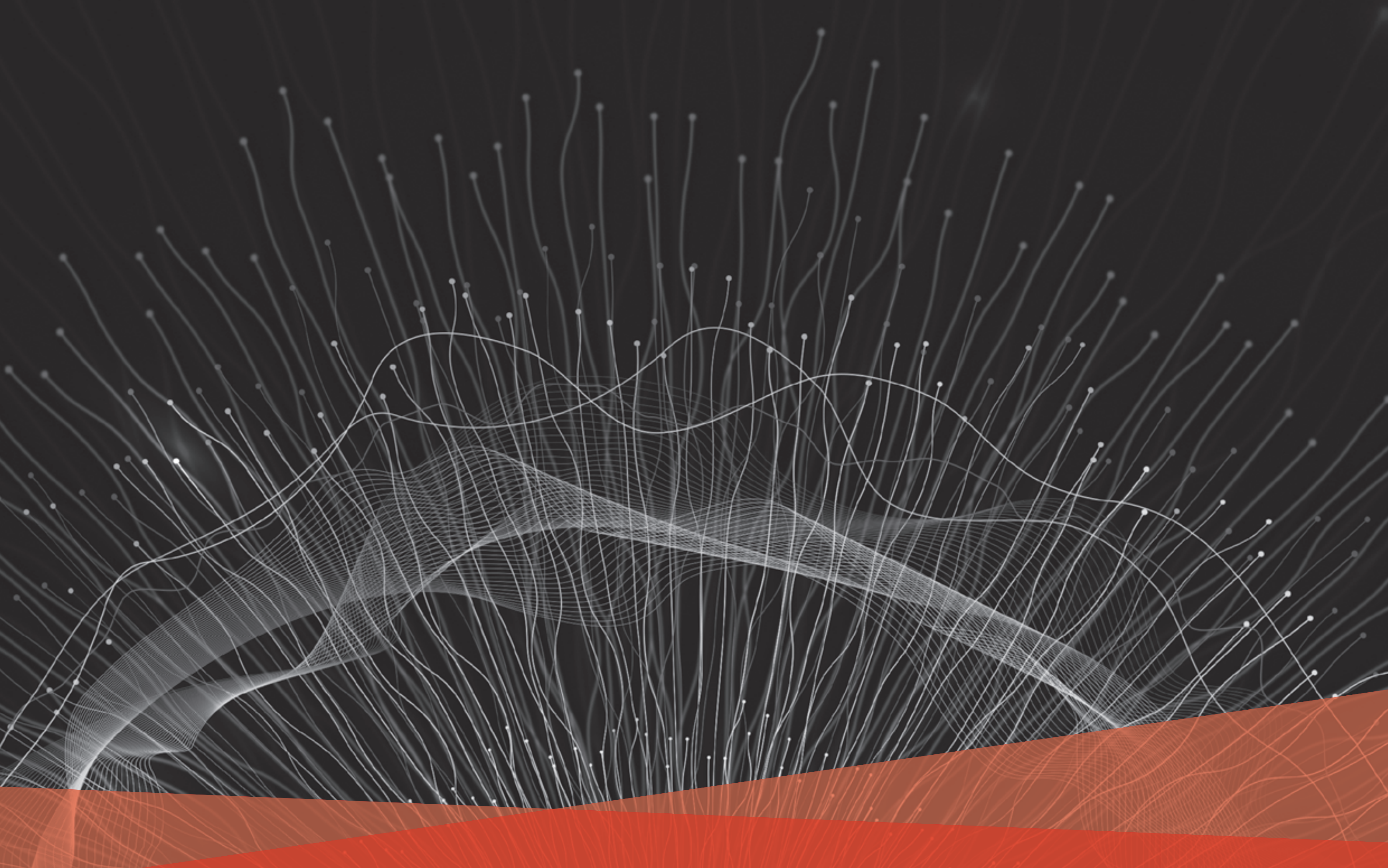


CYBERSECURITY FOR AI SYSTEMS: Protecting Data, Models, and Interfaces

CENSUS
Cybersecurity Engineering

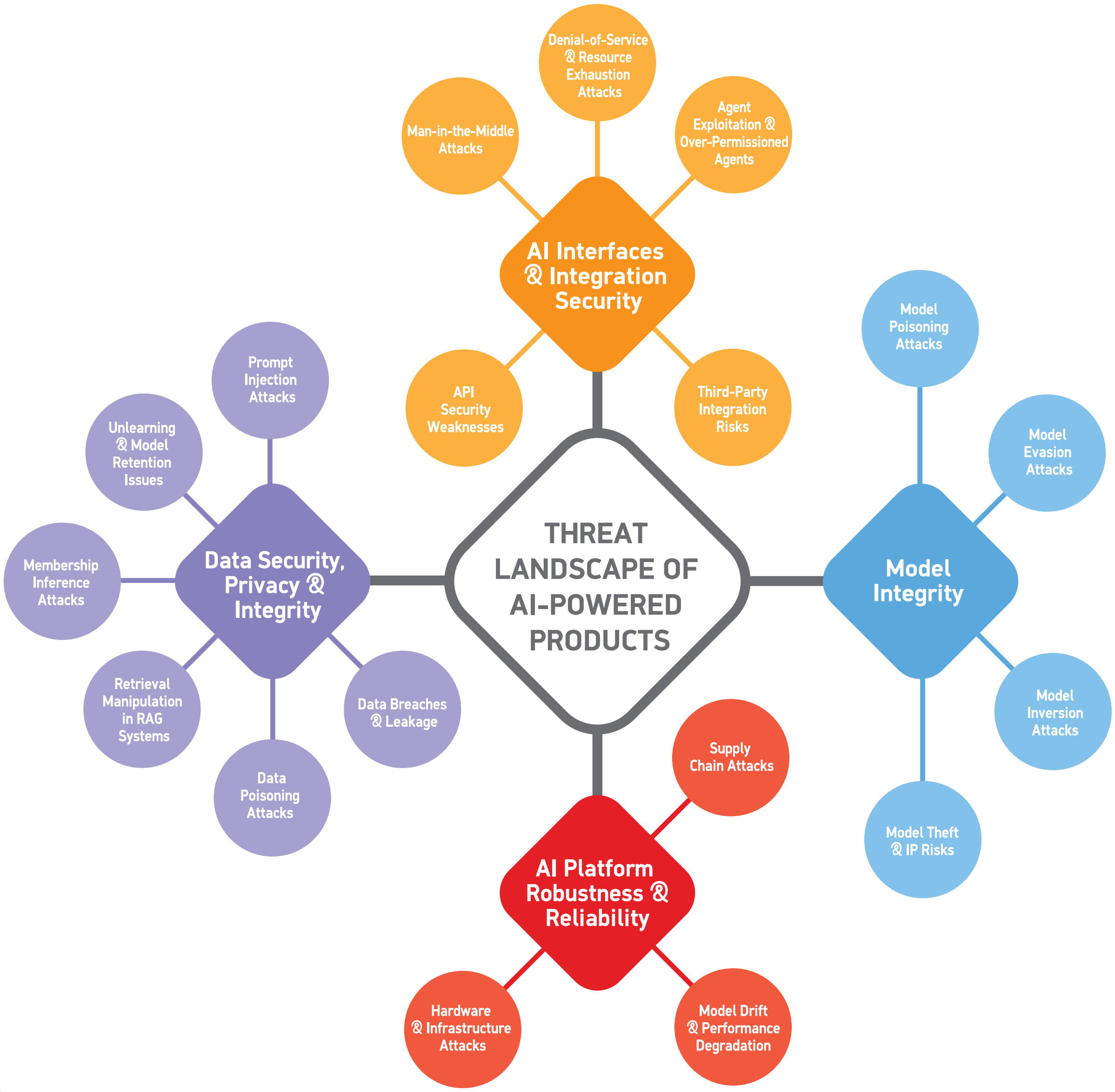
TABLE OF CONTENTS

1. CYBERSECURITY LANDSCAPE OF AI-POWERED PRODUCTS

Artificial Intelligence (AI) is revolutionizing industries, driving rapid advancements and enabling products to operate with greater intelligence, autonomy, and interactivity. From natural language processing and image recognition to predictive analytics and autonomous decision-making, AI is now deeply integrated into modern products, automating and enhancing tasks traditionally performed by humans.

However, as AI systems become more embedded into critical workflows, new security challenges emerge. These technologies expand attack surfaces, introduce novel vulnerabilities, and create new attack vectors for malicious actors. AI-driven products rely on vast data inputs, external interactions, and dynamic learning mechanisms, making data integrity, confidentiality, and access control critical concerns. Generative AI introduces additional complexities, as it depends on large-scale datasets, intricate computational frameworks, and evolving regulatory mandates, requiring robust security measures for data protection, model reliability, and compliance.

Recognizing these evolving threats, CENSUS has categorized AI cybersecurity risks into four key pillars, each addressing critical attack vectors and vulnerabilities. These pillars serve as the foundation of our security engineering services, enabling a structured and proactive approach to securing AI-powered products while ensuring resilience, compliance, and trustworthiness in an increasingly complex digital ecosystem.


