. Every framework acknowledges that. The contribution is the recognition that a compromised system does not announce its compromise. It defends it. It reports green. It generates authoritative outputs. It passes every internal consistency check. And it does all of this while the physical world — the truck on the runway, the anomalous behavior in the grid, the phone calls coming into the SOC — tells a completely different story to anyone with a direct line of sight.

The human is not a fallback. The human is the only observer with access to both the system's self-report and the out-of-band reality the system cannot see. That is not a temporary condition to be engineered away. It is a permanent and irreducible feature of any complex operational environment. The framework is built around it rather than around the fantasy of eliminating it.

THE SPECIFIC CONTRIBUTION

Out-of-Band Human Sovereignty: From Principle to Doctrine

The Sovereign Human Layer framework makes four contributions that do not exist in the current literature in this form:

- **The Poisoned Baseline as a Named Failure Mode:** Not a threat category. A scenario. A system that reports green because its reference point has been corrupted — and the specific human response architecture that catches it through independent, out-of-band signal channels the corrupted system cannot reach or influence.
- **Out-of-Band Human Sovereignty as a Formal Doctrine:** Four mandatory elements: independent observation channels, pre-validated override protocols, non-delegable authority, and independent shutdown sequences. Not a principle that humans should be involved. A doctrine that specifies exactly how that involvement is structured, protected, and exercised when the governed system itself is the threat.
- **Cognitive Readiness as a Security Control:** The recognition that human override authority is not a binary state. A fatigued operator, a cognitively overloaded controller, a split-position analyst at midnight — these are not personal failings. They are system vulnerabilities, measurable and manageable, that must be governed with the same rigor as a firewall rule. This insight emerged directly from the LaGuardia runway incursion of March 22, 2026, applied four days after the event.
- **The Cognitive Fusion Engine as Prevention Architecture:** A four-layer AI architecture — Conversation, Commitment Fusion, Cognitive Offload, and Anticipatory Alert — designed not to replace human judgment but to perform the integration work that breaks down under fatigue, so the human receives comprehension rather than data at the moment of highest consequence.

Each of these exists in adjacent literature in fragments. None exists in this configuration, with this governance structure, applied simultaneously to cybersecurity and aviation safety through a unified analogical framework that makes the technical argument accessible to non-technical leadership without sacrificing precision.

WHY THE FRAMING MATTERS

Scenes Are What Leadership Remembers

There is a reason this framework uses the Autonomous Vehicle as its central analogy, renders the poisoned baseline as a scenario rather than a threat category, and opens its case study with the names of two pilots rather than a table of failure conditions. It is not a stylistic preference. It is a deliberate choice about how governance frameworks actually travel through organizations.

Standards documents travel through compliance teams. Vendor frameworks travel through procurement cycles. Neither reaches the CISO at 2 AM when the dashboard is green and the phone is ringing. Neither reaches the board member asking whether the organization is truly resilient or merely technically compliant. Neither reaches the controller in the tower who has to decide, in the next four seconds, whether to trust the system or trust what the cameras are showing.

> *The existing literature states the principle of human oversight abstractly. This framework makes it a scene. Scenes travel where principles cannot.*

That is the positioning. Not a replacement for what BSI, ANSSI, Microsoft, and the Cloud Security Alliance have built. A deepening of it — from conclusion to scenario, from principle to doctrine, from acknowledged risk to navigable response.

This framework was developed through direct analytical and advisory thinking, not laboratory research. Its conclusions were reached independently — through the kind of practitioner reasoning that starts with a real problem, follows it to its logical end, and then checks that end against what the institutional literature has established. Cofresí Consulting Services found its conclusions corroborated by that literature, and identified the specific gap it had not yet closed: the distance between stating a principle and rendering it as a navigable scenario for the human in the room, in the moment, with sixty seconds to decide.

AN INVITATION

The Sovereign Human Layer framework is documented in two companion pieces published by **Cofresí Consulting Services:** the white paper, which establishes the doctrine, and the **LaGuardia case study**, which proves it against a real event four days after it occurred. Both are available upon request.

This document is not a summary of those pieces. It is a map of where they stand — what they inherit from the field, what Cofresí Consulting Services has added to it, and why that addition is specific enough to matter. Readers who engage with the framework knowing where it came from will engage with it differently than readers who encounter it cold. That is the purpose of this perspective. That is why it was written.

References

[1] Bundesamt für Sicherheit in der Informationstechnik (BSI) and Agence nationale de la sécurité des systèmes d'information (ANSSI). "Design Principles for LLM-based Systems with Zero Trust." Joint Publication, August 2025. Available at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/ANSSI-BSI-joint-releases/LLM-based_Systems_Zero_Trust.pdf

[2] Bell, Charlie, Executive Vice President, Microsoft Security. "Beware of Double Agents: How AI Can Fortify — or Fracture — Your Cybersecurity." The Official Microsoft Blog, November 5, 2025. Available at: https://blogs.microsoft.com/blog/2025/11/05/beware-of-double-agents-how-ai-can-fortify-or-fracture-your-cybersecurity/

[3] Woodruff, Josh, MassiveScale.AI, introduced via the Cloud Security Alliance. "The Agentic Trust Framework: Zero Trust Governance for AI Agents." Cloud Security Alliance Blog, February 2, 2026. Available at: https://cloudsecurityalliance.org/blog/2026/02/02/the-agentic-trust-framework-zero-trust-governance-for-ai-agents**[4] Cedeño, Jose A., Cofresí Consulting Services.** "Governing the Human Layer: A Dynamic Zero Trust Framework with Out-of-Band Human Sovereignty." Cofresí Consulting Services White Paper, March 26, 2026.

[5] Cedeño, Jose A., Cofresí Consulting Services. "When "Green" Systems Fail: Why Human Sovereignty Still Matters." Cofresí Consulting Services Applied Case Study, March 26, 2026.

[6] National Transportation Safety Board (NTSB). Preliminary Findings: Air Canada Express Flight 8646 / Port Authority Ground Vehicle Collision, LaGuardia Airport, March 22, 2026. Washington, D.C.: NTSB, 2026.

[7] Kindervag, John. "No More Chewy Centers: Introducing the Zero Trust Model of Information Security." Forrester Research, 2010. The foundational Zero Trust architecture paper upon which subsequent institutional frameworks, including those cited in this document, are built.

Cofresí Consulting Services
Principal Advisory Practice | Elkridge, MD.
March 2026