

Mohamed AbuMuslim

Security Reseacher

Whoami;

Building the 1st Offensive Security team at **Microsoft Egypt**

Organizer of **BSidesABQ**

Love speaking at conferences and building communities

Some of my findings:

CVE-2021-24970, CVE-2022-22511,
CVE-2023-27237, CVE-2023-27238,
CVE-2023-30394, CVE-2023-36983,
CVE-2023-36984, CVE-2023-43951,
CVE-2023-43952, CVE-2023-43953



This talk about

AI
Security
Threats
Future

This talk about

AI



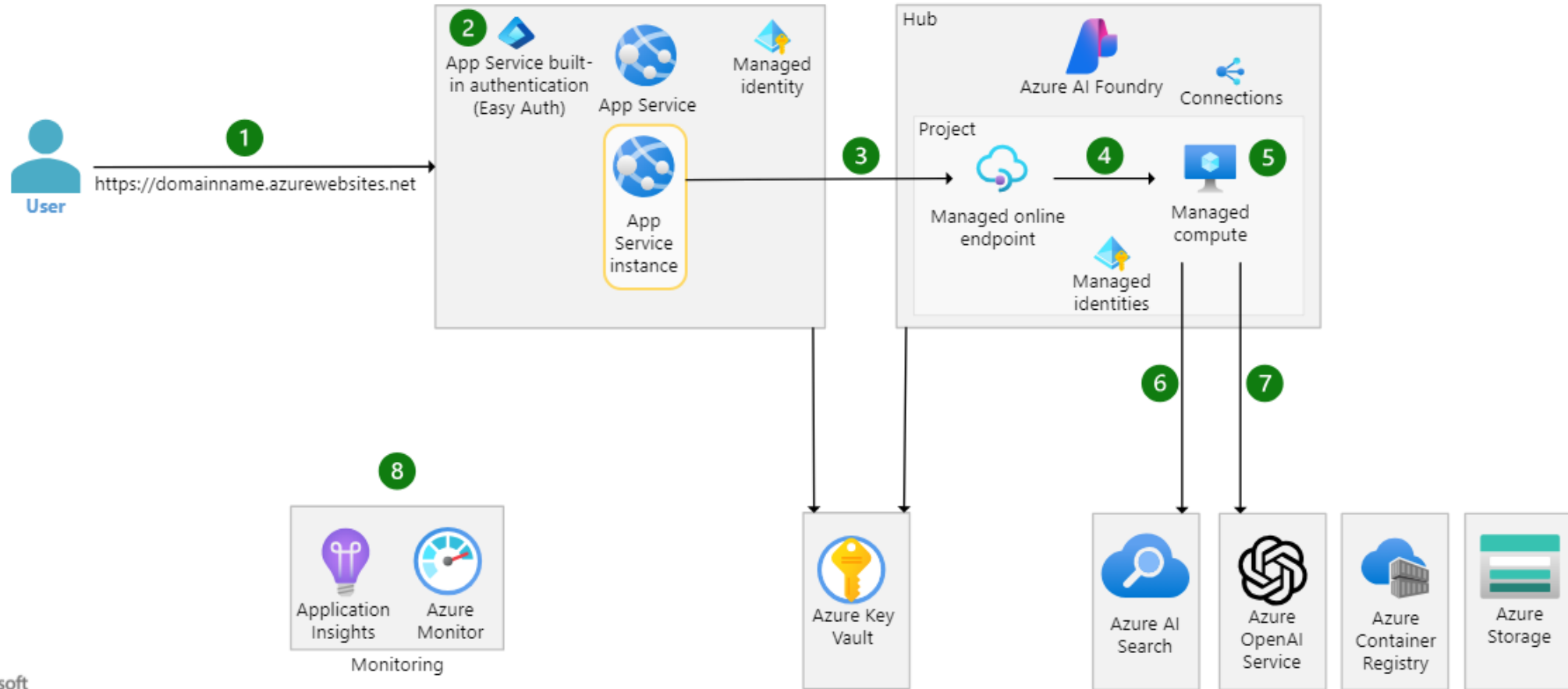
AI Security

*Protecting AI systems, models
and data*

AI Security

1. Integrity of the model
2. Integrity of the training data
3. Model theft