THREAT LANDSCAPE OF AI-POWERED PRODUCTS
AI Interfaces & Integration Security
Man-in-the-Middle Attacks
Denial-of-Service & Resource Exhaustion Attacks
Agent Exploitation & Over-Permissioned Agents
API Security Weaknesses
Third-Party Integration Risks
Model Integrity
Model Poisoning Attacks
Model Evasion Attacks
Model Inversion Attacks
Model Theft & IP Risks
Data Security, Privacy & Integrity
Prompt Injection Attacks
Unlearning & Model Retention Issues
Membership Inference Attacks
Retrieval Manipulation in RAG Systems
Data Poisoning Attacks
Data Breaches & Leakage
AI Platform Robustness & Reliability
Supply Chain Attacks
Hardware & Infrastructure Attacks
Model Drift & Performance Degradation

1.1. Data Security, Privacy and Integrity

AI security extends beyond traditional data protection mechanisms, as AI systems treat data as a foundational asset—not just information exchanged between components, but the core of learning, inference, and decision-making. This data-centric nature of AI creates persistent security risks across its entire lifecycle, from training and fine-tuning to inference and deployment.

- **Data Breaches and Leakage** - Unauthorized access or exfiltration of sensitive training data or inference outputs can expose proprietary or personal information. AI products may inadvertently retain, expose, or leak confidential data due to poor sanitization, weak access controls, or excessive output exposure.
