Testimony of Royal Hansen,
Vice President of Privacy, Safety, and Security Engineering
U.S. House Committee on Homeland Security
Subcommittee on Cybersecurity and Infrastructure Protection &
Subcommittee on Oversight, Investigations, and Accountability
December 17, 2025

Chairmen Garbarino, Ogles, Brecheen; Ranking Members Thompson, Swalwell, Thanedar; and Members of the Committee and Subcommittees: thank you for the opportunity to speak with you today. My name is Royal Hansen, and I serve as Vice President of Privacy, Safety, and Security Engineering at Google. Our team is responsible for building and scaling the foundational technology to keep billions of people safe online.

Thank you for holding this important hearing. We welcome the opportunity to provide information about Google's efforts to secure its own artificial intelligence, protect its customers' workloads, and use artificial intelligence to strengthen cyberdefense and enhance our collective security.

Securing our Artificial Intelligence

Google's AI principles, published in 2018 and updated this year, describe our commitment to developing technology responsibly and in a manner that is built for safety, enables accountability and upholds high standards of scientific excellence. We have built on this work through our Secure AI Framework, as well as with extensive model hardening and various governance measures. This comprehensive approach means we secure all components of the AI ecosystem including data, infrastructure, applications, and models.

The Secure AI Framework (SAIF)

SAIF is our framework for integrating security and privacy measures into machine learning and generative AI applications and it governs how we embed controls throughout the AI system stack from data, infrastructure, application and models. The framework, which is designed to ensure that AI models are secure by design, has six core elements:

- **Expand strong security foundations to the AI ecosystem.** Leverage secure-by-default infrastructure protections and expertise built over the last two decades to protect AI systems, applications and users. At the same time, develop organizational expertise to keep pace with advances in AI and start to scale and adapt infrastructure protections in the context of AI and evolving threat models. For example,

injection techniques like SQL injection have existed for some time, and organizations can adapt mitigations, such as input sanitization and limiting, to help better defend against prompt injection style attacks.

- **Extend detection and response to bring AI into an organization's threat universe.** Detect and respond to evolving AI-related cyber incidents by extending threat intelligence and other capabilities. For organizations, this includes monitoring inputs and outputs of AI systems to detect misuses, and using threat intelligence to anticipate attacks. This effort typically requires collaboration with trust and safety, threat intelligence, and counter abuse teams.

- **Automate defenses to keep pace with existing and new threats** Harness the latest AI innovations to improve the scale and speed of response efforts to security incidents. Adversaries will use AI to scale their impact, so it is important to use AI and its current and emerging capabilities to stay nimble and cost effective in protecting against them. It is important to remember that the vast majority of successful attacks - whether AI-enabled or not- prey on legacy systems; AI can help defenders modernize and address issues at a scale and speed that has historically proved challenging.

- **Harmonize platform level controls to ensure consistent security across the organization.** Align control frameworks to support AI risk mitigation and scale protections across different platforms and tools to ensure that the best protections are available to all AI applications in a scalable and cost efficient manner. At Google, this includes extending secure-by-default protections to AI platforms like Vertex AI and Security AI Workbench, and building controls and protections into the software development lifecycle. Capabilities that address general use cases, like Perspective API, can help the entire organization benefit from state of the art protections.

- **Adapt controls to adjust mitigations and create faster feedback loops for AI deployment.** Constantly test implementations through continuous learning and evolve detection and protections to address the changing threat environment. This includes techniques like reinforcement learning based on incidents and user feedback, and involves steps such as updating training data sets, fine-tuning models to respond strategically to attack attempts, and allowing the software that is used to build models to embed further security in context (e.g. detecting anomalous behavior). Organizations can also conduct regular Red Team exercises to improve safety assurance for AI-powered products and capabilities. These are exactly the techniques we have used to defend Gmail, the Play Store and Chrome with AI at scale for many years.

- **Contextualize AI system risks in surrounding business processes.** Conduct end-to-end risk assessments related to how organizations will deploy AI. This includes

an assessment of the end-to-end business risk, such as data lineage, validation and operational behavior monitoring for certain types of applications. In addition, organizations should construct automated checks to validate AI performance. Nearly all businesses are increasingly digital - AI will only accelerate that trend. The controls required to mitigate risks in these processes must keep pace - some of which will be digital and some will be procedural.

