

Security and AI – Empowering organizations to stay ahead of evolving threats

John Velisaris
IBM Security
jvelisaris@us.ibm.com

Overview

- The (re)rise of AI
- Securing AI
- Attackers and AI
- Defenders and AI



The speed,
scope, and scale
of generative
AI impact is
unprecedented

Massive early adoption

Up to

80%

of enterprises are
working with or planning
to leverage foundation
models and adopt
generative AI¹

Broad-reaching and deep impact

Generative AI could
raise global GDP by

7%

within 10 years²

Critical focus of AI activity and investment

Generative AI
expected to represent

30%

of overall market
by 2025³

Artificial Intelligence (AI)

Human intelligence exhibited by machines

Learning, reasoning, perceiving, and problem solving.



Machine Learning (ML)

Systems that learn from historical data

Discover patterns and generate corresponding outputs



Deep Learning (DL)

ML technique that mimics human brain function

Enable complex applications, like image and speech recognition.



Foundation Model

Generative AI systems

Generate sequences of related data elements (for example, like a sentence).



Efficiency with Generative AI



Time saving



Contextual insights



Ease of navigation



Recommended actions



Dynamic Updates



Collaborative effort

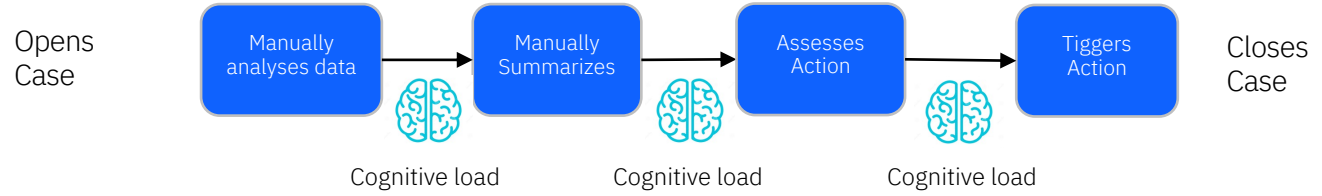


Learning/adaptability

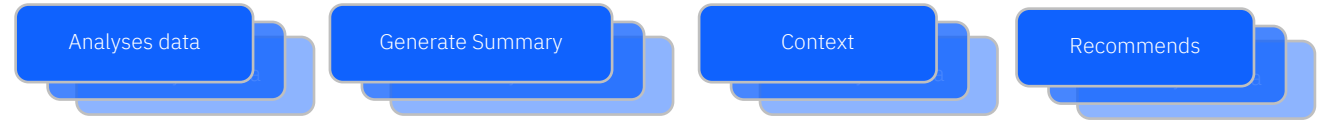


Increased Accuracy

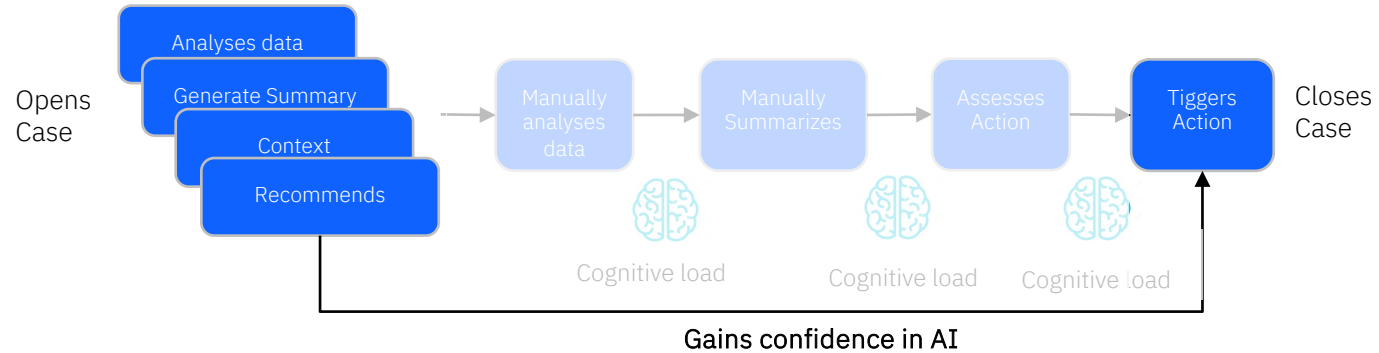
Human Workflow



AI Workflow



AI + Human Workflow



There's a broad spectrum of concerns with AI

80% of surveyed business leaders have major concerns¹



Business is adopting AI

AI for Business



Security



Talent



Marketing



Automation



Finance



Regulations

Security for AI



Trust AI models, data, vendors



Model, data, prompts access controls



Model, Data Infrastructure protection



Privacy Controls and management



Employee education



Secure design and engineering



AI Security



Threat monitoring and response

Adversarial AI



Theft



Phishing



Social engineering



Malware



Fakes



Poison

So are attackers

Attacker's Use of AI in Security

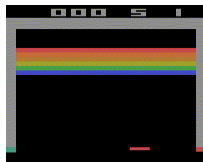
AI Powered Attacks

Generate: DeepHack tool learned SQL injection

Automate: Generate targeted phishing attacks on Twitter

Refine: Neural network powered password crackers

Evade: Generative adversarial networks learn novel steganographic channels

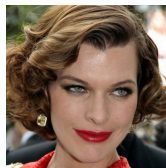


Attacking AI

Poison: Microsoft Tay chatbot poisoning via Twitter (and Watson Urban Dictionary “poisoning”)