- **Data Poisoning Attacks** - Attackers inject manipulated or biased data into the training pipeline, corrupting model behavior and leading to misclassifications, bias reinforcement, or backdoor exploits. These attacks pose severe risks in safety-critical applications such as autonomous vehicles, finance, and healthcare.
- **Retrieval Manipulation in RAG Systems** - RAG models depend on external knowledge sources, making them vulnerable to adversarial data poisoning. Attackers can inject deceptive, biased, or harmful information into indexed databases, leading to misleading AI-generated outputs.
- **Membership Inference Attacks** - Attackers attempt to determine whether specific data points were used in training, posing privacy risks in regulated industries such as healthcare, finance, and legal applications. These attacks can lead to re-identification, inference-based privacy violations, or regulatory non-compliance.
- **Unlearning and Model Retention Issues** - AI models struggle to forget previously learned data, even when deletion requests are enforced. This creates compliance risks, particularly under GDPR's right to be forgotten, where organizations may be unable to fully erase sensitive data from an AI system's decision-making process.
- **Prompt Injection Attacks** - Attackers manipulate AI prompts to coerce models into disclosing sensitive information or executing unintended actions. Indirect prompt injection occurs when adversarial commands are embedded in external data sources, such as compromised web content or API responses, leading to malicious execution within AI agents.

