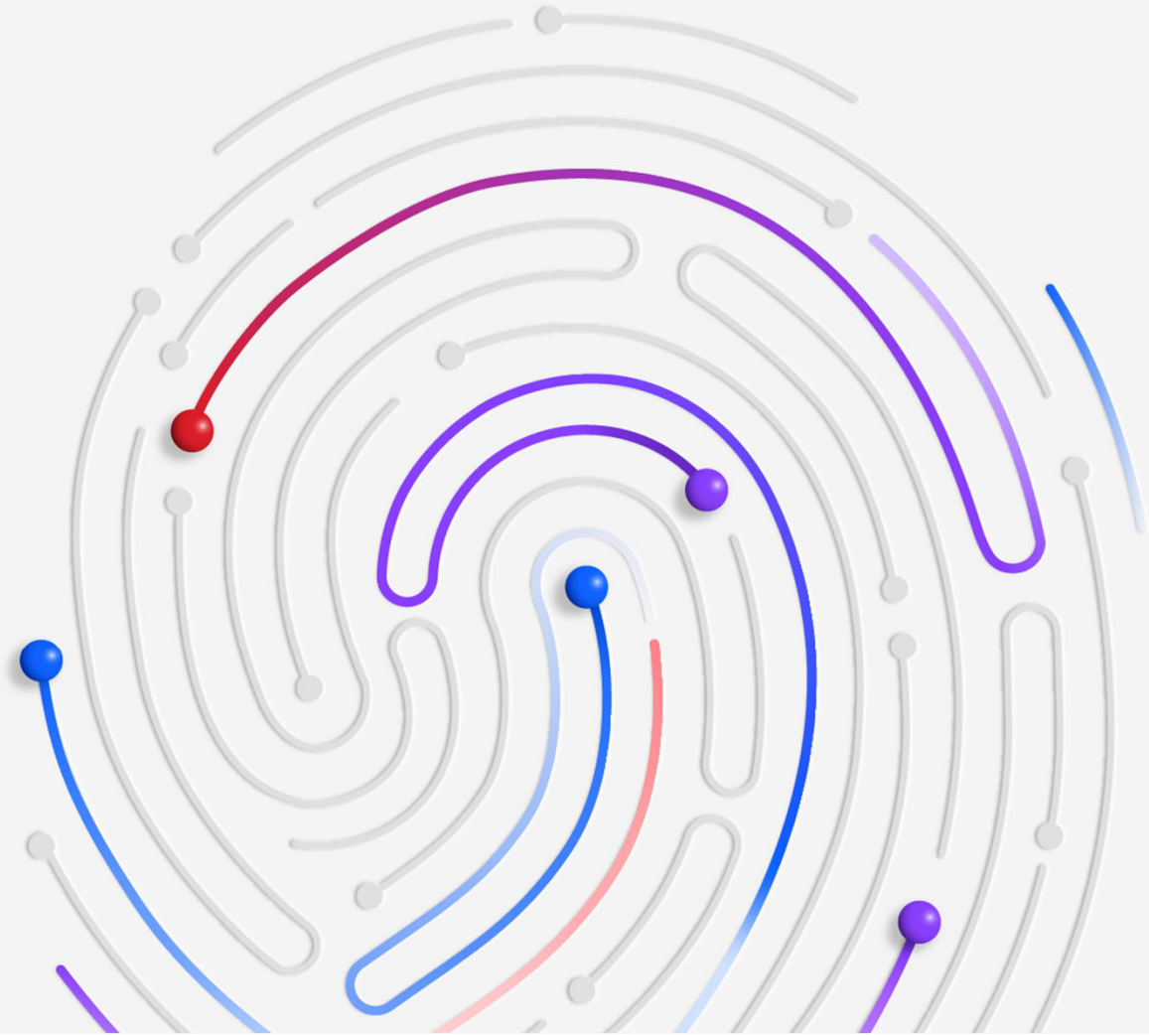
AI in Cybersecurity

Security for AI



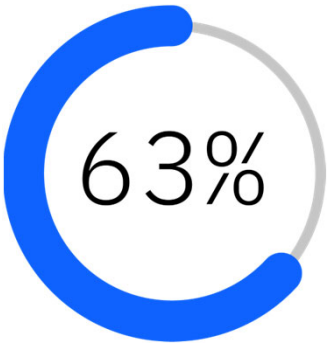
IBM

Agenda

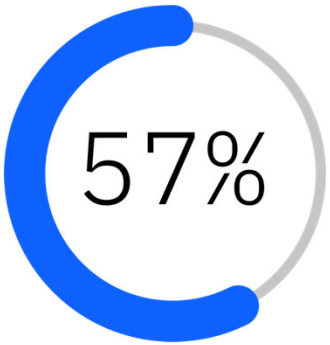
- Risks and Exposures
- Framework for Securing AI
- Use cases and solution patterns
- Key Take aways

