Testimony for the hearing “Security to Model: Securing Artificial Intelligence to Strengthen Cybersecurity”

Jonathan Dambrot, CEO and Co-Founder, Cranium AI

Before the *House Homeland Security Subcommittee on Cybersecurity and Infrastructure Protection*
June 12, 2025

Chairman Garbarino, Ranking Member Swalwell, and distinguished members of the Subcommittee,

Thank you for the opportunity to testify today on the crucial matter of securing artificial intelligence to strengthen our nation’s cybersecurity. My name is Jonathan Dambrot, and I serve as the CEO and Co-Founder of Cranium, an AI security and governance platform built to enable safe, resilient, and innovative AI adoption across enterprises and critical infrastructure. At Cranium, we believe that the future of secure AI begins not just with awareness of risk, but with a foundational shift to transparency, accountability, and continuous security throughout the lifecycle of every AI system. As the U.S. competes to lead in global AI development, we must advance innovation hand-in-hand with strong governance to ensure that AI progress reinforces our national security and democratic values.

AI is transforming every industry, promising significant benefits, but it also introduces new cyber risks. As an enterprise focused on AI security, Cranium’s perspective is that we must secure AI systems both before they are deployed and throughout their operational lifecycle. This means building security into AI “by design” from the outset and maintaining robust defenses and oversight “by default” after deployment. My testimony will outline Cranium’s mission and technology capabilities, how they advance AI security, and policy considerations for fostering innovation while safeguarding our systems. I will also align with themes raised by important initiatives, such as the Secure by Design principles, Secure by Demand procurement, the importance of guardrails, and proactive risk mitigation, and address emerging threats, including autonomous AI agents and foreign AI models, like China’s DeepSeek.

Artificial intelligence is now deeply embedded into our digital infrastructure. From foundation models and generative systems to embedded AI in third-party software, the pace and scale of adoption has accelerated beyond what many expected. **This proliferation has introduced**

new risks—complex supply chains, unmonitored shadow AI deployments, and misaligned third-party integrations—that are difficult to address without robust, AI-native security strategies.

Cranium's Platform and Approach

At Cranium, we have developed an end-to-end platform that supports visibility, risk assessment, continuous monitoring, and model governance for AI systems. Our Detect AI product helps organizations automatically map and inventory their AI assets, including internal and external models, embedded third-party APIs, and training datasets, providing what we call an **AI Bill of Materials, or AI-BOM.** This capability is vital in today's environment, where unmanaged and undocumented AI systems often operate in mission-critical environments.

Our AI security capabilities take that visibility a step further, using **automated risk analysis and AI-specific threat intelligence to identify vulnerabilities in models, pipelines, and data flows.** Cranium Arena, our **adversarial testing and red-teaming** product, simulates real-world attack scenarios, including model evasion, prompt injection, and adversarial data poisoning. We developed Arena because organizations need to identify weaknesses not only in their own models, but in the external models and services they depend on—a critical need in an era of supply chain infiltration and rapidly evolving AI systems.

Through our AI Card product, we provide **structured documentation of an AI system'**s **security posture, compliance readiness, and development history,** offering both internal stakeholders and external regulators a clear, verifiable view of the AI system's lifecycle and associated controls. This promotes trust, facilitates audits, and aligns with emerging transparency requirements from the NIST AI Risk Management Framework, the EU AI Act, and forthcoming U.S. regulatory guidance.

Secure-by-Design Development and Emerging Threats

Securing AI before deployment must be a central principle. Just as we expect bridges to be built with safety in mind from day one, we need AI developers to integrate security into the very architecture of AI models and applications. Unfortunately, today, this is more the exception than the rule. Studies by global cybersecurity bodies have found that many enterprise AI and data science teams often neglect to consider security concerns during the design stage, largely due to a lack of awareness about the unique threats to AI systems. In other words, many AI projects start without a threat model or security requirements, focusing only on functionality and accuracy. This gap at the front end of model development is where serious risk can breed, from training data vulnerabilities to models that are overly permissive in what they will output, to a lack of audit mechanisms.

To mitigate AI risk early, we must change the development culture and tooling. **Security should be treated as a first-class concern in model design and training, just as performance or accuracy is**. This includes conducting AI-specific threat assessments and red-team exercises prior to deployment, as well as building “guardrails” into the model. For example, developers should consider how their model could be misused or attacked (such as through adversarial inputs or prompt injections) and implement constraints or filters to prevent these attacks. By catching issues in a controlled test environment, we could avoid costly fixes or incidents later. In essence, this brings the mindset of a security researcher into the AI lab from the outset.