1.2. Model Integrity

The integrity of AI models is essential to maintaining trust, reliability, and security. A compromised model can result in erroneous outputs, flawed decision-making, and systemic vulnerabilities, leading to catastrophic failures in critical applications.

- **Model Poisoning Attacks** - Adversaries inject maliciously crafted data during training or fine-tuning, corrupting the learning process and causing incorrect or adversary-controlled predictions. In adaptive AI models, attackers can gradually manipulate behavior over time, influencing financial models, security systems, or autonomous decision-making.
- **Model Evasion Attacks** - Attackers craft subtle input modifications to mislead AI models, forcing misclassifications or manipulated responses. These attacks can bypass AI-driven security defenses, interfere with image recognition, and create self-reinforcing vulnerabilities in self-learning AI agents.
- **Model Inversion Attacks** - Attackers reverse-engineer AI models to extract sensitive training data, posing severe privacy risks for products that handle proprietary datasets, financial transactions, or medical records.
- **Model Theft and IP Risks** - Threat actors attempt to steal, replicate, or extract proprietary AI models via query-based model extraction, API abuse, or insider threats. In agentic AI systems, attackers can gradually extract model parameters, training logic, or decision heuristics, leading to intellectual property theft, unauthorized AI replication, and competitive disadvantages.

1.3. AI Interfaces and Integration Security

The integration of AI into broader systems exposes attack surfaces at APIs, external services, and data exchange points. Agentic AI architectures, which autonomously interact with external environments, further increase risks, necessitating robust security controls to prevent unauthorized execution and data exposure.

- **API Security Weaknesses** - Poorly secured APIs can allow unauthorized access to AI functions, enabling attackers to extract sensitive data, manipulate behavior, or exploit model vulnerabilities.
- **Man-in-the-Middle Attacks** - Attackers intercept and alter data exchanges between AI models, cloud services, and edge devices, compromising confidentiality and trust.
- **Denial-of-Service & Resource Exhaustion Attacks** - AI products require high-performance computing resources, making them vulnerable to DoS attacks that degrade performance or disrupt real-time AI operations.
- **Agent Exploitation & Over-Permissioned Agents** - Attackers exploit vulnerabilities in agentic AI decision-making forcing unintended actions. Weak safeguards around permissions enable malicious API chaining, tricking the agent into executing a sequence of attacker-defined tasks.
- **Third-Party Integration Risks** - AI often relies on third-party APIs, plugins, and external data sources, making supply chain security essential to prevent backdoors, data poisoning, or exfiltration risks.

