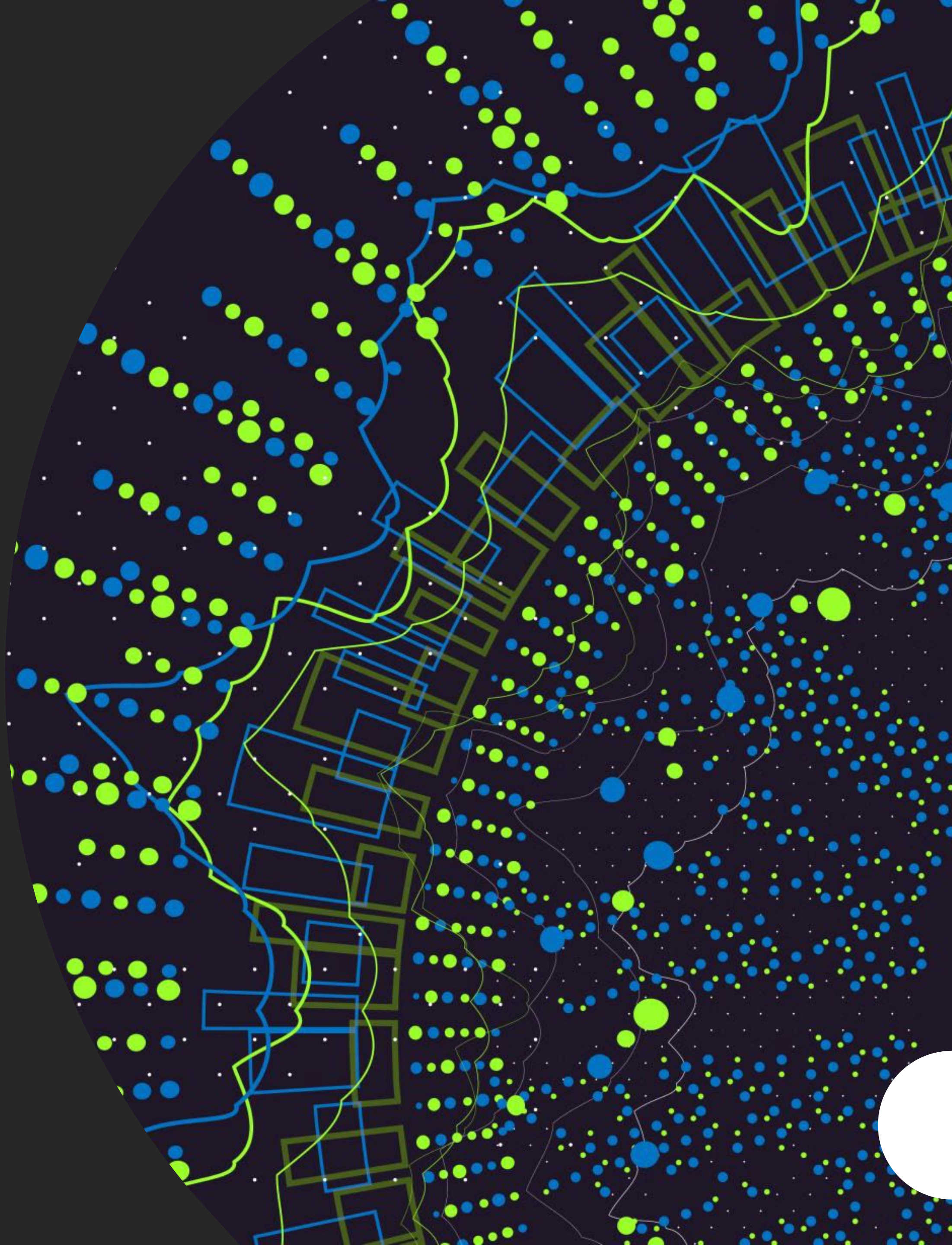
AI Safety Strategies: Mitigating Risks and Addressing Challenges

Professor/Chair Kyle Jones



Introduction

Kyle Jones - MS, A+, Net+, Security+, CYSA+, ITIL, Strata

Chair/ Professor

Sinclair Community College

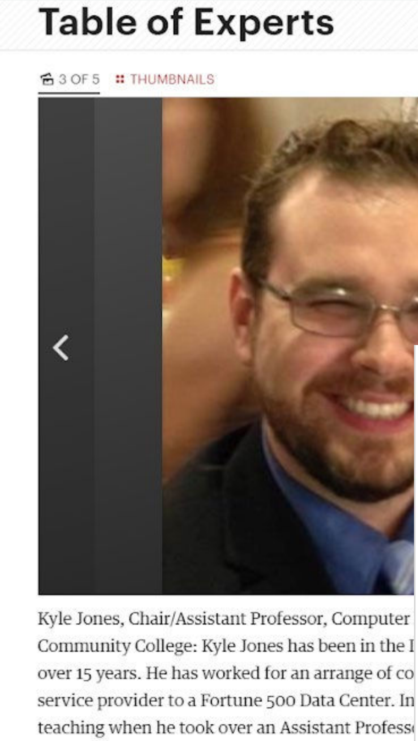


Table of Experts

Kyle Jones, Chair/Assistant Professor, Computer
Community College: Kyle Jones has been in the I
over 15 years. He has worked for an arrange of co
service provider to a Fortune 500 Data Center. In
teaching when he took over an Assistant Profess

Dayton Daily
Complete. In-Depth. Dı

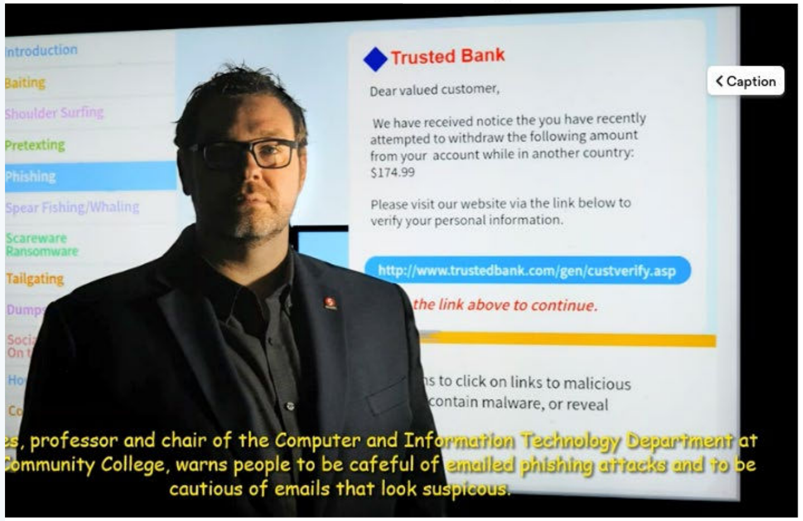
Limited time offer! Get unlimited digital access for

Data breaches see slight d
year in 20

Sinclair Community College Selected for Innovative Training Program to Strengthen Cybersecurity Education in High Schools

Sinclair Community College is one of three institutions selected by the U.S. Department of Education to offer a new training program designed to strengthen cybersecurity education in high schools. The CTE CyberNet Program is designed to give teachers knowledge and resources they can use to effectively prepare students for cybersecurity courses and careers.