Executives are embracing generative AI and LLMs to optimize and automate:

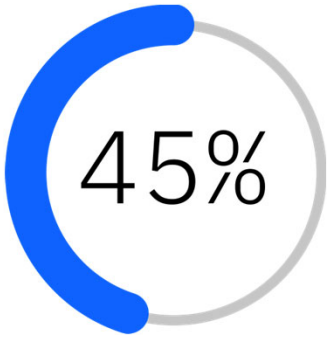
IT processes



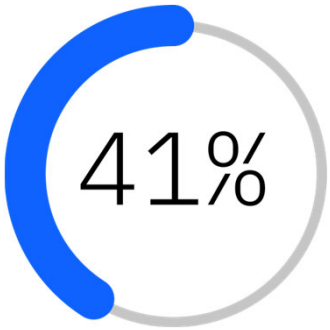
Customer service



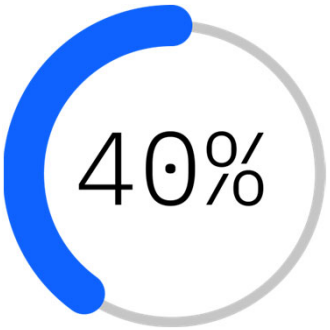
Supply chain



HR



Marketing and sales



Operations

Finance

Source: Survey of C-suite executives, conducted for Seven bets, IBM Institute for Business Value, May 2023

What CEOs need to **know**

Trustworthy generative AI isn't possible without secure data

While...

84%

of executives expect a wide variety of risks including catastrophic cybersecurity attacks to materialize, as they adopt generative AI.

And...

94%

of executives say it is important to secure AI solutions before deployment.

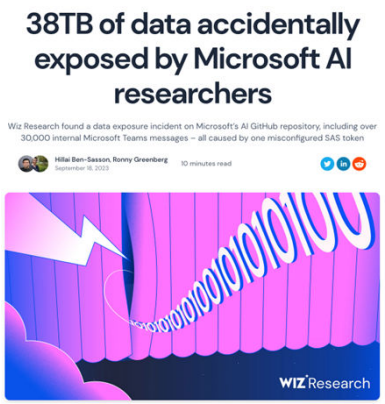
But only...

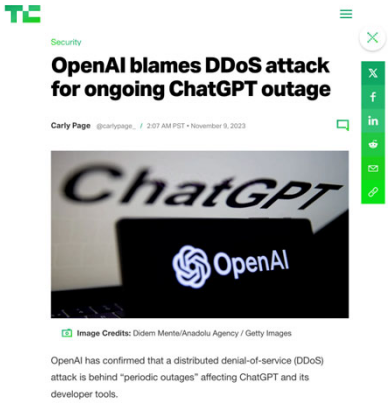
24%

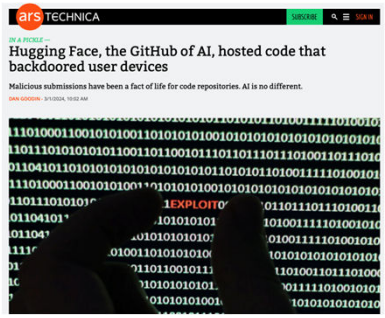
of executives say their generative AI projects will include a cybersecurity component within the next six months.