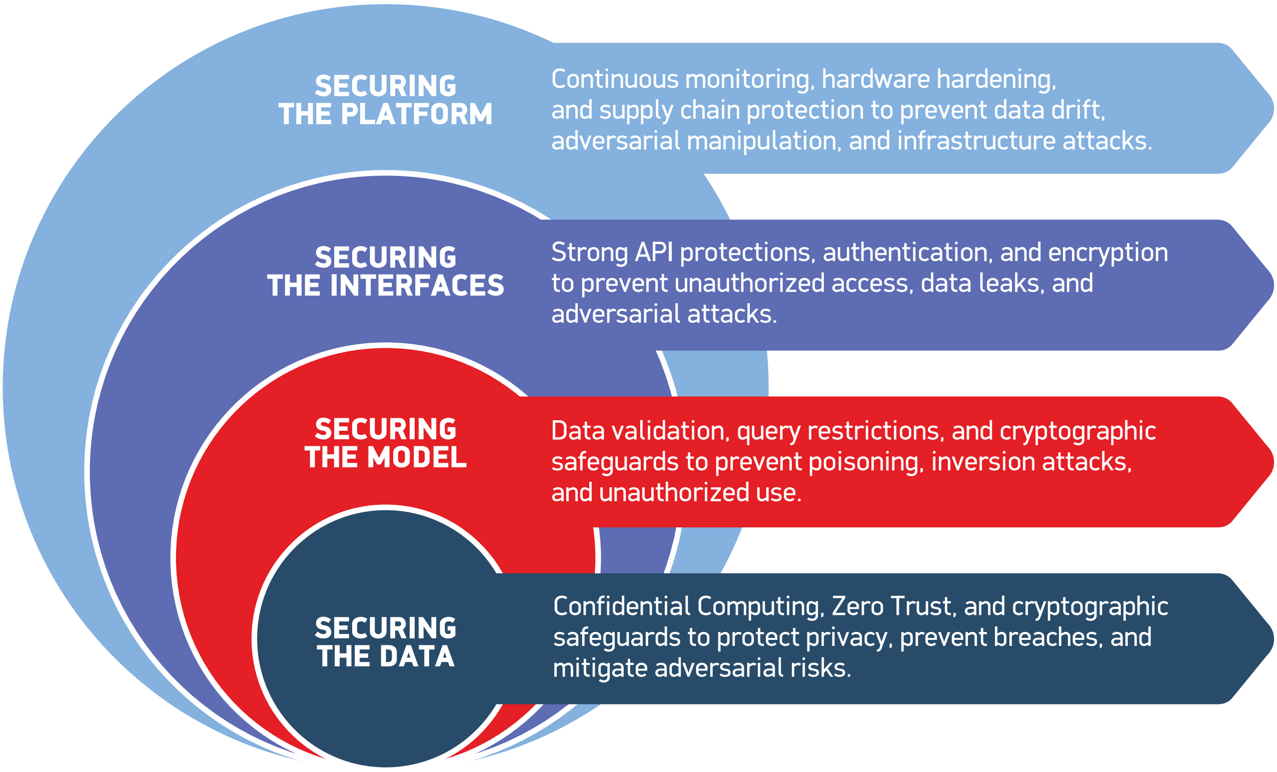
1.4. AI Platform Robustness and Reliability

AI platform resilience is critical for maintaining availability, accuracy, and security in adversarial environments. As AI systems evolve, ensuring robust defenses against performance degradation, model drift, and infrastructure vulnerabilities is paramount.

- **Model Drift and Performance Degradation** - AI models degrade over time as data distributions shift, leading to reduced accuracy. Attackers can intentionally manipulate drift to alter model behavior without immediate detection.
- **Hardware and Infrastructure Attacks** - AI relies on specialized computing environments (e.g., GPUs, TPUs, IoT devices), making them vulnerable to firmware exploits, misconfigurations, and adversarial payload injections.
- **Supply Chain Attacks** - AI depends on third-party libraries, pre-trained models, and external frameworks, making supply chain security crucial to prevent malicious dependencies, poisoned datasets, or covert backdoors.

2. MULTI-LAYERED AI SECURITY: A STRATEGIC APPROACH TO RESILIENCE

As AI technology becomes deeply embedded across industries, it introduces new security risks that extend beyond traditional cybersecurity challenges. Ensuring compliance, resilience, and trustworthiness requires organizations to adopt a proactive, multi-layered security approach that embeds robust protections at every level of the AI architecture, from data security and model integrity to secure deployment pipelines and API defenses.



SECURING THE PLATFORM

Continuous monitoring, hardware hardening, and supply chain protection to prevent data drift, adversarial manipulation, and infrastructure attacks.

SECURING THE INTERFACES

Strong API protections, authentication, and encryption to prevent unauthorized access, data leaks, and adversarial attacks.

SECURING THE MODEL

Data validation, query restrictions, and cryptographic safeguards to prevent poisoning, inversion attacks, and unauthorized use.

SECURING THE DATA

Confidential Computing, Zero Trust, and cryptographic safeguards to protect privacy, prevent breaches, and mitigate adversarial risks.

