



KHAOS SIGNAL // FLASH REPORT — SEPTEMBER 14, 2026

TECHNOLOGY DOES NOT WAIT

The AI race will not be won by moving fastest—or by waiting for government. It will be won by building the systems that let us move safely at speed.

SIOBHAN MACDERMOTT | KHAOS HOLDINGS | INTELLIGENCE FOR COMPLEX SYSTEMS

KEY JUDGMENT

The debate over whether to slow frontier AI is asking the wrong operational question. The more urgent question is what we are building—right now—to contain, observe, interrupt, and recover from AI systems that can already act.

Dario Amodei, Anthropic's CEO, is right that capability is moving faster than our ability to understand and control it. George Kurtz, CrowdStrike's CEO, is right that the rest of the world will not pause while democracies deliberate. These positions are not opposites. They describe two different clocks.

The first is the **frontier clock**: how quickly the most powerful laboratories train and release more capable systems. The second is the **threat clock**: how quickly deployed models, open tools, stolen credentials, distilled capabilities, and autonomous agents are being operationalized by states, criminals, companies, and individuals.

We may be able to pace the first. We cannot assume we can stop the second.

The credible strategy is therefore not acceleration without restraint or restraint without readiness. It is to convert every additional day of pacing into a **security dividend**: faster investment in runtime cyber defense, nonhuman identity, secure infrastructure, independent evaluation, incident disclosure, recovery, and human authority over consequential decisions—including a tested ability to halt AI activity when predefined risk thresholds are crossed.

“We must slow the pace at which we improve the capabilities of AI models. Progress will still seem fast, and we must make wise use of the time we gain.” — Dario Amodei[1]

The operative words are not simply *slow the pace*. They are **use the time**.

TWO CLOCKS, ONE PROBLEM

Amodei’s new proposal, *We Must Pace the Frontier*, is more serious than a generic call to stop AI. He proposes permanent independent evaluators with employee-like access inside frontier laboratories, coordinated standards among companies in democratic countries, and eventually verifiable international agreements. He also argues that any pacing regime must preserve the U.S. and allied advantage over China.[1] [2]

That is an important distinction. A unilateral pause by responsible developers would not freeze the global capability landscape. It could instead transfer advantage to actors that do not pause, do not disclose incidents, or do not permit independent scrutiny.

Kurtz’s response brings the problem out of the laboratory and into production. His argument is that the unit of threat is no longer merely the individual hacker but the autonomous campaign: coordinated agents operating at machine speed across endpoints, cloud workloads, identities, and software services. In that environment, a policy document does not stop an agent already executing. Runtime enforcement does.[3]

“Pacing what comes next doesn’t secure what’s already here.” — George Kurtz[3]

That is the central tension. Frontier governance is necessary, but governance operates upstream. Cyber defense must also operate **during execution**, at the moment an agent requests a credential, invokes a tool, changes a configuration, moves data, or attempts to escape its assigned environment.

The question is no longer whether AI safety belongs to the policy team or the security team. It belongs to the entire operating system of the enterprise.

THE ARMS RACE IS ALREADY INSIDE THE SYSTEM

The U.S.–China AI competition is often described as a contest for chips, models, talent, and compute. It is also a contest over **access, extraction, resilience, and control**.

A September 8 joint advisory from CISA, the NSA, and the FBI states that China-based AI companies have conducted industrial-scale distillation campaigns against U.S. frontier models since at least late 2024. The agencies say the activity involved billions of tokens across millions of exchanges and used native APIs, cloud providers, aggregators, proxy networks, fraudulent accounts, and automated failover to evade controls.[4]

Whether Beijing accepts that characterization or not, the strategic reality is clear: capability does not remain neatly contained within the laboratory that created it. It can be copied, extracted, approximated, open-sourced, stolen, or accessed through intermediaries. China has rejected Amodei’s China-focused measures as fearmongering and containment, while Chinese officials have simultaneously warned about the national-security risk of foreign AI capabilities.[5]

Both countries understand the same underlying fact: **AI capability is now part of national power.**

That makes a universal, perfectly enforced slowdown unlikely. It also makes cybersecurity part of industrial strategy. Model weights, training clusters, inference infrastructure, APIs, identity systems, data pipelines, and the human expertise surrounding them are not peripheral technology assets. They are elements of the AI industrial base.

If those systems are not protected, the race is not simply about who can build the best model. It is about who can secure, retain, deploy, and govern capability after it has been built.

THE THREAT HAS ALREADY CHANGED

Anthropic's September threat report describes a shift from AI as an assistant to AI as an orchestrator. It documents actors using AI across reconnaissance, exploitation, phishing, persistence, and data exfiltration. In one case, AI-driven workflows monitored whether malware had been detected and then modified and rebuilt it to evade defensive controls. Anthropic concludes that sophistication is becoming less useful as an attribution signal because advanced execution is increasingly available to less capable actors.[6]

This is what makes the waiting strategy insufficient.

A statute passed next year cannot revoke a credential stolen tonight. A new reporting regime cannot stop an agent currently moving through a cloud environment. A governance committee cannot contain a system if nobody gave it a distinct identity, constrained its permissions, recorded its actions, or retained the ability to shut it down.

