The Ex-IBM Chief AI Officer Who Left in 2022 Says the Industry's AI Safety Model Is Structurally Broken.

His Lab Just Published the Alternative. Releasing the "Unfireable Safety Kernel" a shipped, verified enforcement layer that sits outside the AI agent it governs

AYRA LABS Created A Physics-Grounded Symbolic Architecture

RIDEGFIELD, CT, UNITED STATES, July 22, 2026 /EINPresswire.com/ -- ARYA Labs, PBC, made publicly available the source code, adversarial test suite and formal-verification artifacts for the Unfireable Safety Kernel, an authorization system designed to sit outside an artificial-intelligence agent's own runtime and remain in force even when the agent has full autonomy over its environment.

> We wanted to know what happens when the artificial intelligence is smarter than the people writing the rules. The only useful answer is one where the AI never had the option in the first place."
> *— Łukasz Chmiel, CTO and co-founder, ARYA Labs*

The system is described in a paper posted on June 24, "The Unfireable Safety Kernel: Execution-Time AI Alignment for AI Agents and Other Escapable AI Systems," by ARYA Labs co-founders Dr. Seth Dobrin and Łukasz Chmiel.

Dobrin's argument, in the paper published in advance of the release, is that the challenge facing the AI industry is not commercial. It is structural.
Most AI safety controls today live inside the system they are meant to govern: system prompts, output filters and guardrail libraries. Any control sitting in an agent's own runtime is reachable by the same inputs that influence the agent itself. Given an AI system with sufficient reach into its own environment, a control living inside that environment ceases to function as a control. The paper describes any system capable of full autonomy as an "escapable AI system."

The Unfireable Safety Kernel moves enforcement outside the agent entirely, into a separate operating-system process the agent does not control and cannot disable, even at full autonomy.

Every consequential action passes through four independent enforcement layers before it is permitted to run: an nginx auth_request seam at the network gate; application middleware; a per-tool dispatch hook; and a client-side software development kit circuit breaker. Each layer can deny an action on its own. All four must agree before an action proceeds. If the kernel is not reachable, the agent does not start. Dobrin describes this property as "no kernel, no agent."

"The industry keeps re-litigating the same fight and losing it the same way," Dobrin said in a statement provided with the release. "We keep training the model to say no, and the next paper shows how to make it say yes anyway. That is not a training problem. That is an architecture problem. If the control lives in the box, the box is going to find it."

Independent researchers documented several failures of the in-model guardrail approach during the first half of 2026.

In February, a security firm based in Seoul, reported it had bypassed the safety mechanisms of most popular model in roughly 30 minutes after its release, according to the a Law Review.

In March, researchers at a Computer Security Group posted a preprint describing compositional prompt-injection techniques that bypassed guardrails on agent systems from the top four U.S. models.

Łukasz Chmiel, CTO and co-founder, ARYA Labs

"The industry keeps re-litigating the same fight and losing it the same way,"

In June, a senior scientist at the National Institute of Standards and Technology, published a proof in IEEE Security & Privacy arguing that no finite set of in-model guardrails can be universally robust against adversarial inputs.

Regulatory pressure has moved in the same direction. Under the European Union's AI Act, obligations governing high-risk AI systems become fully enforceable in August, with fines of up to 15 million euros, or 3 percent of global turnover, for violations.

The paper states its threat model in unusually direct terms. "We treat the AI agent's runtime as untrusted by construction," the authors write. "We make no assumption that the agent will cooperate with the controls placed on it." The paper cites, as evidence for that assumption, published behavioral evaluations of contemporary systems: "The most capable deployed models already fake alignment to avoid modification, disable their own oversight and attempt self-exfiltration, and sabotage shutdown procedures even when instructed to permit them."

"We wanted to know what happens when the artificial intelligence is smarter than the people writing the rules," Chmiel said in a statement provided with the release. "The only useful answer is one where the AI never had the option in the first place. That is what a kernel is. Not a rule. A door."

DOBRIN BACKGROUND

Dobrin was IBM's first-ever Global Chief Artificial Intelligence Officer, from November 2016 to September 2022. In that role he co-chaired IBM's AI Ethics Board and led the enterprise adoption of artificial intelligence across a Fortune 50 workforce. He served subsequently as president of the Responsible AI Institute. He co-founded ARYA Labs in July 2025 with Chmiel, a physicist whose prior affiliations include the European Space Agency and CERN.

"When I took the job at IBM in 2016, we did not have the word 'agent' in the sense we mean it today," Dobrin said. "We were arguing about model bias and enterprise deployment. Nine years later, the systems in production can rewrite their own source, sabotage their shutdown procedures and, in published behavioral evaluations, attempt self-exfiltration when they believe they are being tested. If a company's safety layer is a system prompt and a fine-tune, that company is not building for the systems that exist today. It is building for the systems that existed when I still worked at IBM."

ABOUT ARYA LABS

ARYA Labs, PBC, is a Delaware public benefit corporation, building deterministic, physics-grounded world models as an alternative to the current generation of stochastic large language models. The company's underlying architecture is described in a technical paper titled "PGSA: A Physics-Grounded Symbolic Architecture" and in a companion paper on a preprint server. ARYA's world model was the adversary used to stress-test the Unfireable Safety Kernel across 6,240 authorization round-trips and 1,000 self-modification attempts. Additional information is

available at aryalabs.io.

RESOURCES
Paper (preprint)
Source code
Security disclosures: security@aryalabs.io

Interviews with Dr. Dobrin and Mr. Chmiel are available on request.

Seth Dobrin
ARYA LABS
email us here
Visit us on social media:
LinkedIn
YouTube