> “By adopting defense-in-depth principles and zero-trust security models, organizations can proactively mitigate evolving cyber threats and establish a strong foundation for secure, scalable, and resilient AI deployments across enterprise, cloud, edge, and mixed criticality environments.”

2.1. Securing the Data

A **data-centric security approach** is essential to protect confidentiality, integrity, and privacy throughout the AI lifecycle. **Confidential Computing** safeguards data-in-use during training and inference using trusted execution environments, secure virtualization, and GPU attestation to prevent unauthorized access. **Zero Trust Architecture** enforces continuous authentication and strict access controls, ensuring only verified entities can interact with sensitive datasets.

AI solutions leveraging retrieval-augmented generation introduce external data risks, making cryptographic verification and content filtering critical in preventing adversarial manipulation. Federated learning reduces reliance on centralized storage, limiting exposure to large-scale data breaches, while differential privacy techniques ensure models do not memorize individual training data, mitigating membership inference and re-identification risks.

2.2. Securing the Model

Ensuring AI model integrity is vital to maintaining trust, accuracy, and resilience against adversarial attacks. Robust data validation prevents model poisoning by auditing and sanitizing training datasets, eliminating maliciously manipulated inputs. Query restrictions mitigate model inversion attacks, preventing adversaries from extracting sensitive parameters or training insights.

To protect intellectual property, model watermarking embeds cryptographic markers, enabling detection of unauthorized use, theft, or tampering. Secure deployment ensures AI models remain unaltered through code signing, encryption, and runtime integrity verification.

2.3. Securing the Interfaces

AI-powered products heavily rely on APIs to connect with external systems, making interface security a critical concern. APIs represent high-risk attack surfaces, where unauthorized access, data leaks, and adversarial prompt injections can compromise AI operations.

To mitigate these risks, organizations must implement multi-factor authentication (MFA), granular access control, and end-to-end encryption. As cyber threats evolve, continuous authentication, API rate limiting, and protocol hardening become essential to defend against API hijacking, adversarial queries, and unauthorized model access.

A defense-in-depth strategy, integrating **Trustworthy Computing** principles, cryptographic key management, and AI-specific security controls, ensures secure, scalable, and resilient AI deployments.

2.4. Securing the Platform

AI systems must integrate continuous monitoring and adaptive retraining to mitigate the impact of data drift, preventing accuracy degradation and manipulation by adversaries. Hardware security hardening, firmware integrity verification, and host attestation strengthen AI processing environments, protecting against external and internal threats.

Since AI systems depend on external dependencies, securing the software supply chain is crucial. Cryptographic integrity checks, dependency validation, and origin tracking help detect malicious package injections, tampered datasets, or adversarial backdoors that could compromise AI operations.

3. CENSUS CYBERSECURITY ENGINEERING CAPABILITIES

CENSUS delivers comprehensive cybersecurity solutions tailored to AI-powered products. Our services ensure end-to-end protection, secure AI architectures, and enhance data pipelines with evolving security frameworks.

3.1. Security by Design

Security must be embedded from inception, ensuring AI-powered products integrate robust protections at every stage. Security architecture reviews, adversarial resilience assessments, and cryptographic safeguards form the foundation of hardened AI products.

To reduce security risks, AI solutions undergo comprehensive attack surface profiling to identify applicable threats. CENSUS employs an engineering-driven threat modeling approach, analyzing AI-specific threats such as data poisoning, prompt injection, and RAG retrieval manipulation. By leveraging structured security domain abstractions, organizations can apply targeted security controls, enforce zero trust architectures, and implement defense-in-depth strategies.

3.2. End-to-End Security Assessment

CENSUS Security Posture Assessment (SPA) provides a 360-degree evaluation of an AI-powered system's security readiness, resilience, and compliance. This assessment covers all AI layers—from model development and data pipelines to APIs, infrastructure, and runtime environments.