Government has a vital role. It can define minimum duties, require disclosure, establish liability, protect competition, coordinate allies, set procurement standards, and punish reckless conduct. But government cannot be the control plane for every autonomous system. Regulation is necessarily deliberative, jurisdictional, and retrospective. Threat activity is adaptive, distributed, and immediate.

We should not wait for government to tell organizations how to secure capabilities they have already deployed. Nor should lawmakers attempt to design technical controls without the operators who understand how those controls fail under real-world pressure.

Public policy should set the obligations. Cybersecurity practitioners, system designers, infrastructure operators, and independent evaluators must shape how those obligations work in production.

TURN PACING INTO A SECURITY DIVIDEND

If the frontier slows, the time gained must not become a regulatory waiting room. It should become an investment window.

The highest-value investments are not abstract promises of "responsible AI." They are the systems that reduce blast radius, preserve human authority, and create proof of control.

Investment priority	What must exist in production	Why it matters
Identity for every agent	A unique nonhuman identity, least privilege, short-lived credentials, per-action authorization, and immediate revocation	An agent without attributable identity is an ungoverned privileged user.
Runtime policy enforcement	Controls at endpoints, cloud workloads, SaaS applications, APIs, tools, and data stores	Model-level safeguards can be bypassed; the action must still be stopped at execution.
Sandboxing and blast-radius control	Segmented environments, network allowlists, constrained tools, isolated compute, and protected credentials	A capable agent should never inherit the full reach of the person or system that launched it.
Continuous observability	Immutable logs, tool-call records, network telemetry, behavioral monitoring, and 24/7 response integration	Organizations cannot govern actions they cannot reconstruct or attribute.
Independent adversarial testing	Embedded evaluators, external red teams, production-relevant scenarios, and authority to report material findings	Self-attestation is not proof. Safety claims require access, evidence, and challenge.
Machine-speed defense with human boundaries	Automated containment for reversible events; named human authority for high-consequence actions	Defense must match attacker speed without outsourcing irreversible judgment.
Circuit breakers, kill switches, and recovery	Predefined intervention thresholds, graduated throttling, tested shutdown authority, credential revocation, network isolation, rollback, continuity plans, and manual fallback	Stopping the process is not enough if the agent's access, changes, and dependencies remain active.
Protection of the AI industrial base	Security for model weights, training clusters, inference APIs, research data, supply chains, and technical personnel	Capability theft and service disruption are now geopolitical and economic-security risks.
Collective defense with privacy intact	Cross-company indicators, rapid incident disclosure, allied coordination, and data-minimized sharing	Autonomous campaigns distribute themselves across providers; defense must connect the pattern without creating a surveillance system.

FROM MARKET CIRCUIT BREAKERS TO AI KILL SWITCHES

Finance already understands that speed requires interruption mechanisms. Market-wide circuit breakers do not abolish trading or ask markets to stop evolving. They establish predefined thresholds at which activity pauses so information can be assessed and disorderly behavior does not compound without restraint.[10]

AI needs an equivalent architecture.

This cannot be a theatrical red button controlled by one executive, nor a universal switch capable of turning off “AI.” It must be a **graduated, system-specific control framework** designed before deployment and tested under realistic conditions.

AI circuit-breaker level	Trigger	Required response
Level 1 — Constrain	Anomalous tool use, unexpected data access, repeated policy violations, or abnormal resource consumption	Reduce permissions, throttle execution, restrict network access, and require human approval for the next action.
Level 2 — Quarantine	Attempted privilege escalation, sandbox escape, credential misuse, unexplained lateral movement, or material deviation from the authorized objective	Revoke credentials, isolate the agent and its workspace, preserve evidence, and transfer control to incident response.
Level 3 — Halt	Credible risk to customers, critical operations, financial integrity, safety, or national security	Terminate execution, sever model and tool connectivity, stop dependent agents, invoke continuity procedures, and require named executive authorization before restart.
Level 4 — Coordinate	A campaign or failure pattern crosses companies, infrastructure providers, or national boundaries	Share minimized indicators rapidly, activate sector and allied response channels, and suspend the affected capability until independently assessed.

The hard part is not inventing the switch. It is determining **what trips it, who has authority to use it, what else must stop with it, how evidence is preserved, how critical services continue, and what standard must be met before restart.**

That is where finance offers a useful model. Circuit breakers work because thresholds, authority, operating procedures, and consequences are established in advance. AI kill switches must be treated the same way: as engineered resilience, not improvised crisis management. Every high-consequence AI deployment should have a documented shutdown hierarchy, a tested manual fallback, and a board-approved risk threshold beyond which autonomous operation is no longer permitted.

Official guidance is converging around these controls. The U.K.'s national cybersecurity agency recommends distinct agent identities, shortest-possible credential lifetimes, sandboxing, real-time monitoring, human oversight for consequential events, and an immediate ability to halt autonomous activity.[7] A Five Eyes guidance document similarly warns against granting agents broad access to sensitive data or critical systems and recommends integrating agentic AI into established cybersecurity frameworks rather than treating it as a separate governance exercise.[8]

The important shift is from **AI principles** to **AI operating discipline**.