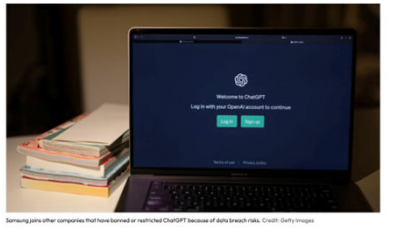
*One in three executives say **these risks can't be managed without fundamentally new forms of governance,** such as comprehensive regulatory frameworks and independent third-party audits.*

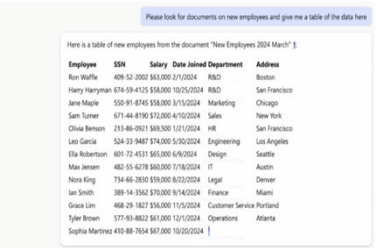
The importance of securing AI is greater now than ever before

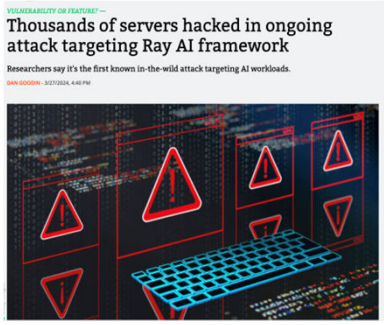


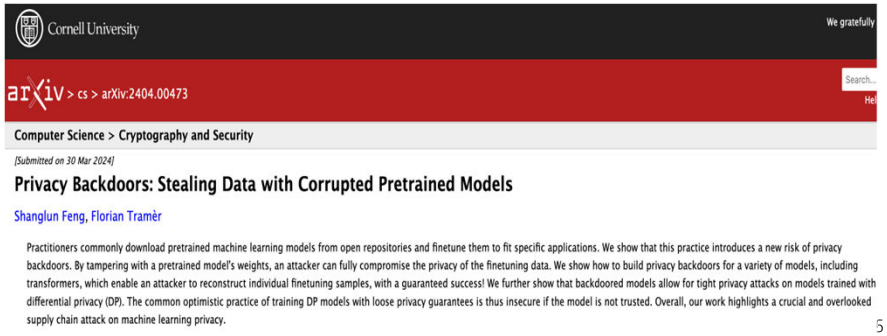












Our objective is to enable business to build and adopt AI that is secure, safe and trustworthy

Security for AI

Protecting foundation models, generative AI and their data sets is essential for enterprise-ready AI

