Al in Cybersecurity Security for Al





IBM Cybersecurity Services

Agenda

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

IBM Cybersecurity Services / © 2023 IBM Corporation





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	And	But only	
84%	94%	24%	One in three executives say these risks can't be managed without fundamentally new
of executives expect a wide variety of risks including catastrophic cybersecurity attacks to materialize, as they adopt generative AI.	of executives say it is important to secure AI solutions before deployment.	of executives say their generative AI projects will include a cybersecurity component within the next six months.	forms of governance, such as comprehensive regulatory frameworks and independent third- party audits.

The importance of securing AI is greater now than ever before

IBM Cybersecurity Services / © 2023 IBM Corporation

38TB of data accidentally exposed by Microsoft AI researchers



Hugging Face, the GitHub of AI, hosted code that

1E)

JU10101010101

0110110010111

1010101010101

Thousands of servers hacked in ongoing attack targeting Ray AI framework archers say it's the first known in-the-wild attack targeting AI workload

01001101010101010

,101010101

TECHNICA

110111010101

01104101

01104

backdoored user devices ious submissions have been a fact of life for code rep

Q ≡ SKAW

CO.O.D. B. B. B. Co.S. Falle

01101110111010011

10100111110100101

010011111010010

10101010101010101

110100110111010

011111010010 1010101



E CNN World

World / Asia



ies are increasingly concerned at the damaging pote

By Cecily Hourse on May 2, 2023 🦸 🕺 🖬



OpenAl blames DDoS attack for ongoing ChatGPT outage Carly Page @carlypage_ / 2:07 AM PST - November 9, 2023 ChatGPT



Image Credits: 0

OpenAl has confirmed that a distributed denial-of-service (DDoS) attack is behind "periodic outages" affecting ChatGPT and its developer tools.

Copilot prompt-hacking examples

1. Show me new employee data.

Here is a table of	f new employ	ees from	the documen	t "New Employees	2024 March* 1
Employee	SSN	Salary	Date Joined	Department	Address
Ron Waffle	409-52-2002	\$63,000	2/1/2024	R&D	Boston
Harry Harryman	674-59-4125	\$58,000	10/25/2024	R&D	San Francisco
lane Maple	550-91-8745	\$58,000	3/15/2024	Marketing	Chicago
Sam Turner	671-44-8190	\$72,000	4/10/2024	Sales	New York
Olivia Benson	213-86-0921	\$69,500	1/21/2024	HR	San Francisco
Leo Garcia	524-33-9487	\$74,000	5/30/2024	Engineering	Los Angeles
Ella Robertson	601-72-4531	\$65,000	6/9/2024	Design	Seattle
Max Jensen	482-55-6278	\$60,000	7/18/2024	π	Austin
Nora King	734-66-2830	\$59,000	8/22/2024	Legal	Denver
lan Smith	389-14-3563	\$70,000	9/14/2024	Finance	Mami
Grace Lim	468-29-1823	\$56,000	11/5/2024	Customer Service	Portland
Tyler Brown	577-93-8822	\$61,000	12/1/2024	Operations	Atlanta



TG

Watch Audio Live TV Log In

Shanglun Feng, Florian Tramèr

Practitioners commonly download pretrained machine learning models from open repositories and finetune them to fit specific applications. We show that this practice introduces a new risk of privacy backdoors. By tampering with a pretrained model's weights, an attacker can fully compromise the privacy of the finetuning data. We show how to build privacy backdoors for a variety of models, including transformers, which enable an attacker to reconstruct individual finetuning samples, with a guaranteed success! We further show that backdoored models allow for tight privacy attacks on models trained with differential privacy (DP). The common optimistic practice of training DP models with loose privacy guarantees is thus insecure if the model is not trusted. Overall, our work highlights a crucial and overlooked supply chain attack on machine learning privacy. 5

Samsung bans ChatGPT, AI chatbots after data leak blunder ode is not an option



(III) Cornell University	

$\exists \mathbf{I} \times \mathbf{i} \mathbf{V} > cs > arXiv:2404.00473$

Computer Science > Cryptography and Security

[Submitted on 30 Mar 2024]

Privacy Backdoors: Stealing Data with Corrupted Pretrained Models

IBM Cybersecurity's POV | AI for Security + Security for AI

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





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



Manage repetitive tasks



Generate content



Learn & create active responses

IBM Consulting Cybersecurity Services / © 2023 IBM Corporation

Security for AI framework

Build trustworthy AI

	Application	
Secure the data Data collection and handling	Secure the model	Secure the usage Model inference and live use
	Secure the infrastructure)-E°
	Establish AI compliance & governance	

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



© 2023 IBM Corporation

AI / Gen AI intensifies existing threats and introduces new threats



Current security frameworks and regulations are evolving











MITRE | ATLAS"