A concrete example involves large language models (LLMs) that enterprises fine-tune for internal use (say, a customer service chatbot). Rather than deploying the model and hoping for the best, a secure-by-design approach would utilize red-teaming and adversarial simulation to subject the chatbot to a battery of abuse scenarios during development and testing, verifying whether it can be prompted to reveal confidential information or if it exhibits biases or unsafe behaviors. If the red-team simulation shows that the model can be tricked (as often happens with out-of-the-box models), developers can then implement guardrails, such as content filters, refined training data, or adjusting the model’s temperature and response rules. **The goal is that by the time the AI system goes live, it has been hardened against foreseeable attacks or failures. This practice is analogous to unit-testing and QA in traditional software, extended to the AI context.**

Embracing **Secure by Design for AI also means empowering those building AI with better knowledge and incentives.** We must ensure education is prioritizing security and by disseminating secure AI development guidelines (such as those recently published by the US CISA and the UK’s NCSC). The government could help with this crucial endeavor by supporting frameworks and tools (like NIST’s AI Risk Management Framework) that make it easier to integrate security in AI development, and by funding research into AI security techniques. Security shouldn’t be viewed as a blocker to innovation, but as a prerequisite for sustainable innovation. Indeed, when done right, security improves quality: The same design decisions that make AI systems secure by default also lead to more resilient, higher-quality code that is cheaper to maintain in the long run. **In short, security and innovation can go hand-in-hand—there does not have to be a tradeoff.**

Yet, unlike traditional software, red teaming AI models often involves probing behaviors protected by access controls or license agreements, which has historically exposed researchers to legal risk. The recent **DMCA Section 1201**[1] rulemaking rejected an exemption for AI "trustworthiness" research, meaning researchers conducting critical red-teaming could still face legal uncertainty or liability when circumventing technological protection measures. This highlights the urgent need for **safe harbor protections**, as advocated by OpenPolicy[2], which would **legally shield good-faith actors engaged in testing for bias, harmful outputs, or security vulnerabilities in AI systems**. Without such protections, we risk chilling the very research needed to secure and audit AI models before deployment.

Further, Congress and this Subcommittee can play a pivotal role in promoting secure-by-design principles for AI. Foster the development of outcomes-based security practices rather than relying on burdensome checklists. Rather than relying on "check-the-box" compliance, we should start measuring actual product security outcomes. Regulations or procurement rules should focus on whether AI systems have been rigorously tested, what vulnerabilities were found and fixed, and how the system behaves under stress, not just whether a paperwork process was followed. For example, an outcomes-based approach could require that any AI system used in critical infrastructure undergo an independent red-team assessment and that the findings be addressed, rather than requiring a specific certification that may not reflect real security. This incentivizes genuine risk reduction, enabling outcomes-based security by generating concrete evidence, such as vulnerability scan results and compliance scores, that demonstrate the security posture. **We urge policymakers to encourage such evidence-driven security accountability in AI development.**

Beyond pre-deployment security, we emphasize the need for **continuous protection throughout the operational lifecycle of AI systems**. Just as traditional cybersecurity recognizes the importance of ongoing patch management and threat monitoring, AI systems require persistent oversight. **Models evolve, inputs shift, and adversaries adapt, requiring a governance framework that is not static but dynamic.**

As we evaluate how best to secure the nation's AI infrastructure, we must also confront the growing threat posed by **autonomous AI agents.** These agentic systems, while promising for automating workflows and enhancing productivity, also introduce a new and complex class of security risks. A compromised or maliciously directed AI agent could autonomously conduct cyber operations at machine speed. Imagine an AI system that iteratively probes a target network for vulnerabilities, adapts to defensive measures in real time, exfiltrates data,

[1] DMCA Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies https://www.govinfo.gov/content/pkg/FR-2024-10-28/pdf/2024-24563.pdf

[2] See OpenPolicy comments to the Copyright Office concerning the Ninth Triennial Proceeding on section 1201 exemptions, Class Four https://shorturl.at/9zavx

conceals its activity, and even generates new malware variants on the fly, **without human intervention.** This capability would enable end-to-end automated hacking at a scale and velocity that far exceeds current threats, and it would lower the skill barrier for malicious actors. We are already seeing precursors of this evolution, with AI being used to create more convincing phishing campaigns and to refine malware that evades detection. The next stage is full autonomy in offensive operations.

To counter this risk, **security must be embedded throughout the lifecycle of AI agents—from the moment they are conceived and built to the moment they are deployed and run.** At the **build-time phase,** this means establishing full visibility and accountability for who is building which agents, for what purpose, and with what data. Organizations must maintain an **up-to-date inventory of agents** under development and classify them according to risk level, functionality, and data access. In parallel, there must be robust **governance of training data**: identifying and securing sensitive datasets, understanding data provenance, and preventing data leakage or manipulation. These practices align with long-standing security principles reflected in the OWASP Top 10 Web Application Security Risks[3], including the need to prevent broken access control, mitigate vulnerable and outdated components, and defend against injection attacks, even in the context of training and fine-tuning AI models.