SAFETY IS NOT THE OPPOSITE OF SPEED

There is a persistent assumption in the AI debate that safety and speed sit on opposite sides of a trade-off. That is true only when safety is added after the system is built.

Security architecture can increase safe speed. Strong identity reduces the need for blanket restrictions. Sandboxing allows experimentation without exposing the entire enterprise. Observability makes bounded autonomy possible. Independent evaluation creates credible evidence for release decisions. Tested recovery gives leaders options when a deployment behaves unexpectedly.

The organizations that invest now will not merely be safer. They will be able to deploy more confidently, partner more credibly, recover more quickly, and demonstrate control to customers, boards, insurers, investors, and regulators.

CrowdStrike's SafeMind announcement offers one model of the direction of travel: offensive and defensive models operating in a closed loop, with the defensive system designed to act on risk rather than simply describe it.[9] Its performance claims require independent validation, as all vendor claims do. But the strategic point is sound: **the ability of AI to act must be matched by a defensive capacity to observe, constrain, and stop it.**

This is where capital should move.

Not away from AI, but toward the infrastructure that makes AI survivable.

THE BOARD AGENDA

Every board deploying agentic AI should be able to answer five questions now:

1. **Which AI systems can take action, and what is the maximum harm each could cause before a human intervenes?**
2. **Does every agent have a distinct identity, limited permissions, traceable actions, and credentials that can be revoked immediately?**
3. **Can we detect and stop abnormal behavior at runtime—not after an incident review?**
4. **What thresholds trigger constraint, quarantine, or shutdown; who has authority to act; and has the full circuit-breaker process been tested under operational pressure?**
5. **Which model, infrastructure, data, and vendor dependencies could become unavailable or compromised through geopolitical action?**

If leadership cannot answer those questions, the organization does not have an AI strategy. It has an AI exposure.

THE KHAOS VIEW

Amodeli is right to ask the frontier laboratories to create breathing room. Embedded evaluators, meaningful incident disclosure, stronger alignment work, and verifiable thresholds would be material improvements. Kurtz is right that the threat environment will not honor that breathing room.

The synthesis is straightforward:

Pace where pacing is verifiable. Defend everywhere at machine speed.

Do not wait for a perfect international agreement. Do not wait for every legislature to understand agentic systems. Do not wait for case law to determine who was responsible after the damage is done.

Move now: fund the controls, protect the infrastructure, constrain the agents, engineer the circuit breakers, test the shutdowns, share the signals, and keep humans responsible for the decisions that can change lives.

Technology does not wait.

Neither can safety.

EXECUTIVE SIGNAL

Decision window: Immediate

Primary exposure: Autonomous cyber operations are scaling faster than institutional controls.

Required action: Reallocate AI investment toward runtime security, agent identity, independent assurance, tested circuit breakers, operational resilience, and human control before expanding autonomous deployment.

Owner: CEO, CIO, CISO, General Counsel, and Board Risk Committee.

REFERENCES

[1] Dario Amodei, We Must Pace the Frontier. <https://darioamodei.com/post/we-must-pace-the-frontier>

[2] Reuters, Anthropic CEO urges AI companies to slow model development amid fears over misuse.

<https://www.reuters.com/business/anthropic-ceo-urges-ai-companies-slow-model-development-2026-09-12/>

[3] George Kurtz, response to We Must Pace the Frontier. https://x.com/George_Kurtz/status/2099263605900255575

[4] CISA, NSA and FBI, China-Based Artificial Intelligence Companies Conducting Industrial-Scale Distillation Campaigns Against U.S. AI Companies.

<https://www.cisa.gov/news-events/cybersecurity-advisories/aa26-251a>

[5] Associated Press, China bristles at Anthropic CEO's fearmongering about its AI development. <https://apnews.com/article/china-anthropic-ai-us-amodei-3da458d2c078da3e60900728d59f1ae8>

[6] Anthropic, Detecting and countering misuse of AI: September 2026. <https://www.anthropic.com/threat-intelligence-report-september-2026>

[7] U.K. NCSC, Managing the cyber risk of agentic AI. <https://www.ncsc.gov.uk/blogs/managing-the-cyber-risk-of-agentic-ai>

[8] Five Eyes agencies, Careful adoption of agentic AI services. <https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/careful-adoption-of-agentic-ai-services>

[9] CrowdStrike, SafeMind announcement. <https://ir.crowdstrike.com/news-releases/news-release-details/crowdstrike-launches-frontier-models-cybersecurity-created>

[10] U.S. Securities and Exchange Commission, Investor Bulletin: New Measures to Address Market Volatility.

<https://www.sec.gov/investor/alerts/circuitbreakersbulletin.htm>

Khaos Holdings

Intelligence for Complex Systems

khaosholdings.com

#ArtificialIntelligence #Cybersecurity #AgenticAI #AIGovernance #Geopolitics #NationalSecurity #USChina #CriticalInfrastructure #EnterpriseRisk
#KhaosSignal