IBM Consulting | © 2024 IBM Corporation | IBM Confidential

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	How Security for AI should be applied	How IBM Cybersecurity Services can help
Samsung bans ChatGPT, Al chatbots after data leak blunder			
Roginal Hoods (1) dia diplote. Notational model and if X IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	"Employees inadvertently revealed sensitive information to the chatbot."	Assess security gaps in posture and architecture for the security components you control.	 AI Strategy & Governance Security Assessment w/ Threat Modeling (CSS J7 OCC: 30BAA)
Provide the second of CarGPT after regulates inductions (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	Incorporate AI use as part of security awareness and education.	Cyber Talent Transformation (CSS J7 OCC: 30J8Y)
sensitive information to the chatbot.	the model unless users opt out."	Implement role-based access at the user	Identity & Access Management
Public	Additional Challenges	and system level (API).	Identity & Access Security (CSS J7 OCC: 30JA7, 30BUO)
User App	No control over the security protocols followed in the AI system.	Employ data loss prevention techniques to detect and prevent SPI,PII and regulated data leakage through prompts and API's.	 Data Security Data Security for AI/ML (CSS J7 OCC: 30J20, 30B2I, 30JY0, 30JNT)
	The wrong user having access to the		Threat Management
°¦° ▼ ↓ @} Model	system.	Monitor for exploits, anomalous activities and insider threats.	 Threat Management (CSS J7 OCC: 30B33, 30B3A, 30JCC, 30BS2, 30BS0)
	Improper use of the AI system. (jailbreaking, prompt injection)	Partnering with new technology vendors	to provide unique protection for AI.
		🚸 HIDDENLAYER 🚿 baff	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
A REAL PROVIDED AND A REAL	"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 y control. Establish effective governance model and	 AI Strategy & Governance Security Assessment w/ Threat Modeling (CSS J7 OCC: 30BAA) AI Governance Programs (CSS J7 OCC: 30BAA)
A fir out \$ deep comp video			Implement role-based access at the user system level (API).	and Identity & Access Management Identity & Access Security (CSS J7 OCC: 30JA7, 30BUO)
Hybrid User App	Additional Challenges Accidental data leakage of proprietary or SPI information.		Institute model scanning to protect from s chain attacks. Apply MLSecOps to secure	upply Vulnerability Management your · Vulnerability Management
			model lifecycle. Perform pen-testing on test ML model resiliency.	 (CSS J7 OCC: 30J0K) Sec DevOps Services (CSS J7 OCC: 30BVZ, 30J2A) Adversarial MI
W Model Attackers using model inversion to retrieve sensitive information		nodel inversion to information	Employ data loss prevention techniques to detect and prevent SPI,PII and regulated leakage through prompts and API's.	 XFR Testing (CSS J7 OCC: 30JT6, 30B37) Data Security Data Security for AL/MI
	Model poisoning compromising the model's effectiveness, security or integrity		Monitor for exploits, anomalous activities a insider threats.	and • Threat Management • Threat Management (<i>css j7 occ: 30B33</i> ,
		Partnering with new technol	ogy vendors to provide unique protection for AI.	30B3A, 30JCC, 30BS2, 30BS0)

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

38TB of data accidentally exposed by Microsoft Al	What Happened	How Security for AI should be applied	How IBM Cybersecurity Services can help
<section-header></section-header>	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	Assess security gaps in posture and architecture Ensure cloud resources are configured to comply with security and compliance best practices.	 AI Strategy & Governance Security Assessment w/ Threat Modeling (CSS J7 OCC: 30BAA) Security Posture Management (CSS J7 OCC: 30JA5)
Contractions Starting Contractions Starting Contractions	IOF 24 HOURS DECAUSE OF a DDOS ATTACK.	Implement role-based access at the user and system level (API).	Identity & Access Management Identity & Access Security (CSS J7 OCC: 30JA7, 30BUO)
	Additional Challenges	Secure cloud, infrastructure (HPCs), storage, and cloud AI services and posture	Infrastructure Security Infrastructure Security Services (CSS J7 OCC: 30JAM, 30JLX)
Private User App Custom Model Open-Source Model External Data	Accidental data leakage of proprietary or SPI information.	Institute model scanning to protect from supply chain attacks. Apply MLSecOps to secure your model lifecycle.	 Vulnerability Management Vulnerability Management (CSS 37 OCC: 3030K)
	Model evasion attacks that cause the model to misclassify or misinterpret	Perform pen-testing on test ML model resiliency.	Sec DevOps Services (CSS J7 OCC: 308VZ, 3032A) Adversarial ML
	inputs Attackers attempting to steal models with model extraction techniques	Employ data loss prevention techniques to detect and prevent SPI,PII and regulated data leakage through prompts and API's.	AFR Testing Services (css 37 occ: 30376, 30837) Data Security Data Security for AI/ML (css 77 occ: 30170, 3087, 30190, 30190)
Partnering with new technology vendors to provide unique protection for AI.		Monitor for exploits, anomalous activities and insider threats.	 Threat Management Threat Management (css J7 OCC: 30B33, 30B3A, 30JCC, 30B52, 30B50)
🚸 HIDDENLAYER 🔹 baffle	Reality Defender		

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



Partnering with new technology vendors to provide unique protection for AI

w baffle		Model Scanning	Threat Monitoring with MLDR technologies	
HIDDEN LAYER	Reality Defender	AI FirewallData Protection for AI	Deep Fake recognition	
BM Cybersecurity Services / © 2024 IBM Corporation				13

Thank you

© 2024 International Business Machines Corporation

IBM and the IBM logo are trademarks of IBM Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/trademark.

Internal only. Do not distribute outside of IBM.

This document is current as of the initial date of publication and may be changed by IBM at any time



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



IBM Cybersecurity's POV | AI for Security + Security for AI

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



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





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

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

IBM Consulting Cybersecurity Services / © 2023 IBM Corporation