Our SPA service identifies vulnerabilities, verifies security controls, and validates resilience against real-world adversarial attacks. It includes targeted testing such as model inversion resistance evaluations, membership inference risk assessments, and API resilience validation. Through automated and manual testing, product-specific security strategies, and engineering-driven validation techniques, SPA ensures AI vulnerabilities are identified, prioritized, and mitigated.

3.3. Applied Research and Future-Ready Solutions

CENSUS Applied Research anticipates and mitigates emerging AI security threats, exploring next-generation security solutions through cutting-edge cryptography and secure systems. By leveraging post-quantum cryptography, confidential computing, and federated security models, we future-proof AI solutions against evolving cyber threats. Innovations in hardware-backed AI attestation, isolation, and privacy preserving technologies, strengthen the security of autonomous and learning-based systems.

3.4. Security Foundations Consulting

CENSUS translates high-level security strategies into engineering-driven solutions, ensuring AI-powered products align with best practices and regulatory compliance. We develop modular security architecture blueprints, securing AI products through adversarial robustness, trusted execution, and privacy-preserving encryption techniques. Data isolation, multi-tenancy security, and federated learning safeguards further enhance the product's maturity.

4. CENSUS - YOUR TRUSTED PARTNER IN SECURING AI PRODUCTS

CENSUS is a trusted cybersecurity engineering leader, providing cutting-edge security solutions for AI-powered products across enterprise, cloud, edge, and autonomous systems.

4.1. Engineering-Driven Innovation

CENSUS integrates confidential computing, adversarial defense techniques, and cryptographic safeguards to protect AI products and data pipelines. Our solutions future-proof AI applications against intellectual property theft, adversarial exploitation, and unauthorized model replication.

4.2. Unparalleled Domain Expertise

We specialize in AI-specific cybersecurity, ensuring secure data pipelines, model integrity, and adversarial attack resilience. Our expertise spans data privacy, adversarial defense mechanisms, and AI risk management, aligning with global security and compliance frameworks.

4.3. Comprehensive End-to-End Support

CENSUS provides comprehensive, **end-to-end cybersecurity support** throughout the entire product lifecycle, from conceptualization and design to deployment and maintenance. With a focus on lifecycle security management, we implement strategies for continuous security validation, adversarial robustness testing, and adaptive defenses.

4.4. Value-Driven Partner for AI Security

CENSUS fosters a collaborative engagement model, seamlessly integrating with client teams to provide expert security guidance **without disrupting workflows.** Our solutions are designed to address immediate security needs while ensuring long-term adaptability to future AI advancements and evolving threat landscapes. We **support organizations accelerate secure AI adoption** while maintaining compliance, transparency, and trustworthiness. With a proven track record in AI security research, security testing, and applied engineering, CENSUS consistently delivers precision, innovation, and quality.

> “
>
> Choosing CENSUS for secure AI adoption means **partnering with a cybersecurity leader** that drives **innovation,** ensures **compliance,** and builds **resilient, future-ready AI systems** that meet the demands of today and tomorrow.

USA
607 Boylston Street,
Suite 165L, Boston, MA 02116
T. +1 8882 029 846
E. usa@census-labs.com

www.census-labs.com

UK
Unit 2.15 Barley Mow Centre
10 Barley Mow Passage
W4 4PH, London
T. +44 1293 324 069
E. uk@census-labs.com

EUROPE
128 A. Siggrou Ave.,
Athens 117 45, Greece
T. +30 210 220 8989-90
E. eu@census-labs.com

ABU DHABI
Office 110 - First Floor
Incubator Building
Masdar City, Abu Dhabi
United Arab Emirates
T. +971 8003 110 047
E. uae@census-labs.com

DUBAI
Unit GA-00-SZ-L1-RT-201, Level 1
Gate Avenue - South Zone
Dubai International Financial Centre
Dubai, United Arab Emirates
T. +971 8003 110 047
E. uae@census-labs.com