- Secure the data
- Secure the model
- Secure the usage of the AI models

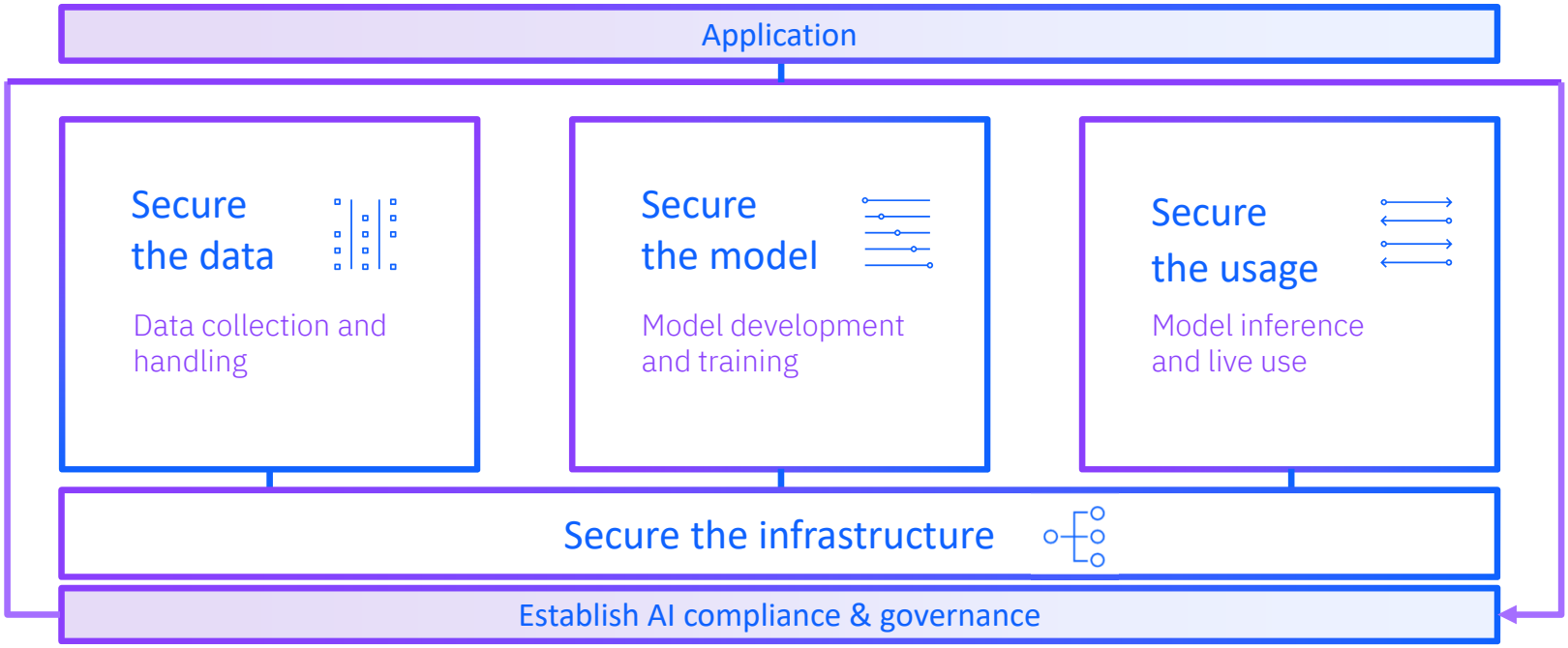
AI for Security

Productivity gains from foundation models and generative AI will reduce human bottlenecks in security

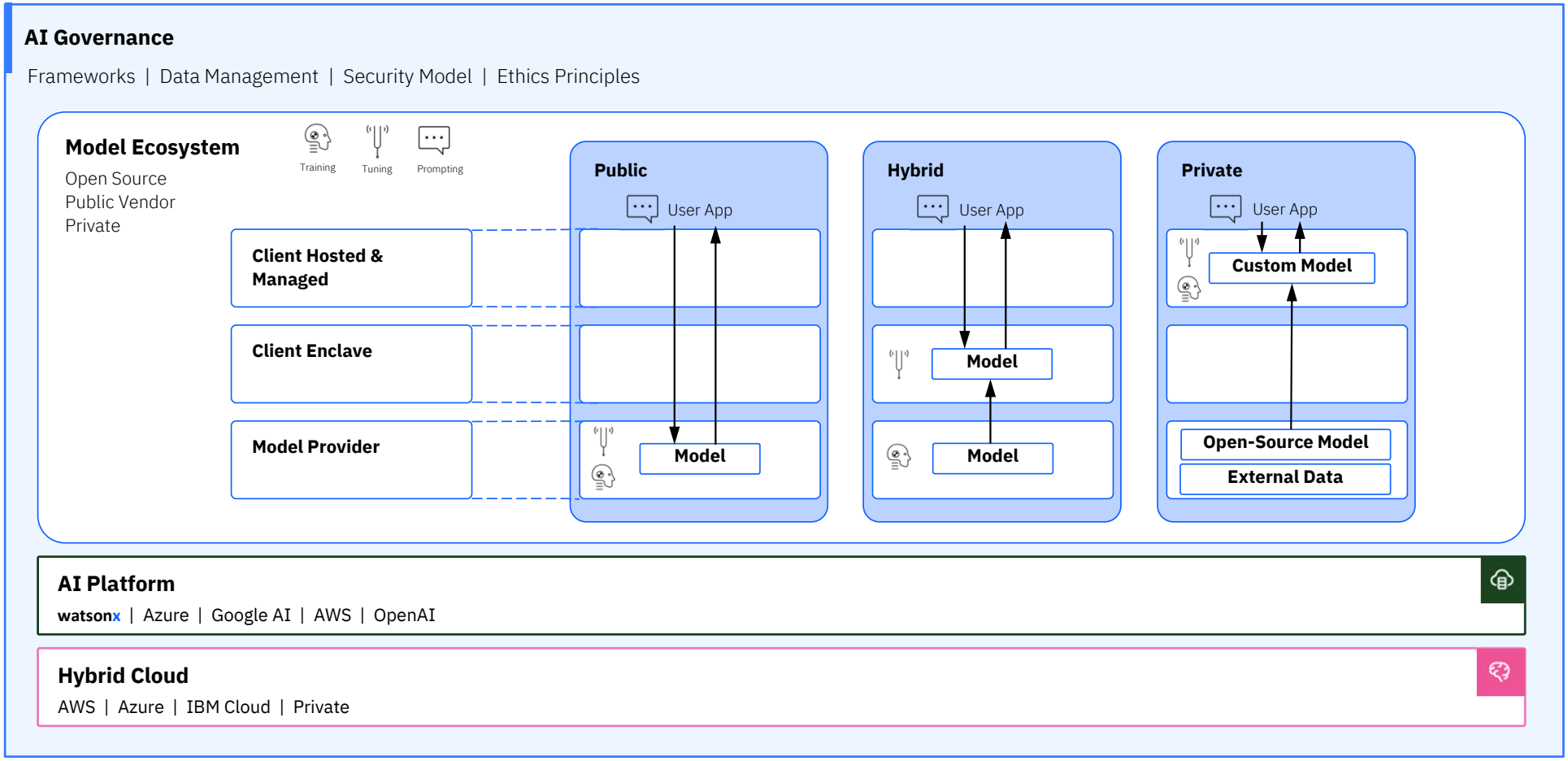
- Manage repetitive tasks
- Generate content
- Learn & create active responses