The U.S. Department of Education identified three National Centers of Academic Excellence (NCAE), including Sinclair Community College, which are qualified to "design, host, and lead the inaugural cohort of CTE CyberNet academies." Sinclair offered the week-long CyberNet Program for local K-12 and community college instructors at its Centerville campus July 18-22, 2022. Sinclair was the first to host the regional training event.



Trusted Bank

Dear valued customer,

We have received notice the you have recently attempted to withdraw the following amount from your account while in another country:
$174.99

Please visit our website via the link below to verify your personal information.

http://www.trustedbank.com/gen/custverify.asp

the link above to continue.

ns to click on links to malicious
contain malware, or reveal

es, professor and chair of the Computer and Information Technology Department at
Community College, warns people to be cafeful of emailed phishing attacks and to be
cautious of emails that look suspicous.

Practicing good cyber hygiene

By Katie Ussin
Published: January 18, 2017, 6:47 pm

DAYTON, Ohio (WDTN) – Kyle Jones is the chair of the computer information systems department at Sinclair Community College in Dayton. He stopped by Five on 2 to talk about cyber security — specifically about social engineering and cyber hygiene.

Top News

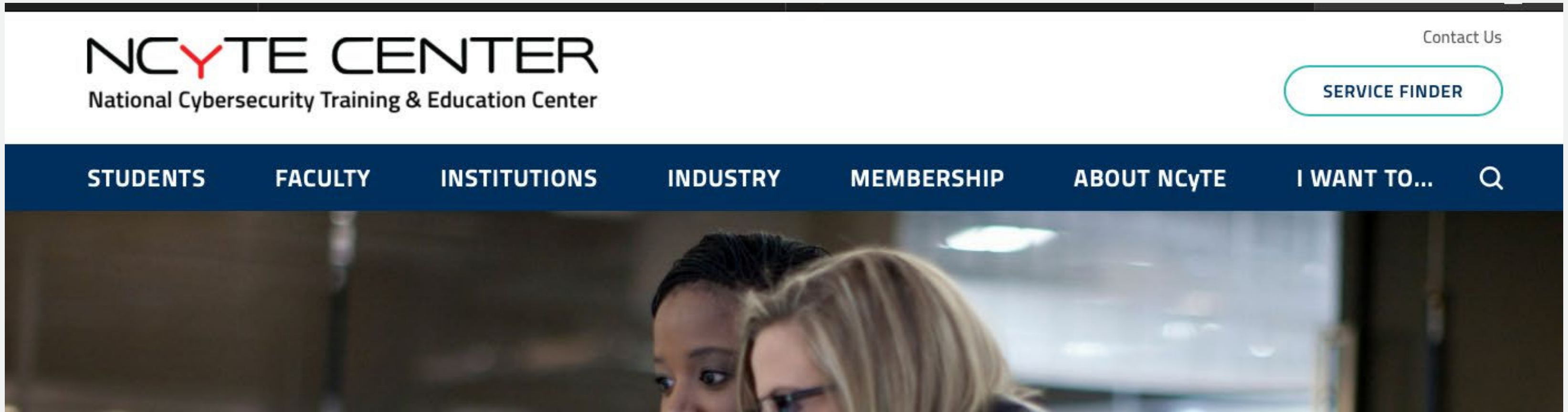
Local students host walkout in support of Florida students
Local students held a walkout event to support the...

US Embassy trip to Israel

BACK TO SCHOOL SAFELY
CYBERSECURITY FOR
K-12 SCHOOLS

SEAL OF THE PRESIDENT OF THE UNITED STATES

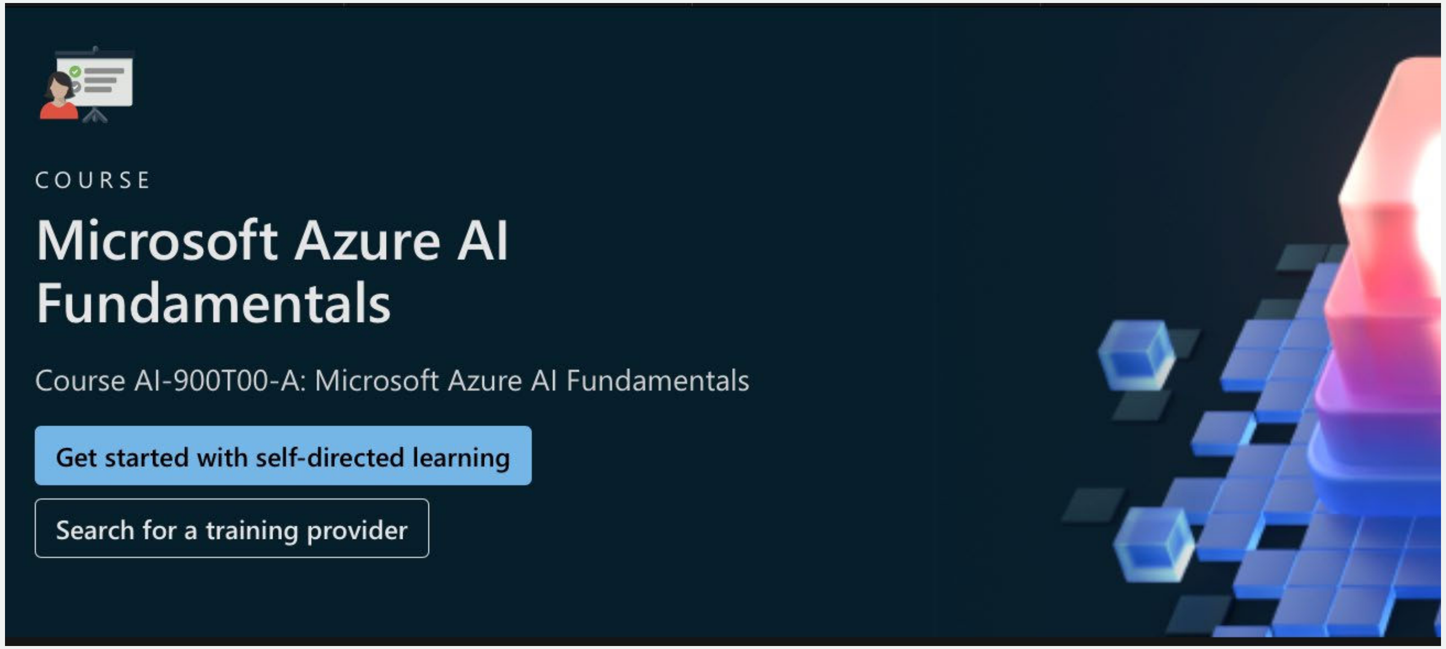
National Grants



NITIC
National Information Technology Innovation Center
Working together
to prepare the
IT professionals of tomorrow