Model Hardening

Our AI models are fine-tuned on large datasets of realistic attack scenarios to build intrinsic resilience. They are taught to recognize and ignore malicious instructions while still following user requests. This is, and will continue to be, an evolving space requiring rapid iterations as attackers innovate.

Over the past decade, we have evolved our approach to translate the concept of red teaming to the latest innovations in technology, including AI. The AI Red Team is closely aligned with traditional red teams, but also has the necessary AI subject matter expertise to carry out complex technical attacks on AI systems. A core part of our security strategy is automated red teaming, where our internal Gemini team constantly attacks Gemini in realistic ways to uncover potential security weaknesses in the model. We fine-tuned Gemini on a large dataset of realistic scenarios, where automated red teaming generates effective indirect prompt injections targeting sensitive information.

Protecting AI models against attacks like indirect prompt injections requires "defense-in-depth" – using multiple layers of protection, including model hardening, input and output checks (like classifiers), and system-level guardrails. Securing advanced AI systems against specific, evolving threats like indirect prompt injection is an ongoing process. It demands pursuing continuous and adaptive evaluation, improving existing defenses and exploring new ones, and building inherent resilience into the models themselves.

Securing AI Workloads

Recent headlines have highlighted several key vulnerabilities and attack vectors targeting private and public sector entities. It is clear that legacy systems, misconfigured cloud environments, and the exploitation of known vulnerabilities remain significant concerns. Email phishing, supply chain attacks, criminal hacking, and state-sponsored cyber espionage further compound these challenges. Our approach to protecting public and private sector entities is built on several core tenets:

- AI-Powered Security: We leverage the power of AI and machine learning to enhance threat detection, automate security operations, and secure AI development.
- Secure by Design: We engineer security into every layer of our infrastructure and services, from custom-designed hardware to advanced encryption techniques. To do this well requires security engineering which goes well beyond checklists and compliance requirements.
- Zero Trust: We ensure that no user or device is inherently trusted, regardless of their location or network. Access is continuously authenticated and authorized based on identity, device health, and context. We developed this approach in the wake of Chinese threat actor attacks on Google over 15 years ago, and it remains as important today.
- Shared Fate: We operate under a clear shared responsibility model, securing the underlying cloud infrastructure while providing tools and guidance for customers to manage their own security. We believe in a "shared fate" where our success is tied to the customer's. We are deeply invested in the collective security outcomes of consumers, companies and countries. We align our goals with the security and resilience of critical operations, particularly where national security is at stake.

Artificial Intelligence and Cybersecurity: Identifying Opportunities and Mitigating Risks

We stand at a critical technological inflection point. Rapid advances in AI are unlocking new possibilities for the way we work and accelerating innovation in science, technology, and beyond. Some of these same AI capabilities, however, can also be deployed by attackers, leading to understandable anxieties about the potential for AI to be misused for malicious purposes. Until recently, our analysis of government-backed threat actor use of AI revealed that threat actors were using generative AI primarily for common tasks like troubleshooting, research, and content generation. Over the past year, Google Threat Intelligence Group has identified an important shift, with adversaries not only leveraging AI for productivity gains, but experimenting with novel AI-enabled malware in active operations.

We have identified malware families that use LLMs to generate malicious scripts, obfuscate their own code to evade detection, and use AI models to create malicious functions on demand, rather than hard-coding them into the malware. This marks a new operational phase of AI abuse, involving tools that dynamically alter behavior mid-execution. While still nascent, this development represents a significant step toward more autonomous and adaptive malware. We have and will continue to publish on these topics, take action and enhance our products to ensure industries and societies as a whole can keep pace with the latest threats.

Today, and for decades, the main challenge in cybersecurity has been that attackers need just one successful, novel threat to break through the best defenses. Defenders, meanwhile, need to deploy the best defenses at all times, across increasingly complex digital terrain — and there is no margin for error. As we have seen in recent years, this is particularly true for legacy technology. This is the “Defender's Dilemma,” and there has never been a reliable way to tip that balance.