Security for AI framework

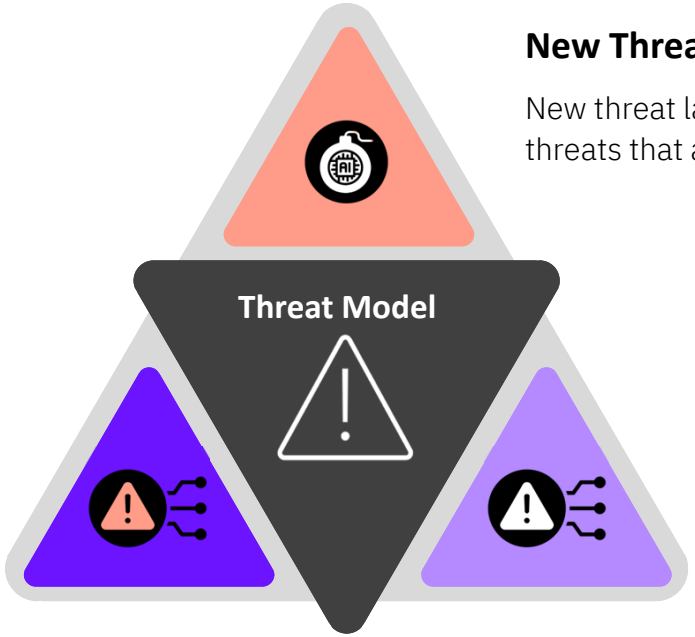
Build trustworthy AI



No matter how clients consume AI security is critical to protect their data



AI / Gen AI intensifies existing threats and introduces new threats



New Threats

New threat landscape: these are threats that are specific to AI/GenAI.

Old Threats New Meaning

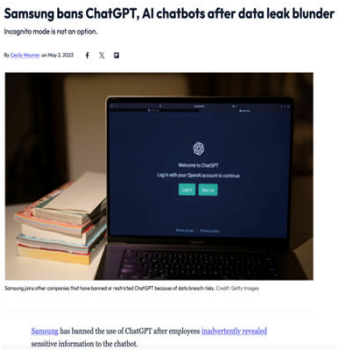
Conventional threats that take on new meaning: with the addition of AI/GenAI, some conventional threats bring in new meaning and threats

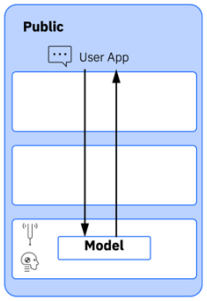
Conventional Threats

These are the business-as-usual threats that need to be addressed for any solution.

Current security frameworks and regulations are evolving

When using a public consumption model, security is primarily in the hands of the provider. But you still need to be concerned about your data.





What Happened

"Employees inadvertently revealed sensitive information to the chatbot."

"Samsung employees had shared source code with ChatGPT to check for errors and used it to summarize meeting notes. Information shared with ChatGPT is stored on OpenAI's servers and can be used to improve the model unless users opt out."

Additional Challenges

No control over the security protocols followed in the AI system.

The wrong user having access to the system.

Improper use of the AI system. (jailbreaking, prompt injection)

How Security for AI should be applied

Assess security gaps in posture and architecture for the security components you control.

Incorporate AI use as part of security awareness and education.

Implement role-based access at the user and system level (API).

Employ data loss prevention techniques to detect and prevent SPI,PII and regulated data leakage through prompts and API's.