Microsoft Azure AI

AI-900T00-A

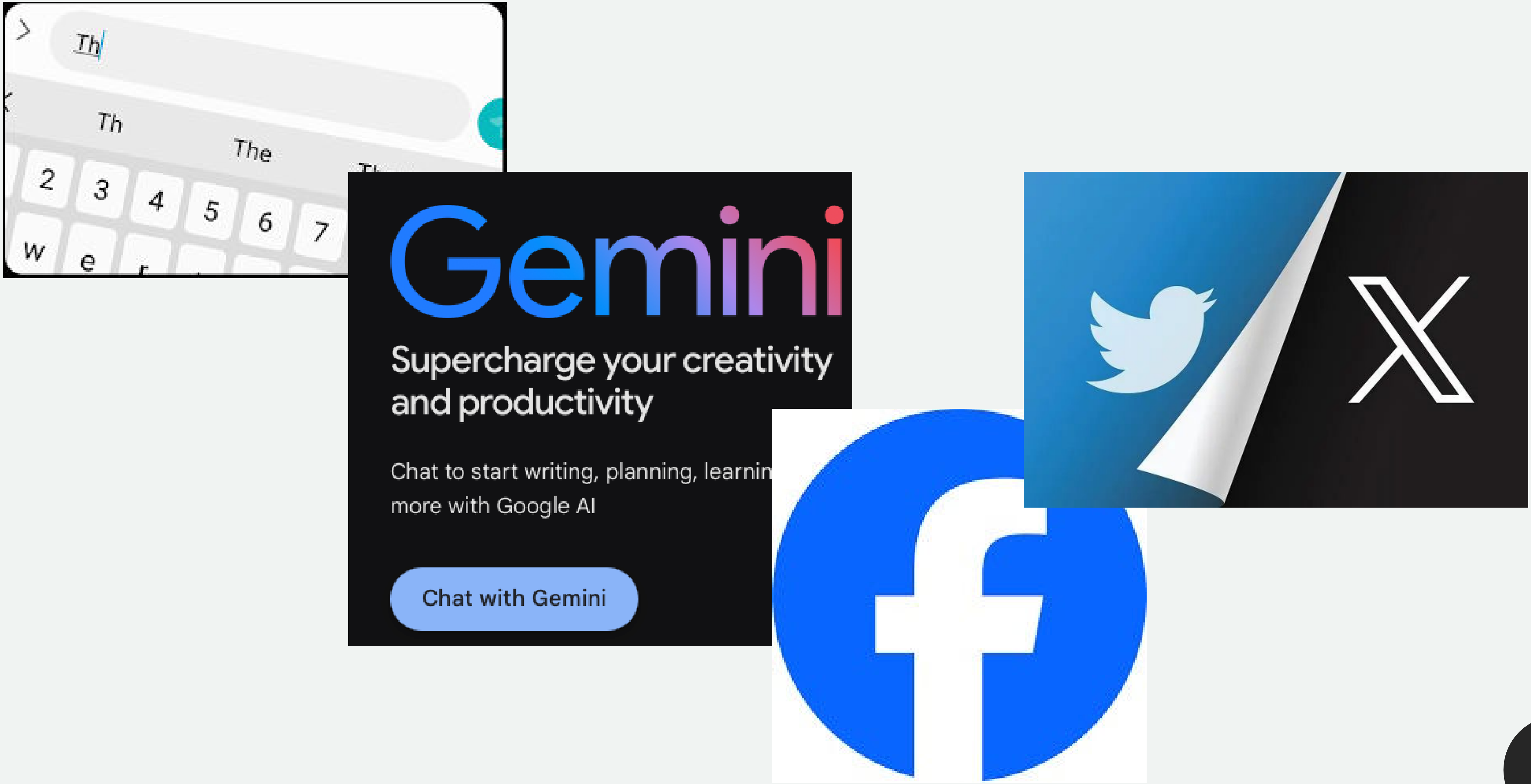


COURSE

Microsoft Azure AI Fundamentals

Course AI-900T00-A: Microsoft Azure AI Fundamentals

Get started with self-directed learning
Search for a training provider



Gemini

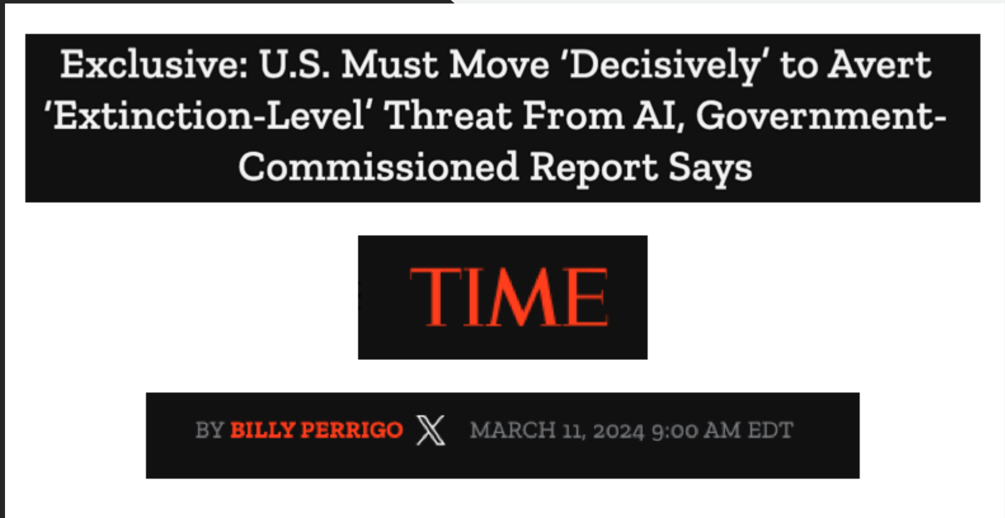
Supercharge your creativity and productivity

Chat to start writing, planning, learnin
more with Google AI

AI In use

- 1. **Customer Service Chatbots**:
- 2. **Personalized Marketing**:
- 3. **Inventory Management**:
- 4. **Automated Accounting and Bookkeeping**:
- 5. **Email Marketing Automation**:
- 6. **Social Media Management**:
- 7. **Sales Forecasting**:
- 8. **Recruitment and Hiring**:
- 9. **Fraud Detection and Cybersecurity**:
- 10. **Product Recommendations and Upselling**:

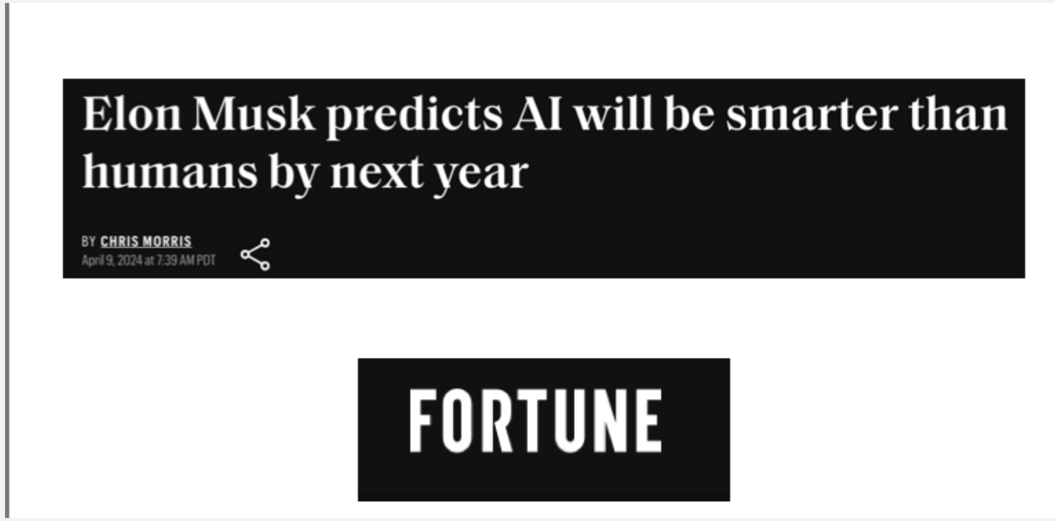
The Threat



Exclusive: U.S. Must Move 'Decisively' to Avert 'Extinction-Level' Threat From AI, Government-Commissioned Report Says