Build-time security also requires accountable risk assessments. These should evaluate the design of AI agents, their integration with third-party components, and the safeguards in place to ensure they function within expected parameters. Importantly, as organizations increasingly rely on low-code and no-code platforms to accelerate AI deployment, these environments must not be overlooked. The OWASP Top 10 for Low-Code/No-Code Security Risks[4] makes clear that abstracted development frameworks can introduce unique vulnerabilities, such as unauthorized code execution, insufficient policy enforcement, and risky use of prebuilt components. **AI agents built in these environments must be held to the same security standards as those developed in traditional pipelines**.

At **runtime**, the focus must shift to ensuring that the AI agent's behavior aligns with its intended function. This involves continuously monitoring how agents operate in real-world environments, assessing whether their actions align with their declared objectives, and identifying anomalies that may indicate compromise or misuse. **When behavior deviates from expectations,** such as an agent attempting to scan networks, alter logs, or communicate with unauthorized endpoints, **those deviations must trigger investigation.**

[3] See OWASP Top 10 Web Application Security Risks https://owasp.org/www-project-top-ten/
[4] See OWASP Top 10 for Low-Code/No-Code Security Risks https://owasp.org/www-project-top-10-low-code-no-code-security-risks/

Runtime guardrails, such as sandboxing, rate limiting, and technical policy enforcement, are crucial for mitigating these risks. These measures align with the core concerns of the OWASP Top 10, particularly those related to insecure design, security misconfigurations, and improper asset management. Securing runtime behavior also requires tools that can **interpret agent decisions and provide meaningful visibility** into how outcomes are generated, even if full explainability remains a challenge.

Importantly, there is no simple solution to these risks. **Relying on "AI to stop AI" is not a viable defense strategy.** While defensive AI can play a role, it cannot fully prevent prompt injections, model drift, or malicious repurposing. You cannot block every injection, and attackers are already experimenting with novel ways to subvert AI systems through their inputs, training data, or deployment contexts. In many cases, AI agents will not be able to reliably detect one another, especially if adversarial agents are designed to evade such detection or mimic legitimate behavior**. The threat of AI-enabled attacks necessitates layered, proactive defenses, rather than reactive arms races.**

Ultimately, securing AI agents requires a paradigm shift. **We must move beyond reactive patching and toward a comprehensive model of proactive governance and control.** That means treating AI agents as autonomous actors with the potential to cause harm, embedding trust mechanisms throughout their lifecycle, and **designing guardrails that are technically enforced, not just documented**. Without securing the build process, runtime protections will always be a step behind. And without establishing behavioral baselines, we will fail to detect when agents deviate from acceptable norms. As we consider how best to govern and secure AI, we must recognize that autonomous agents are not just a theoretical risk—they are an imminent reality. **Our security posture must rise to meet them.**

At Cranium, when we simulate automated AI red-teamers in our Arena platform, we always contain them in controlled environments and define strict boundaries on what they can do. **We need similar thinking industry-wide: those developing agentive AI should integrate safety layers that prevent harmful actions.** The U.S. can lead by articulating clear red lines on the use of autonomous AI, both to guide its own use and to set expectations for others. If we don't address this, we risk an arms race of "AI vs AI" where malicious autonomous agents battle defensive AIs across networks, which could spiral out of control.

Addressing Post-Deployment and Supply Chain Risk

Another stark development is **the rise of advanced AI models from geopolitical rivals that do not share our security or democratic values.** The clearest example is **DeepSeek**, a generative AI model developed in China. DeepSeek has been described as "**a cybersecurity**

wake-up call" because of how it has been leveraged for malicious purposes. DeepSeek is effectively **unrestricted and open-source**, allowing anyone to modify it and remove safety limits.[5] This has made it a boon for cybercriminals. Reports indicate that **DeepSeek enables users to generate fully functional malware from scratch, including ransomware, without needing technical expertise[6].** Beyond enabling cybercrime, DeepSeek also presents a national security and privacy threat. The application is subject to PRC law, and its terms make clear that user data is stored on servers in China. Under China's regime, that means the government can access any data put into DeepSeek. Security analysts identified weak encryption and hidden data transmissions to Chinese state-linked infrastructure within DeepSeek. In short, anyone using this ostensibly "free AI tool" could be unwittingly sending their data straight to Beijing. It is easy to see how this could be weaponized for espionage: an American engineer might use DeepSeek to help write code and inadvertently provide China with sensitive intellectual property, or a government employee might experiment with DeepSeek out of curiosity and leak some internal information.