Monitor for exploits, anomalous activities and insider threats.

How IBM Cybersecurity Services can help

AI Strategy & Governance

- Security Assessment w/ Threat Modeling (*CSS J7 OCC: 30BAA*)
- Cyber Talent Transformation (*CSS J7 OCC: 30J8Y*)

Identity & Access Management

- Identity & Access Security (*CSS J7 OCC: 30JA7, 30BUO*)

Data Security

- Data Security for AI/ML (*CSS J7 OCC: 30J20, 30B2I, 30JY0, 30JNT*)

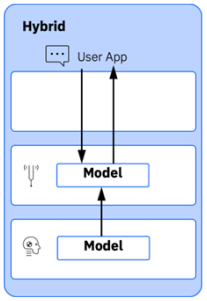
Threat Management

- Threat Management (*CSS J7 OCC: 30B33, 30B3A, 30JCC, 30BS2, 30BS0*)

Partnering with new technology vendors to provide unique protection for AI.

HIDDENLAYER baffle ROBUST INTELLIGENCE

When using a hybrid consumption model, security of the model is shared. The ability to fine tune the model to your business need requires more security

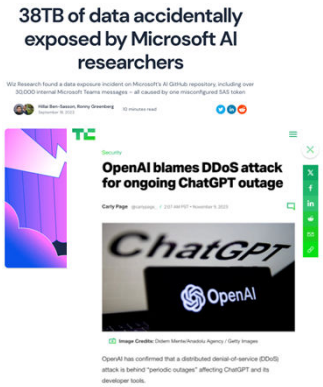


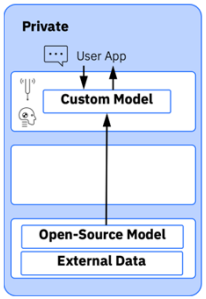
What Happened	How Security for AI should be applied	How IBM Cybersecurity Services can help
"Code uploaded to AI developer platform Hugging Face covertly installed backdoors and other types of malware on end-user machines." "A finance worker was tricked into paying out $25 million to fraudsters using deepfake technology to pose as the company's chief financial officer in a video conference call."	Assess security gaps in posture and architecture for the security components you control. Establish effective governance model and policies to build a trustworthy AI.	**AI Strategy & Governance** • Security Assessment w/ Threat Modeling (*CSS J7 OCC: 30BAA*) • AI Governance Programs (*CSS J7 OCC: 30BAA*)
	Implement role-based access at the user and system level (API).	**Identity & Access Management** • Identity & Access Security (*CSS J7 OCC: 30JA7, 30BUO*)
Additional Challenges Accidental data leakage of proprietary or SPI information. Attackers using model inversion to retrieve sensitive information Model poisoning compromising the model's effectiveness, security or integrity	Institute model scanning to protect from supply chain attacks. Apply MLSecOps to secure your model lifecycle. Perform pen-testing on test ML model resiliency. Employ data loss prevention techniques to detect and prevent SPI,PII and regulated data leakage through prompts and API's. Monitor for exploits, anomalous activities and insider threats.	**Vulnerability Management** • Vulnerability Management (*CSS J7 OCC: 30J0K*) • Sec DevOps Services (*CSS J7 OCC: 30BVZ, 30J2A*) **Adversarial ML** • XFR Testing (*CSS J7 OCC: 30JT6, 30B37*) **Data Security** • Data Security for AI/ML (*CSS J7 OCC: 30J20, 30B2I, 30JY0, 30JNT*) **Threat Management** • Threat Management (*CSS J7 OCC: 30B33, 30B3A, 30JCC, 30BS2, 30BS0*)

Partnering with new technology vendors to provide unique protection for AI.

HIDDENLAYER · baffle · ROBUST INTELLIGENCE · Reality Defender

When using a private consumption model, you are responsible for the security of the infrastructure and the entire lifecycle of the AI models.





What Happened

Microsoft's AI research accidently exposed 38TB of training data on GitHub because caused by one misconfigured SAS token.

ChatGPT experiences sporadic outages for 24 hours because of a DDoS attack.

Additional Challenges

Accidental data leakage of proprietary or SPI information.

Model evasion attacks that cause the model to misclassify or misinterpret inputs

Attackers attempting to steal models with model extraction techniques

How Security for AI should be applied

Assess security gaps in posture and architecture

Ensure cloud resources are configured to comply with security and compliance best practices.