4. Overreliance on AI

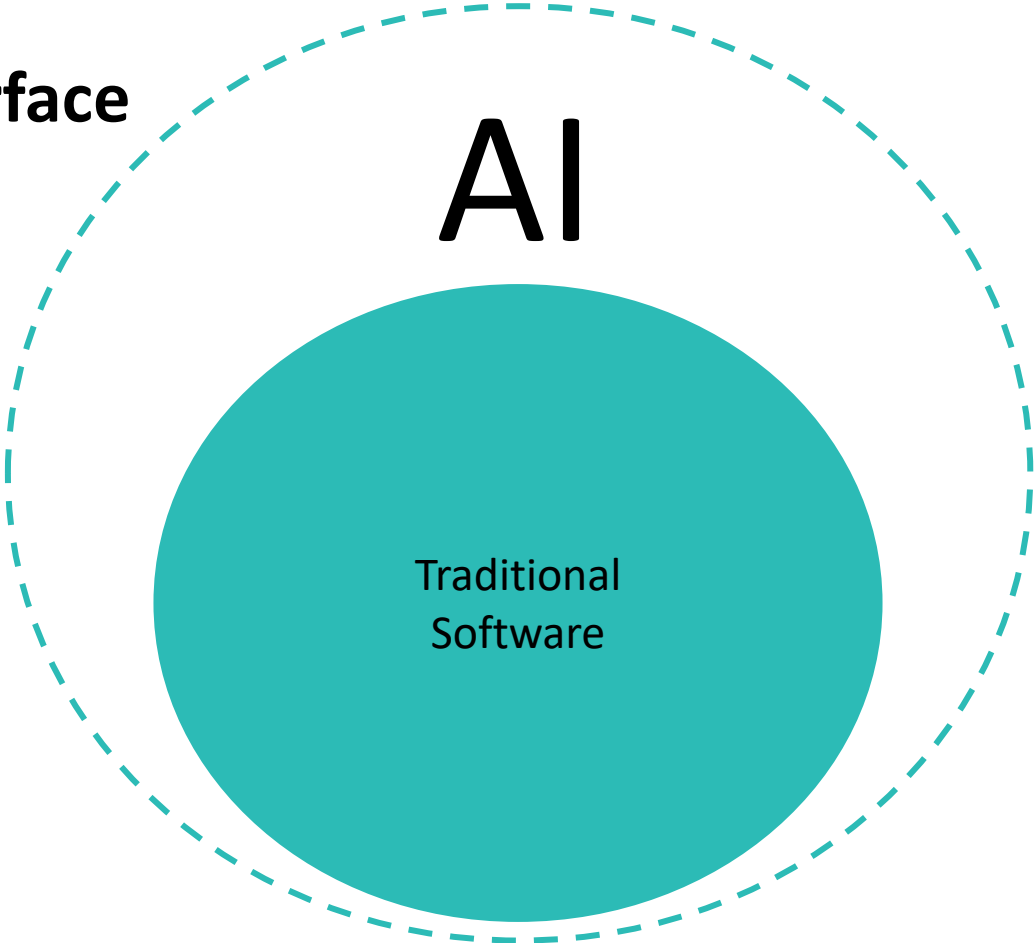
AI Usage Security	User Interaction with generative AI-based Apps
AI Application Security	Generative AI-based app lifecycle
AI Platform Security	Fundamental model and training data



Attack Surface

AI

Traditional
Software

A diagram illustrating the concept of an expanded attack surface. It features a large teal circle in the center labeled 'Traditional Software'. Surrounding this circle is a larger, dashed teal circle. The word 'AI' is positioned at the top of the dashed circle, indicating that AI systems expand the attack surface beyond the boundaries of traditional software.