Our experience deploying AI at scale informs our belief that AI can reverse this dynamic in several ways and enhance our collective security.

- AI allows security professionals and defenders to scale and accelerate their work in threat detection, malware analysis, vulnerability detection, vulnerability fixing and incident response.

- Google’s AI-based efforts like BigSleep have demonstrated AI’s ability to find new zero-day vulnerabilities in well-tested, widely used software. Developed by Google DeepMind and Google Project Zero, Big Sleep can help security researchers find zero-day (previously-unknown) software security vulnerabilities. Since it was introduced last year, it has continued to discover multiple flaws in widely-used software, exceeding our expectations and accelerating AI-powered vulnerability research. With Big Sleep, we have demonstrated how we can find vulnerabilities that defenders don’t yet know about. In this case, we found a vulnerability that the attackers knew about and had every intention of using. We were able to detect and report it for patching before they could exploit it.

- Finding vulnerabilities is only half of the battle. Recently, we developed CodeMender, an AI-powered agent that utilizes the advanced reasoning capabilities of our Gemini models to automatically fix critical code vulnerabilities. CodeMender scales security, accelerating time-to-patch across the open-source landscape. It represents a major leap in proactive AI-powered defense and includes features such as root cause analysis and self-validated patching. This capability in particular will be the most significant security advancement in many years.

Collaboration Toward Responsible Artificial Intelligence Adoption

We believe the private sector, governments, educational institutions, and other stakeholders must work together to maximize AI's benefits while also reducing the risks of abuse. As innovation moves forward, the industry more broadly needs security standards for building and deploying AI responsibly. That's why Google introduced SAIF, as noted above, as a conceptual framework to secure AI systems. Our recent expansion to SAIF 2.0 addresses the rapidly

emerging risks posed by autonomous AI agents and extends our proven framework with new guidance on agent security risks and controls to mitigate them.

In addition, Google co-founded the Coalition for Secure AI (CoSAI), an open-source initiative to help all developers and deployers of AI create and maintain secure by design AI systems and help advance the framework. CoSAI helps foster a collaborative ecosystem to share open-source methodologies, standardized frameworks, and tools. Since its launch, CoSAI has made significant strides in strengthening AI security in collaboration with industry and academia in areas including Software Supply Chain Security for AI Systems; Preparing Defenders for a Changing Security Landscape; AI Security Risk Governance; and Secure Design Patterns for Agentic Systems. We have also supported the MLCommons Association's efforts to develop AI safety benchmarks by contributing funding for the development of a testing platform, as well as technical expertise and resources. ML Commons' shared research infrastructure helps the scientific research community derive new insights for breakthroughs in AI.

Across Google Cloud, we model and promote the adoption of responsible AI data practices that preserve our customers' privacy and support their compliance journey. Robust privacy commitments outline how we protect user data and prioritize privacy and the greater adoption of artificial intelligence rearms their importance. We adhere to a holistic approach to AI risk management and compliance, including focusing on employing an AI risk assessment methodology for identifying, assessing, and mitigating risks; developing and using an automated, scalable, and evidence-based approach for auditing generative AI workloads; and emphasizing human oversight and collaboration in our risk assessments and governance councils.

We use explainability tools to help understand and interpret AI predictions and evaluate potential bias; privacy-preserving technologies such as masking and tokenization and adhering to privacy laws; continuous monitoring and auditing for security vulnerabilities that AI might miss; investing in training programs to bridge the AI knowledge gap; and encouraging "interdisciplinary collaboration" between data scientists, risk analysts, and domain experts is also key.

Cybersecurity has never been a field where perfection is possible. It will remain a dynamic space for years to come, and speed and resilience will be required to defeat and contain innovative attackers. As governments and civil society leaders look to counter evolving threats from cybercriminals and state-backed attackers, we are committed to leading the way in using AI to tip the balance of cybersecurity in favor of defenders.

We appreciate the Committee convening this important hearing. And we look forward to answering your questions.