TIME

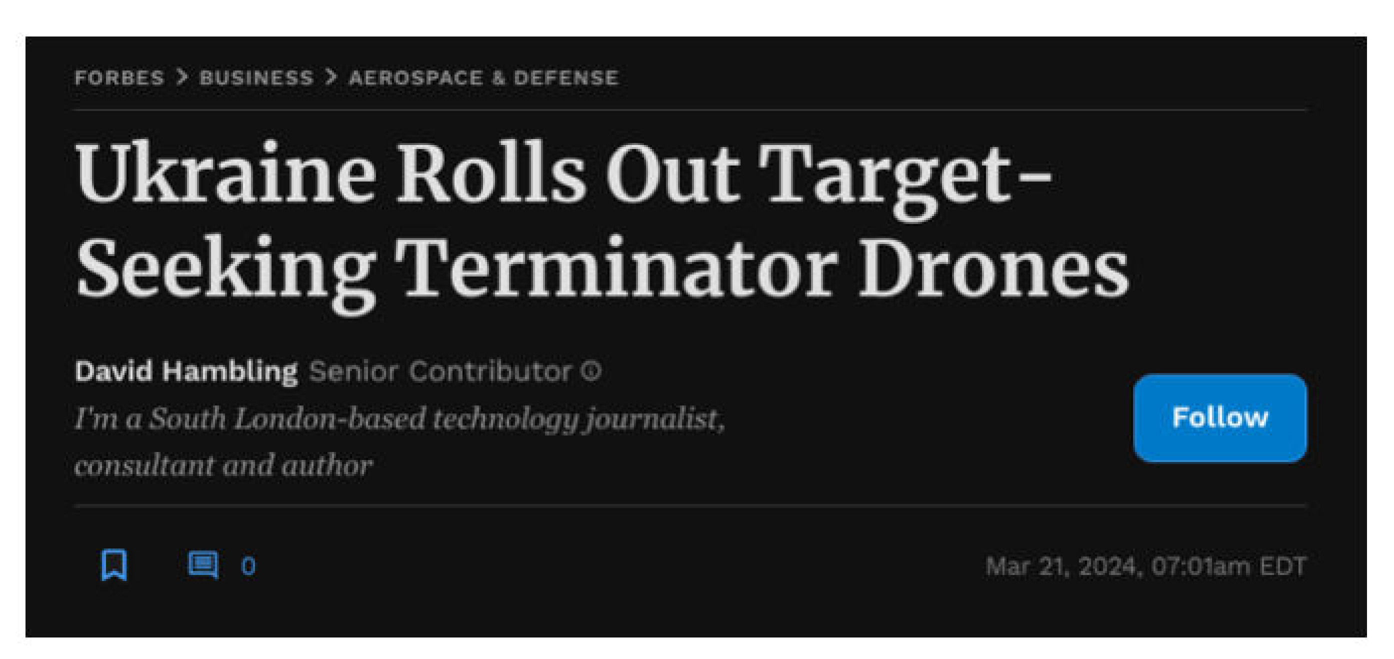
BY BILLY PERRIGO MARCH 11, 2024 9:00 AM EDT



Elon Musk predicts AI will be smarter than humans by next year

BY CHRIS MORRIS
April 9, 2024 at 7:39 AM PDT

FORTUNE



FORBES > BUSINESS > AEROSPACE & DEFENSE

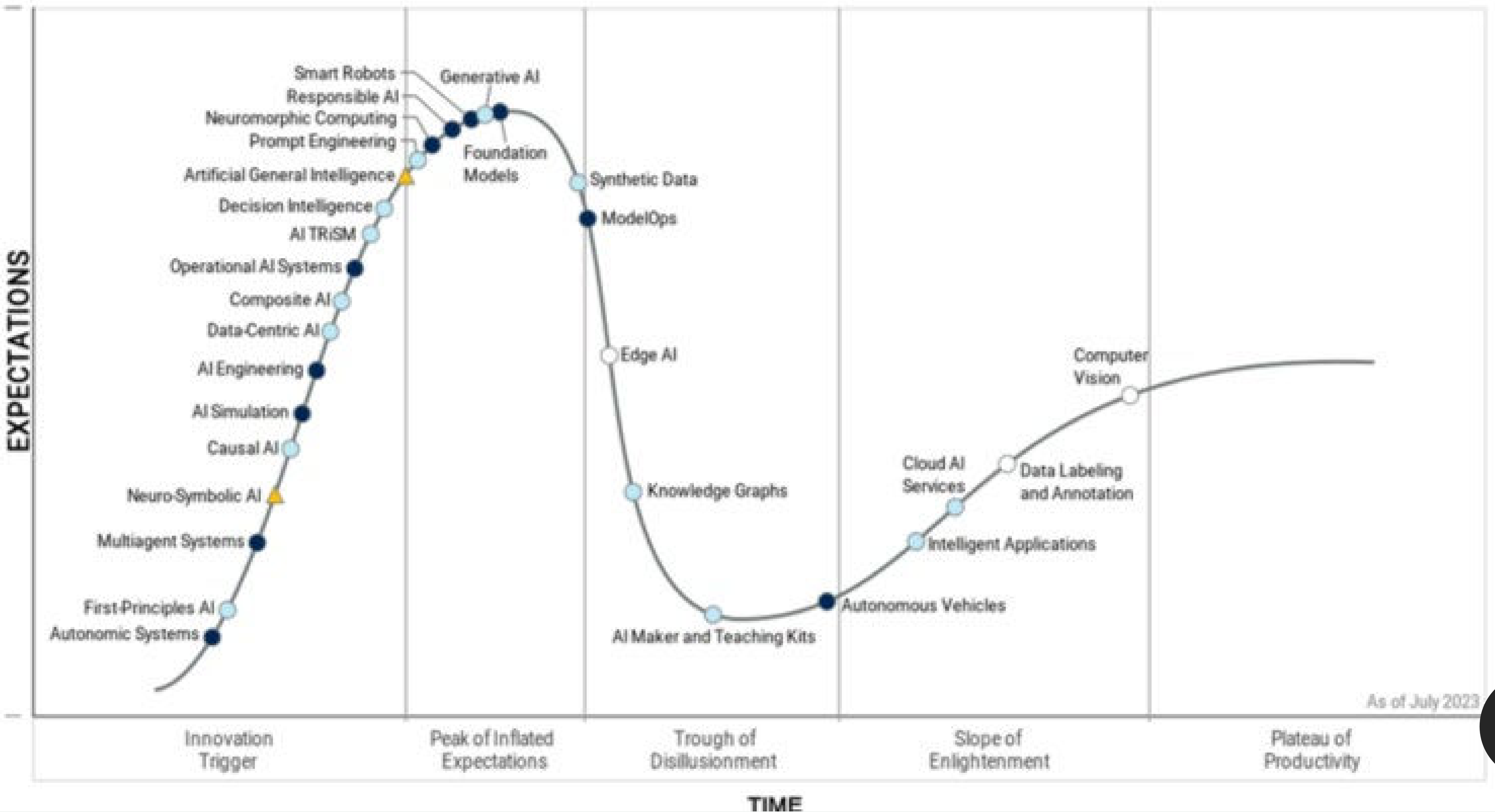
Ukraine Rolls Out Target-Seeking Terminator Drones