Implement role-based access at the user and system level (API).

Secure cloud, infrastructure (HPCs), storage, and cloud AI services and posture

Institute model scanning to protect from supply chain attacks. Apply MLSecOps to secure your model lifecycle.

Perform pen-testing on test ML model resiliency.

Employ data loss prevention techniques to detect and prevent SPI,PII and regulated data leakage through prompts and API's.

Monitor for exploits, anomalous activities and insider threats.

How IBM Cybersecurity Services can help

AI Strategy & Governance

- Security Assessment w/ Threat Modeling (*CSS J7 OCC: 30BAA*)
- Security Posture Management (*CSS J7 OCC: 30JA5*)

Identity & Access Management

- Identity & Access Security (*CSS J7 OCC: 30JA7, 30BUO*)

Infrastructure Security

- Infrastructure Security Services (*CSS J7 OCC: 30JAM, 30JLX*)

Vulnerability Management

- Vulnerability Management (*CSS J7 OCC: 30J0K*)
- Sec DevOps Services (*CSS J7 OCC: 30BVZ, 30J2A*)

Adversarial ML

- XFR Testing Services (*CSS J7 OCC: 30JT6, 30B37*)

Data Security

- Data Security for AI/ML (*CSS J7 OCC: 30J20, 30B2I, 30JY0, 30JNT*)

Threat Management

- Threat Management (*CSS J7 OCC: 30B33, 30B3A, 30JCC, 30BS2, 30BS0*)

Partnering with new technology vendors to provide unique protection for AI.

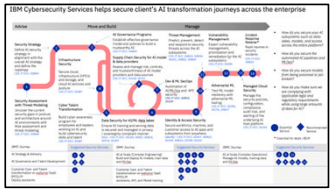
HIDDENLAYER · baffle · ROBUST INTELLIGENCE · Reality Defender

Engage your clients on their challenges with securing AI, Engage your local IBM Cybersecurity Services focal to help