DeepSeek illustrates the broader geopolitical competition around AI. In this arena, the U.S. and like-minded nations cannot be complacent. We must continue to lead in cutting-edge AI development, but also in **embedding security and democratic values into AI**. If we do not, the default AI technologies globally may become those like DeepSeek that are insecure and don't prioritize safety and trust. **It serves as a reminder that AI capability alone is not enough – we must couple capability with responsibility.**

Our best response is to double down on what this hearing is all about: *securing AI to strengthen cybersecurity.* We need to be proactive and forward-looking, anticipating these threats and preparing defenses and norms accordingly. I am encouraged that Congress is now examining these issues. With thoughtful strategy and the cooperation of industry, we can navigate these challenges and ensure the US maintains its leadership in AI innovation and adoption.

Public-Private Collaboration

As AI becomes a competitive frontier among nations, the U.S. must lead not only in AI capability but in AI security standards. The question is no longer who builds AI first, but who builds it securely, transparently, and with accountability. To do that, we must support pro-innovation regulation that incentivizes excellence in AI development while placing necessary

[5] See WIZ report and research on DeepSeek https://shorturl.at/JZcar

[6] See CSIS report on Delving into the Dangers of DeepSeek https://shorturl.at/9w4gV

guardrails around its use. The federal government must establish clear expectations regarding third-party model evaluation, particularly for models developed under foreign jurisdictions.

We also cannot overlook the importance of government capacity. Cranium has worked closely with the Cybersecurity and Infrastructure Security Agency (CISA) as part of its Joint Cyber Defense Collaborative (JCDC) and AI risk initiatives. We contributed to CISA's first AI security tabletop exercise and the development of its AI incident response playbook. These partnerships have been instrumental in building shared understanding and rapid response frameworks.

In conclusion, the future of AI is inseparable from the future of cybersecurity. Our national resilience, economic leadership, and public trust all hinge on how well we govern the development and deployment of artificial intelligence. Cranium remains committed to advancing solutions that secure AI by design and by default. We urge this Subcommittee to continue its leadership in shaping thoughtful, effective, and forward-looking policies that make secure AI not only possible, but inevitable.

Chairman, Ranking Member, and Members of the Subcommittee, **securing artificial intelligence to strengthen cybersecurity is one of the defining challenges and opportunities of our time.** AI will undoubtedly shape the future of our economy and national security. Whether that future is more secure or more dangerous depends on the actions we take today to embed security, accountability, and resilience into our AI ecosystem.

Our mission to enable safe AI adoption, along with our alignment with principles such as Secure by Design and Secure by Demand, ensures that security is woven into AI from the start and continuously reinforced. This commitment to innovation-friendly safeguards protects without stifling progress.

We discussed concrete steps to secure AI before deployment (through proactive design and testing) and after deployment (through ongoing monitoring and incident readiness). We examined emerging threats on the horizon, from rogue AI agents to adversarial nation-state models, that require vigilance and a united front. **And we highlighted how the private sector and government, working together, can set global norms and rapidly improve AI governance in practice.**

I want to underscore a few recommendations derived from today's discussion, particularly:

- **Congress should champion Secure-by-Design and Secure-by-Demand approaches.** Promote policies in procurement, research and development funding, and agency oversight that require or incentivize building security into AI from the ground up. Encourage the demand side—federal buyers—to require evidence of security in the form of model cards, AI risk assessments, or FedRAMP-style attestations.

- **Address emerging threats directly.** Update policies to recognize new risks like autonomous AI-enabled cybercrime, support export controls for high-risk models, and work with allies to establish norms prohibiting the development or use of AI for malicious cyber operations.
- **Promote transparency and standards adoption.** Encourage organizations in critical sectors to adopt frameworks like NIST's AI RMF. Consider implementing rules that mandate incident reporting for major AI-related disruptions, similar to those in traditional cybersecurity, to establish a body of shared knowledge and proactive defenses.
- **Foster innovation through flexible, risk-based, security-first regulation.** Develop regulatory sandboxes that allow developers to test compliance. Continue robust funding for AI safety and security R&D. The U.S. innovation ecosystem is our greatest strength; we should use it to solve the problems AI itself introduces.

In closing, I am optimistic. The fact that we are having this hearing shows we are not destined to relive the mistakes of the past, where technology's risks were realized before its safety was secured. We have the knowledge, tools, and collaborative spirit to get ahead of AI threats. **Companies like Cranium exist precisely to ensure that *security advances in tandem with AI advancements***. We believe the United States can lead the world in both AI innovation **and** AI security. If we do, our cybersecurity will be stronger, our values will be upheld, and our citizens will reap AI's benefits without unnecessary fear.

Thank you for the opportunity to testify and for your leadership on this issue. I look forward to answering any questions you may have.