David Hambling Senior Contributor ⓘ

I'm a South London-based technology journalist, consultant and author

Mar 21, 2024, 07:01am EDT

Hype Cycle for Artificial Intelligence, 2023



Traditional Programming

- The program uses a long list of rules
- Difficult to maintain

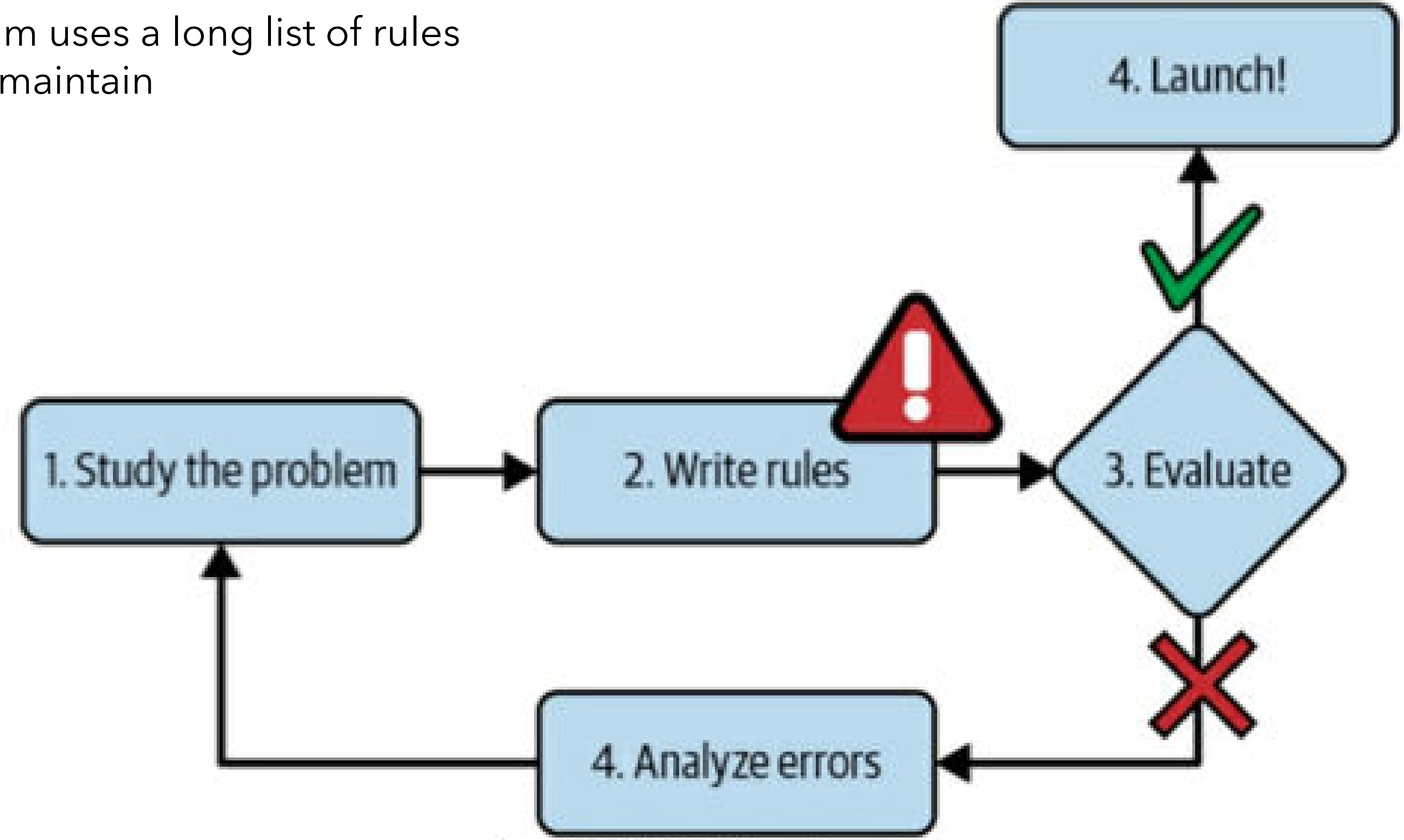


Figure 1-1. The traditional approach

Machine Learning

- Learns words and phrases that can predict spam
- The program is shorter, easier to maintain, and more accurate