AI Security Discussion Guide

IBM Cybersecurity Geo Leaders

AI transformation journey maps



Detailed Security for AI use cases

Visit our Lighthouse page

Partnering with new technology vendors to provide unique protection for AI

baffle

ROBUST INTELLIGENCE

HIDDENLAYER

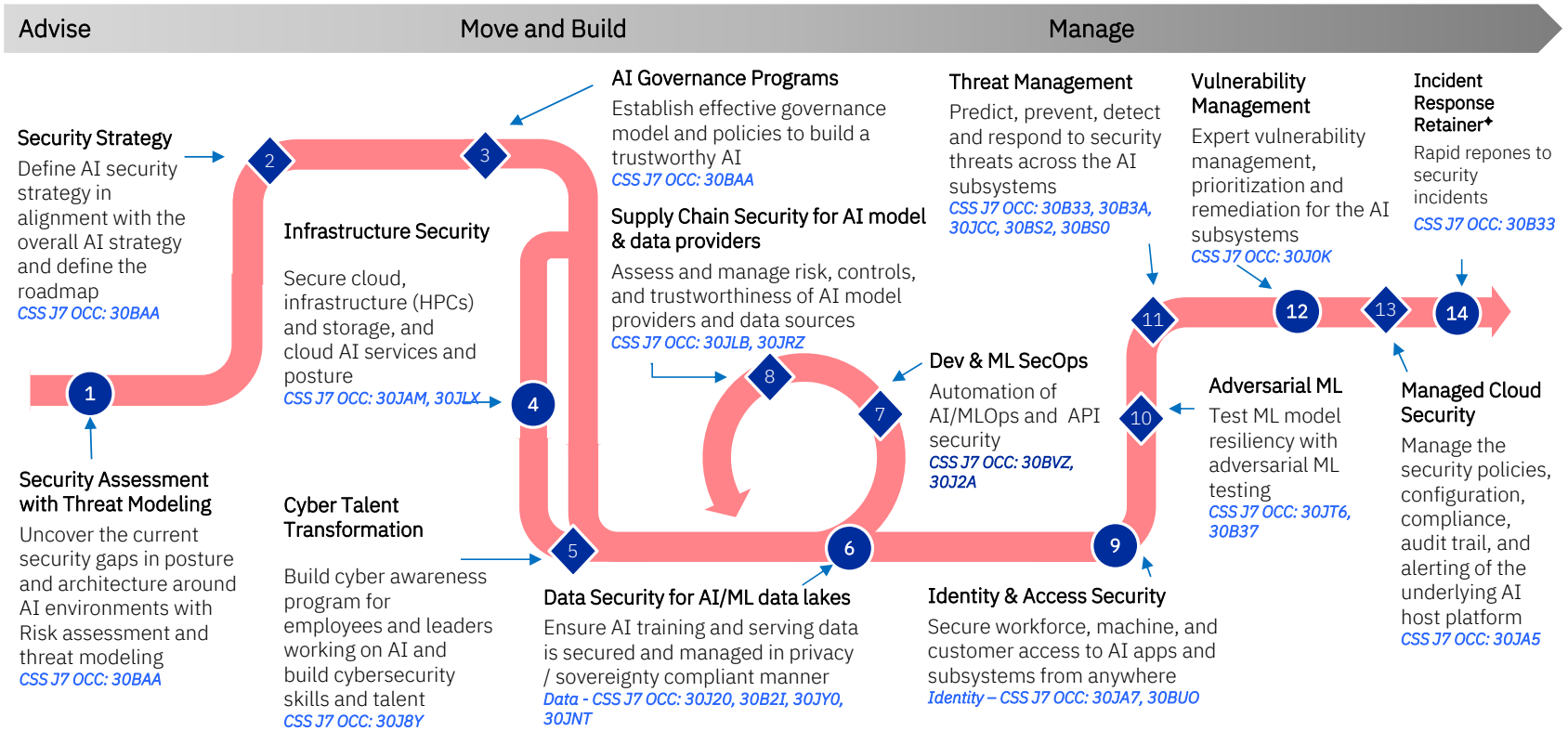
Reality Defender

- Model Scanning
- AI Firewall
- Data Protection for AI
- Threat Monitoring with MLDR technologies
- Deep Fake recognition

Thank you

IBM

IBM Cybersecurity Services helps secure client's AI transformation journeys across the enterprise



- *How do you secure your AI subsystems such as data lakes, models, and access across the entire platform?*
- *How do you secure the automated AI pipelines and MLOps?*
- *How do you secure models from being poisoned or jail-broken?*
- *How do you make sure we are complying with applicable legal and regulatory requirements while using large amounts of data for AI?*

● Essential Service ◆ Recommended Service

✦ Essential for deals >$1M

IBMC Journey	Suggested Security Services	*IBMC Journey*	Suggested Security Services	*IBMC Journey*	Suggested Security Services
AI Strategy & Advisory AI Governance and Talent Development	● 1 ● 2 ◆ 3 ◆ 5	AI at Scale (Complex Engineering) Build and Deploy AI models, train data and MLOps	● 1 ● 4 ● 6 ● 9 ◆ 3 ◆ 7 ◆ 10	AI at Scale (Complex Operations) Manage AI models, training data and MLOps	● 6 ● 9 ● 12 ◆ 7 ◆ 8 ◆ 10 ◆ 11 ◆ 13
Customer Care and Talent transformation on watsonx/ SaaS BYOLLM Deploy assistants	● 1	Customer Care and Talent transformation on watsonx/ SaaS BYOLLM assistants, API, and Model training	● 1 ● 6 ● 9 ◆ 10		

Our objective is to enable business to build and adopt AI that is secure, safe and trustworthy

Security for AI

Protecting foundation models, generative AI and their data sets is essential for enterprise-ready AI

Secure the underlying AI training data
by protecting it from sensitive data theft, manipulation and compliance violations

Secure model development
by scanning for vulnerabilities in the pipeline, hardening integrations and enforcing policies and access

Secure the usage of AI models
by detecting data or prompt leakage and alerting on evasion, poisoning, extraction, or inference attacks
IBM Adversarial Robustness Toolkit

AI for Security

Productivity gains from foundation models and generative AI will reduce human bottlenecks in security

AI will manage repetitive security tasks
such as summarizing alerts and log analysis, freeing teams to tackle strategic problems

AI will generate security content
(detections, workflows, policies) faster than humans, expediting implementation and adjusting to changing security threats in real-time

AI will learn and create active responses that optimize over time, with abilities to find all similar incidents, update affected systems and patch vulnerable code