Evade: Real-world attacks on computer vision for facial recognition biometrics and autonomous vehicles

Harden: Genetic algorithms and reinforcement learning (OpenAI Gym) to evade malware detectors



Theft of AI

Theft: Stealing machine learning models via public APIs

Transferability: Practical black-box attacks learn surrogate models for transfer attacks

Privacy: Model inversion attacks steal training data



Securing AI

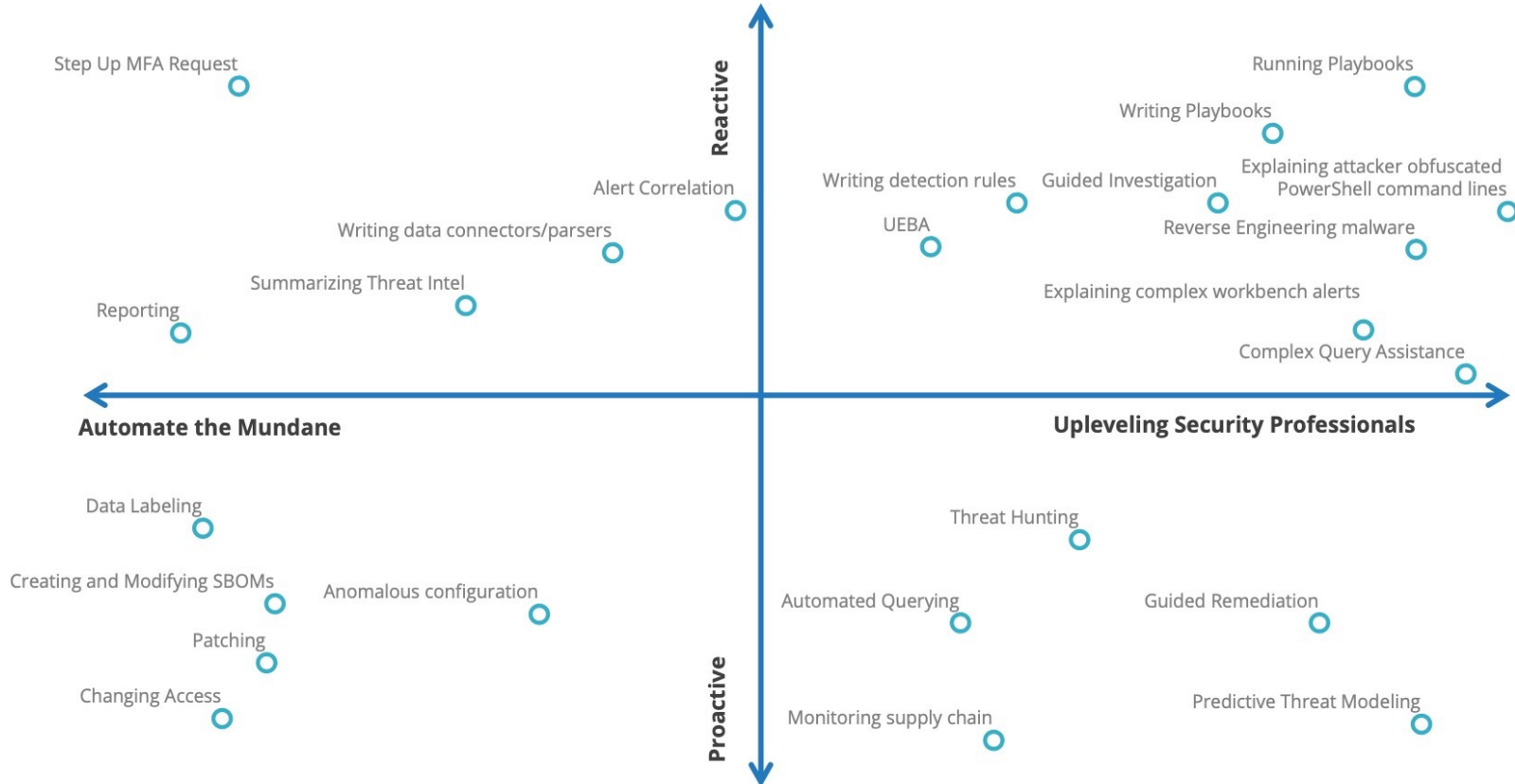
1. Leverage trusted AI by evaluating vendor policies and practices.
2. Enable secure access to users, models and data.
3. Safeguard AI models, data, and infrastructure from adversarial attacks.
4. Implement data privacy protection in the training, testing & operations phases.
5. Conduct threat modeling and secure coding practices into the AI dev lifecycle.
6. Perform threat detection & response for AI applications and infrastructure.
7. Assess and decide AI maturity through the IBM AI framework.