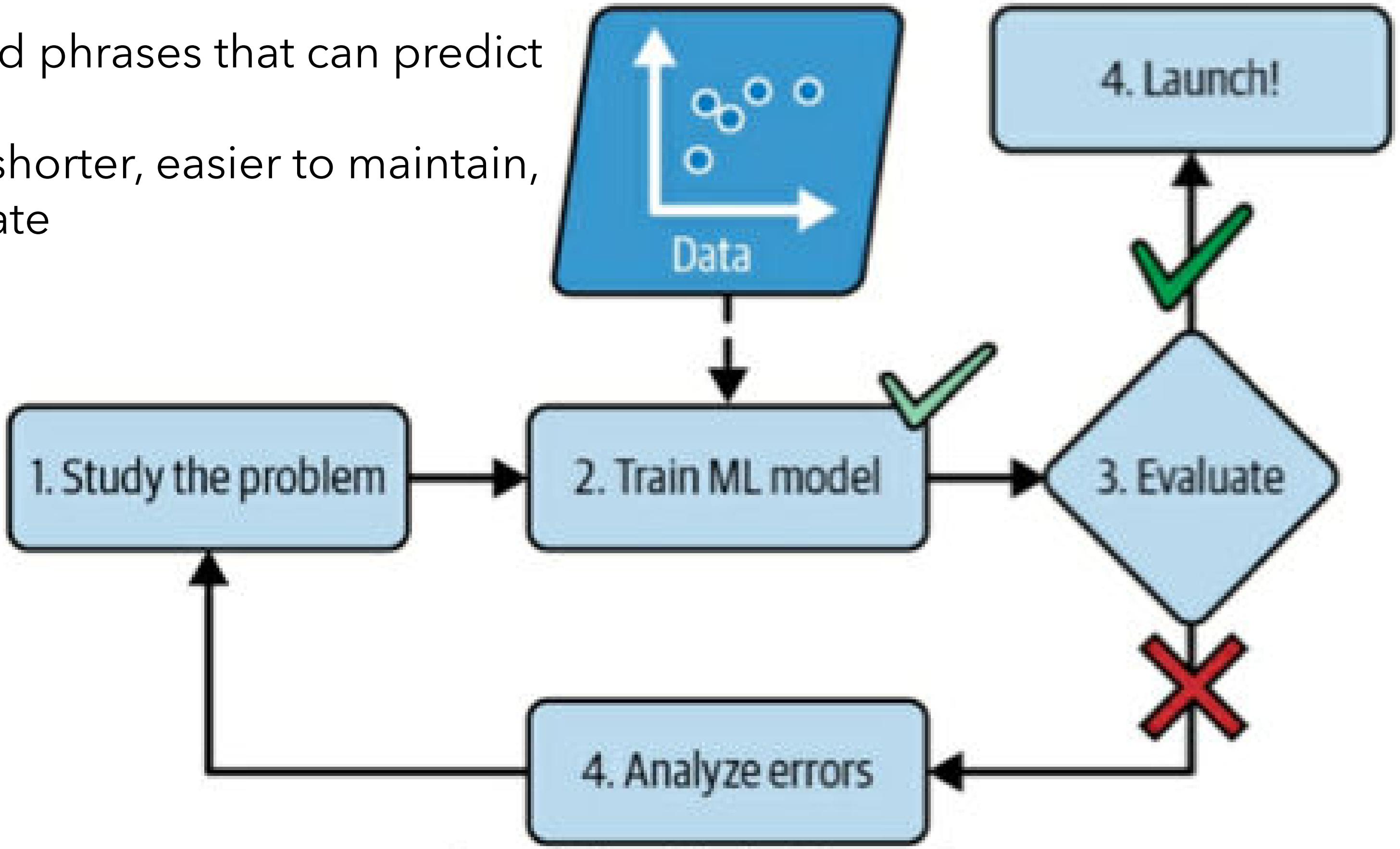
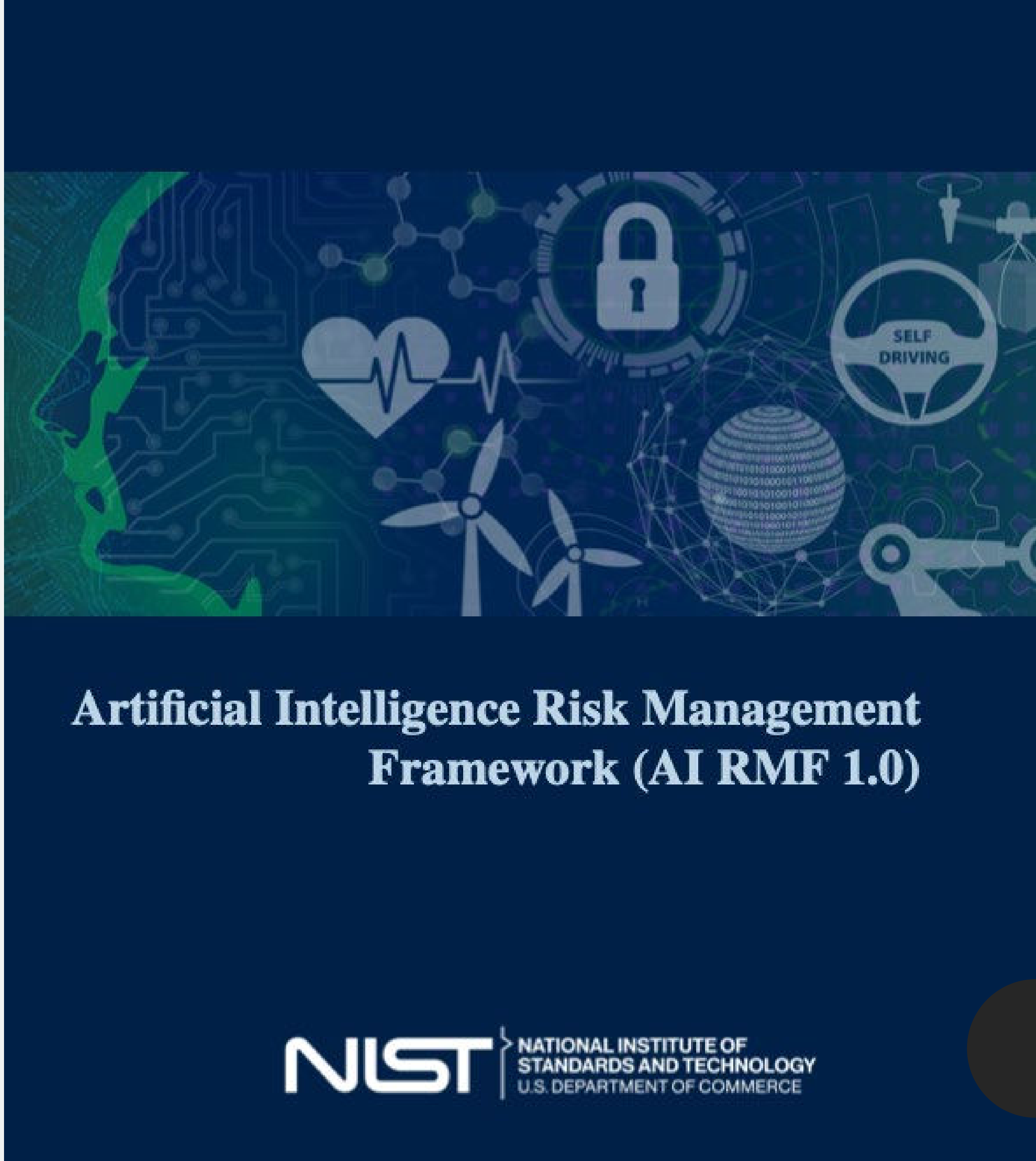


Figure 1-2. The machine learning approach

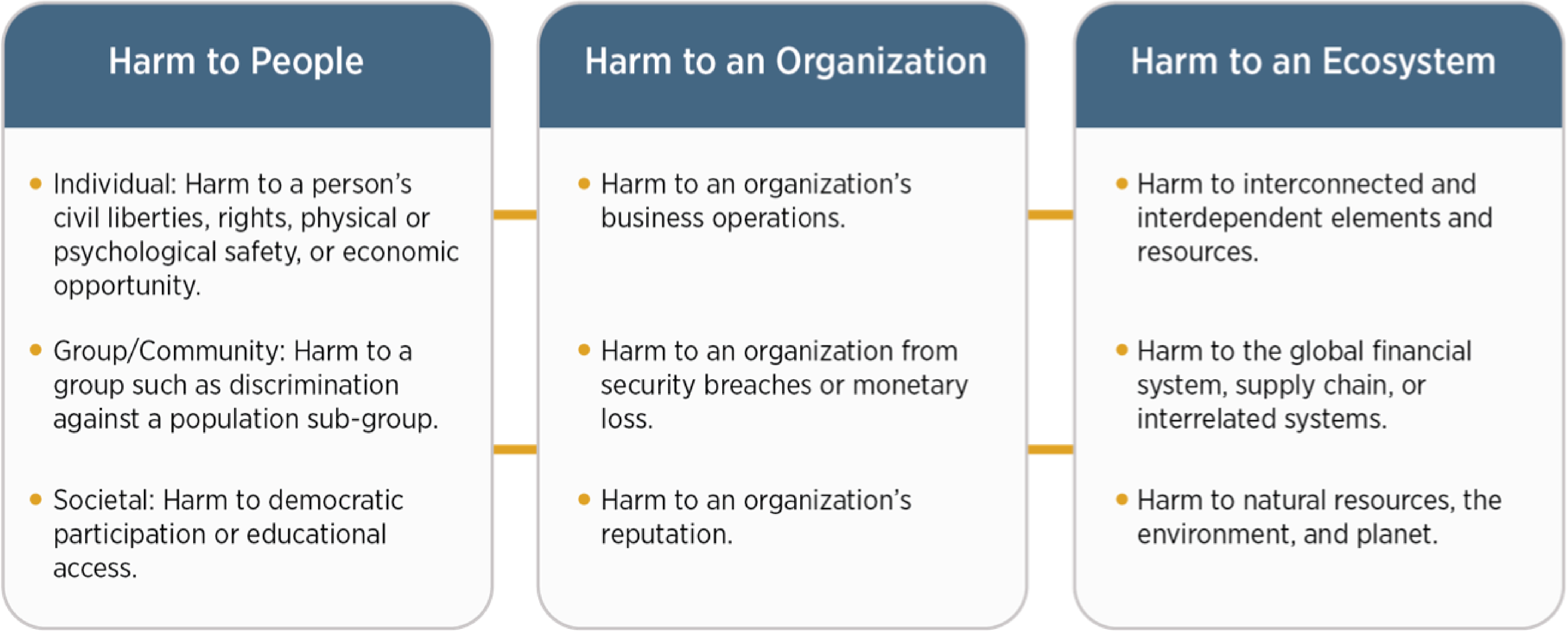
Artificial Intelligence Risk Management Framework (AI RMF 1.0)