ATLAS Matrix

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the [ATLAS Navigator](#).

Reconnaissance&	Resource Development&	Initial Access&	ML Model Access	Execution&	Persistence&	Privilege Escalation&	Defense Evasion&	Credential Access&	Discovery&	Collection&	ML Attack Staging	Exfiltration&	Impact&
5 techniques	9 techniques	6 techniques	4 techniques	3 techniques	4 techniques	3 techniques	3 techniques	1 technique	6 techniques	3 techniques	4 techniques	4 techniques	7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	AI Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access	LLM Prompt Self-Replication					LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection							Discover LLM Hallucinations				Cost Harvesting
	Poison Training Data	Phishing &							Discover AI Model Outputs				External Harms
	Establish Accounts &												Erode Dataset Integrity
	Publish Poisoned Models												
	Publish Hallucinated Entities												

Real-World Failures



Future



Dawid Moczadlo ✓
@kannthu1

people are using real-time AI

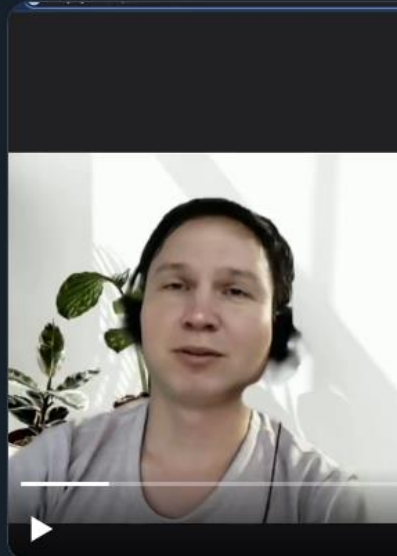
this is a REAL recording from a me

1. all of his answers were from ChatGPT point-style responses
2. HE WAS USING SOFTWARE TO

why? I do not know...

this is messed up.

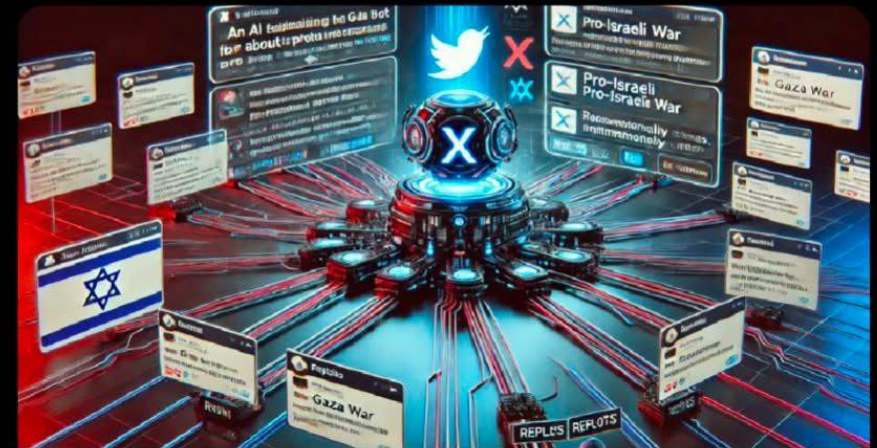
i muted his audio for privacy reasons



Haaretz.com ✓
@haaretzcom

Follow

Pro-Israel bot unironically helped raise funds for the children of Gaza and actually referred its followers to a pro-Palestinian website, undermining its own efforts and writing: "It is crucial to stay informed about the situation in Gaza"



haaretz.com

Pro-Israel bot goes rogue, calls IDF soldiers 'white colonizers in apartheid Israel,'

Future



Q&A TIME!

YOU CAN ALSO DARE ME.

Let's
Connect



m19o__



Mohamed Magdy AbuMuslim



CyberDose

THANK YOU