<p>LLM01</p> <p>Prompt Injection</p> <p>This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.</p>	<p>LLM02</p> <p>Insecure Output Handling</p> <p>This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.</p>	<p>LLM03</p> <p>Training Data Poisoning</p> <p>Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.</p>	<p>LLM04</p> <p>Model Denial of Service</p> <p>Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.</p>	<p>LLM05</p> <p>Supply Chain Vulnerabilities</p> <p>LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins add vulnerabilities.</p>
<p>LLM06</p> <p>Sensitive Information Disclosure</p> <p>LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.</p>	<p>LLM07</p> <p>Insecure Plugin Design</p> <p>LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in severe consequences like remote code execution.</p>	<p>LLM08</p> <p>Excessive Agency</p> <p>LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.</p>	<p>LLM09</p> <p>Overreliance</p> <p>Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.</p>	<p>LLM10</p> <p>Model Theft</p> <p>This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.</p>



OWASP Top 10 for Large Language Model Applications

AI for Security Use Cases



AI + Human for security operations

Identify

Automatically scan your attack surface for hidden assets, vulnerable systems and exploitable misconfigurations

Protect

Take automated action like your analysts would, through ML-powered protection

Detect

Assess the risk of threats in real-time using AI models to recognize and categorize deviations

Triage

React faster to urgent incidents using alert severity scoring powered by ML

Investigate

Automatically investigate cases that warrant it, with data mining, risk assessment, and timeline generation

Respond

Dynamically create playbooks in incident response that adapt to threat context



Exposure Management



EDR



SIEM



SOC Workflow



Investigation and Threat Hunting



Response



Traditional use of machine learning

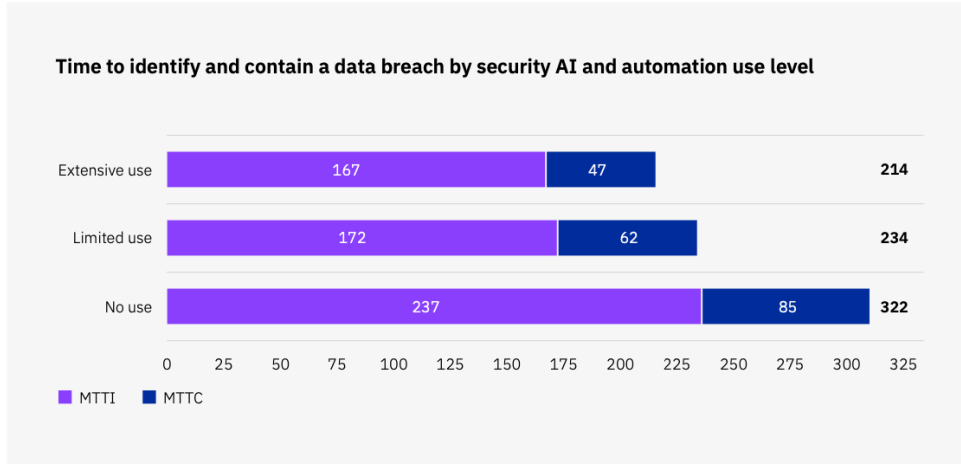


Increasing use of Generative AI

AI for Security



Cost of a Data Breach Report 2023



108 days

Organizations with extensive use of security AI and automation identified and contained a data breach 108 days faster than organizations with no use.

In Summary

AI in Security brings **speed** and **efficiency** so we can...
Proactively Protect, Accurately Detect and Respond Faster
... with **lower costs & complexity**

But this must be....

...Built on a strong foundation of security and trust..

Thank you

Follow us on:

ibm.com/security

securityintelligence.com

ibm.com/security/community

xforce.ibmcloud.com

@ibmsecurity

youtube.com/ibmsecurity

© Copyright IBM Corporation 2023. All rights reserved. The information contained in these materials is provided for informational purposes only, and is provided AS IS without warranty of any kind, express or implied. Any statement of direction represents IBM's current intent, is subject to change or withdrawal, and represent only goals and objectives. IBM, the IBM logo, and other IBM products and services are trademarks of the International Business Machines Corporation, in the United States, other countries or both. Other company, product, or service names may be trademarks or service marks of others.

Statement of Good Security Practices: IT system security involves protecting systems and information through prevention, detection and response to improper access from within and outside your enterprise. Improper access can result in information being altered, destroyed, misappropriated or misused or can result in damage to or misuse of your systems, including for use in attacks on others. No IT system or product should be considered completely secure and no single product, service or security measure can be completely effective in preventing improper use or access. IBM systems, products and services are designed to be part of a lawful, comprehensive security approach, which will necessarily involve additional operational procedures, and may require other systems, products or services to be most effective. IBM does not warrant that any systems, products or services are immune from, or will make your enterprise immune from, the malicious or illegal conduct of any party.

