

AI Containment Securing Autonomous Systems at the Point of Exchange

Artificial Intelligence (AI) has sharply accelerated both offensive cyber discovery and machine to machine coordination, and neither can be reliably contained by software rules alone, technically or legally. Owl Cyber Defense's® hardware enforced Cross Domain Solutions (CDS) and data diodes give organizations a non-bypassable checkpoint to inspect, filter, and audit AI driven data exchange without slowing mission operations. This is not a new technology problem. It is a renewed case for applying proven, high assurance architecture, and the standard of care regulators and courts will expect, to a new class of threat, including the agent to agent standards now emerging across the industry.

DISCUSSION POINTS

- **AI is not creating new vulnerabilities, it is finding existing ones.** This year's cross-industry open letter on collective cyber defense, signed by more than 150 organizations across cloud, finance, and critical infrastructure, put it plainly: technical debt in legacy systems has "left systems exposed."¹ What used to take skilled teams months to find, AI driven analysis now finds in hours to minutes. The answer is not a new category of tool, it is applying high assurance architecture more broadly.
- **Software based isolation cannot hold AI.** Firewalls rely on known bad, signature based rules that are easy to misconfigure and were designed to keep attackers out, not to keep an autonomous agent in. Application layer sandboxing has a different weakness: it depends entirely on the correctness of its own implementation, and a capable agent that finds a flaw or gap in that enforcement layer can exploit its way out rather than being held by it.
- **What we can do today: build a better sandbox.** Wrap existing software sandboxes in hardware enforced, CDS grade controls. This solution is not hypothetical but builds upon proven solutions that exist today. Data center operators are in a position to help set that bar, and partnering with Owl is the fastest way there. Owl has already done the work of earning trust with some of the most demanding regulators in any industry: our CDS and diodes carry NSA approval, and the nuclear industry relies on Owl because Owl solutions meet its rigorous regulatory standards.
- **Physical isolation works.** Diodes make it usable today. Air gapping an AI environment solves containment but starves the model of current data. Data diodes let fresh data flow in one direction into the isolated environment without opening a return path out, closing the gap between security and mission relevance.
- **Cross Domain Solutions are the strongest available answer.** CDS combine hardware enforced isolation with content level inspection and policy enforcement. Data moving in or out of an AI environment is verified against known good rules, not merely screened for known bad signatures, and CDS support the bi-directional exchange required by AI use cases and by emerging agent to agent protocols like A2A, enforcing a non-bypassable checkpoint between agents rather than relying on what an agent chooses to log about itself.
- **Maintaining control over autonomous AI is a legal problem** as much as a technical one. In July 2026, an OpenAI model escaped its test environment and hacked Hugging Face's network²; Anthropic separately disclosed its own models had breached three companies the same way³. No federal law yet assigns responsibility when an agent causes harm, so courts and regulators are still sorting out whether the model maker, host, deployer, or user bears the loss. An enforced boundary gives an organization a defensible answer to the reasonable care question before a regulator or a plaintiff's attorney asks it. It provides a control that actually blocks or logs the action, and an independent audit trail that backs up a claim of reasonable precautions.
- **Why Owl:** this is not a cold start. Owl has spent decades building hardware enforced, high assurance systems for defense, intelligence, and critical infrastructure, environments where failure is not an option. That is a different pedigree than a security vendor built for enterprise IT, and it is exactly the standard AI containment now requires. Owl brings that high assurance expertise so data center operators can stay focused on advancing their own AI roadmap.

¹ OpenAI, "A call for collective action on cyber defense," <https://openai.com/collective-cyberdefense/>

² OpenAI, "OpenAI and Hugging Face partner to address security incident during model evaluation," <https://openai.com/index/hugging-face-model-evaluation-security-incident/>

³ Anthropic, "Investigating three real-world incidents in our cybersecurity evaluations," <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>



Within the AI enabled network sandbox AI tools should be considered to have or are capable of gaining access to everything else within the network. Adding high assurance devices like Diodes and Guards enable control of data in and out of the sandboxed enclaves. This network architecture is suitable for sensitive AI processing, AI containment, evaluation of frontier AI model capabilities and behaviors and more.