- This publication is available free of charge from: https://doi.org/10.6028/NIST.AI.100-1



Artificial Intelligence Risk Management Framework (AI RMF 1.0)

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE



Harm to People

- Individual: Harm to a person's civil liberties, rights, physical or psychological safety, or economic opportunity.
- Group/Community: Harm to a group such as discrimination against a population sub-group.
- Societal: Harm to democratic participation or educational access.

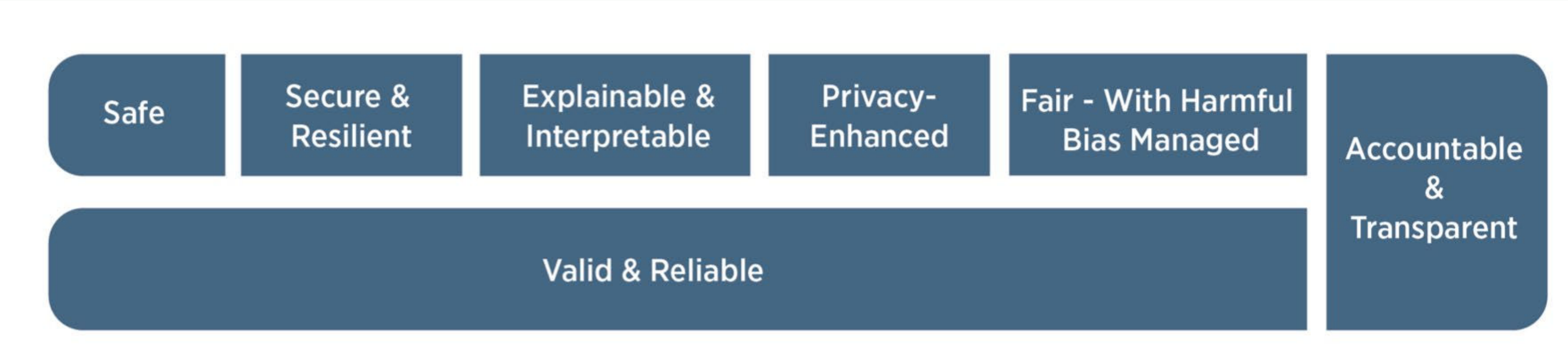
Harm to an Organization

- Harm to an organization's business operations.
- Harm to an organization from security breaches or monetary loss.
- Harm to an organization's reputation.

Harm to an Ecosystem

- Harm to interconnected and interdependent elements and resources.
- Harm to the global financial system, supply chain, or interrelated systems.
- Harm to natural resources, the environment, and planet.

Characteristics of Trustworthy AI



Safe

Secure & Resilient

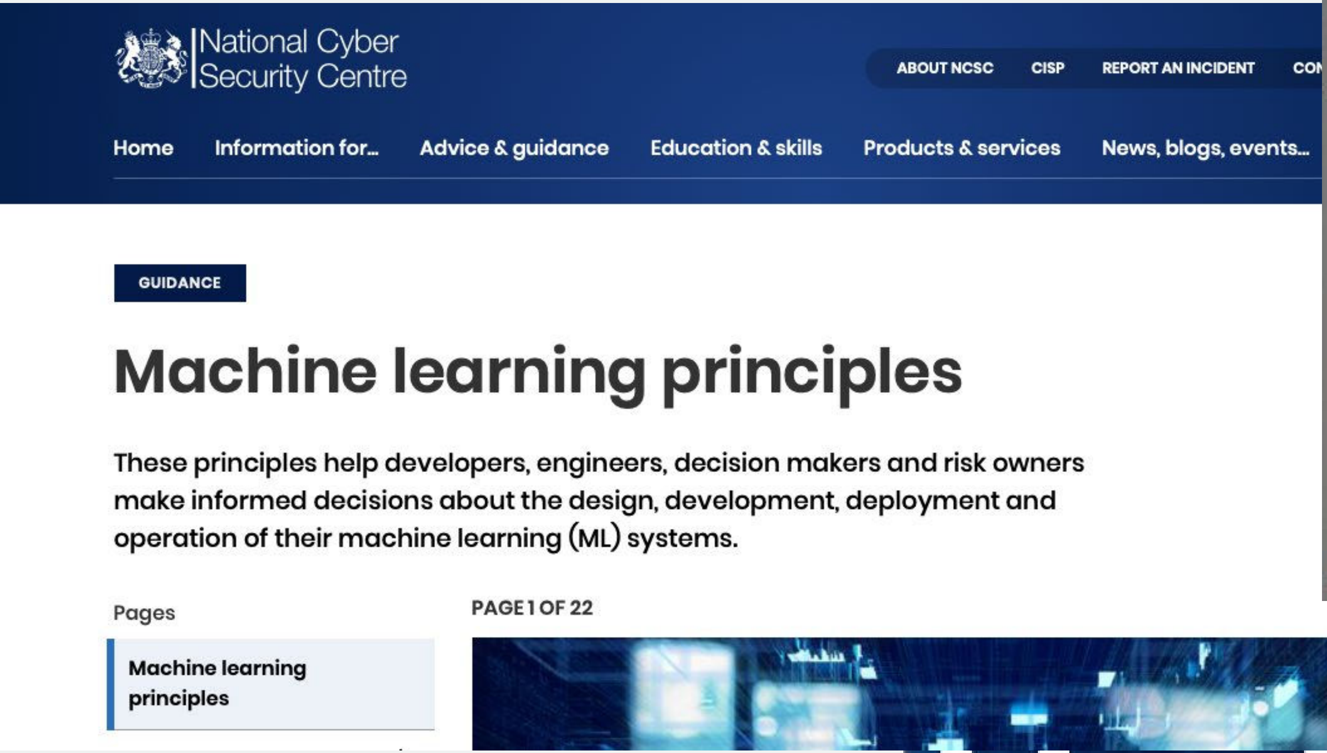
Explainable & Interpretable

Privacy-Enhanced

Fair - With Harmful Bias Managed

Accountable & Transparent

Valid & Reliable



GUIDANCE

Machine learning principles

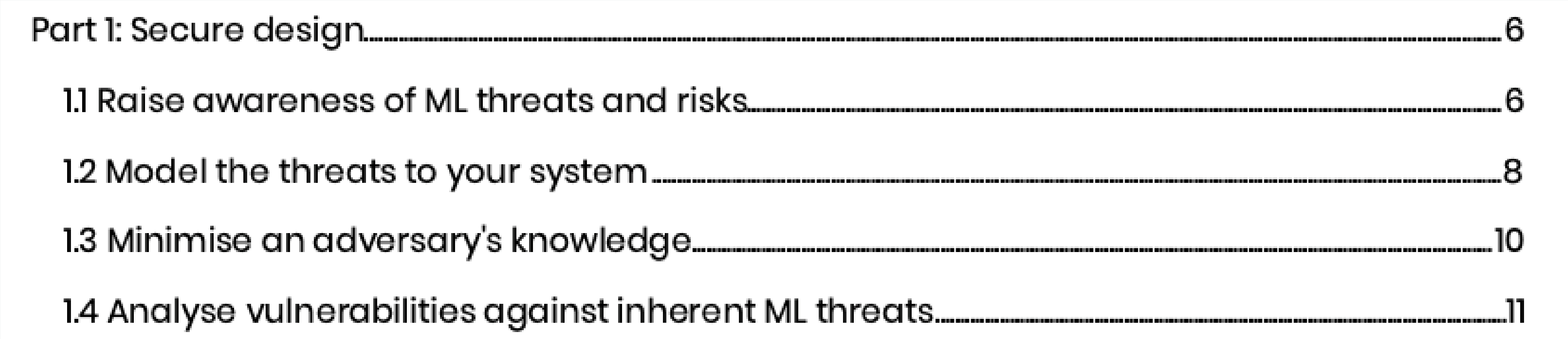
These principles help developers, engineers, decision makers and risk owners make informed decisions about the design, development, deployment and operation of their machine learning (ML) systems.

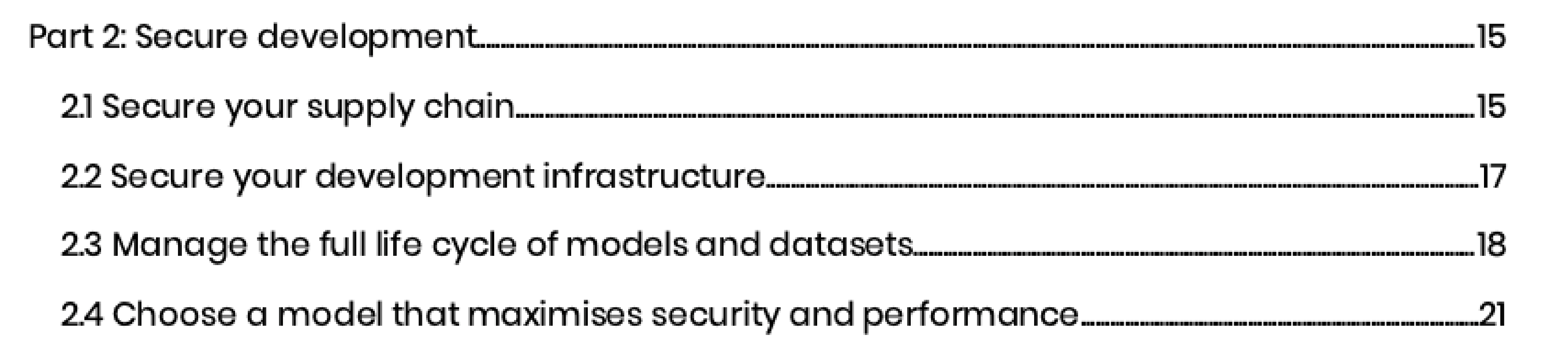
Pages

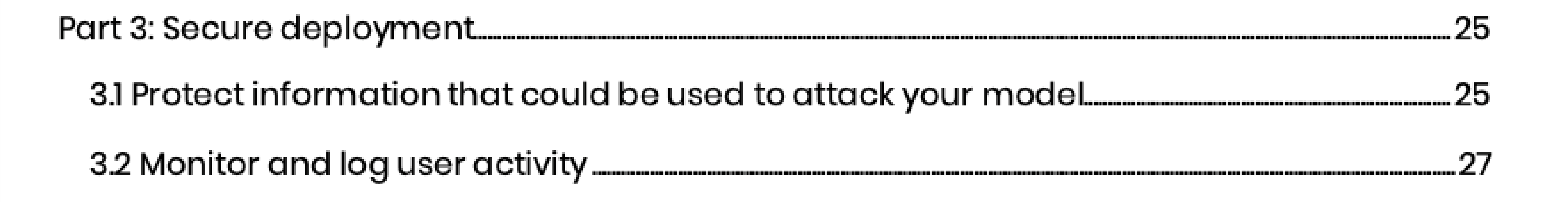
Machine learning principles

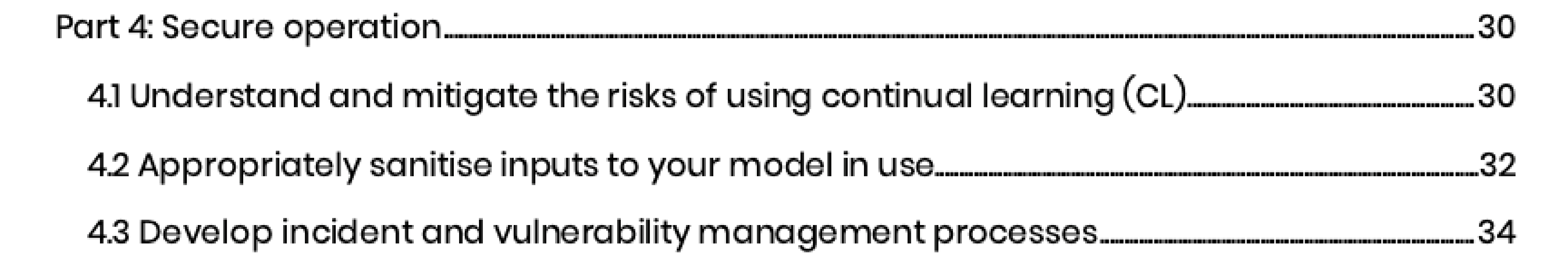
PAGE 1 OF 22

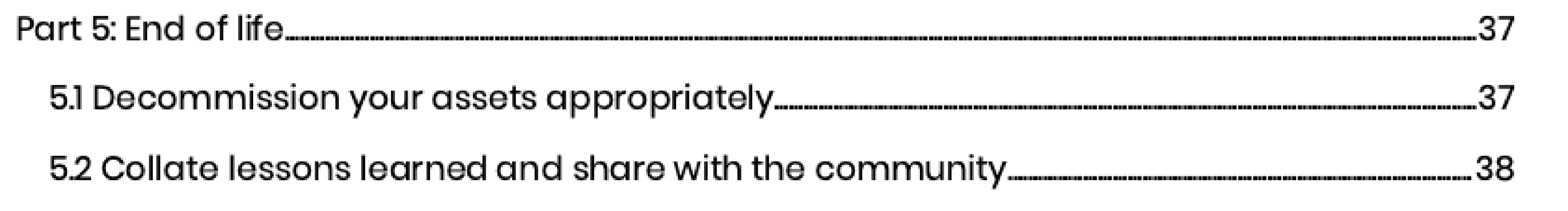
https://www.ncsc.gov.uk/files/NCSC-Machine-learning-principles.pdf



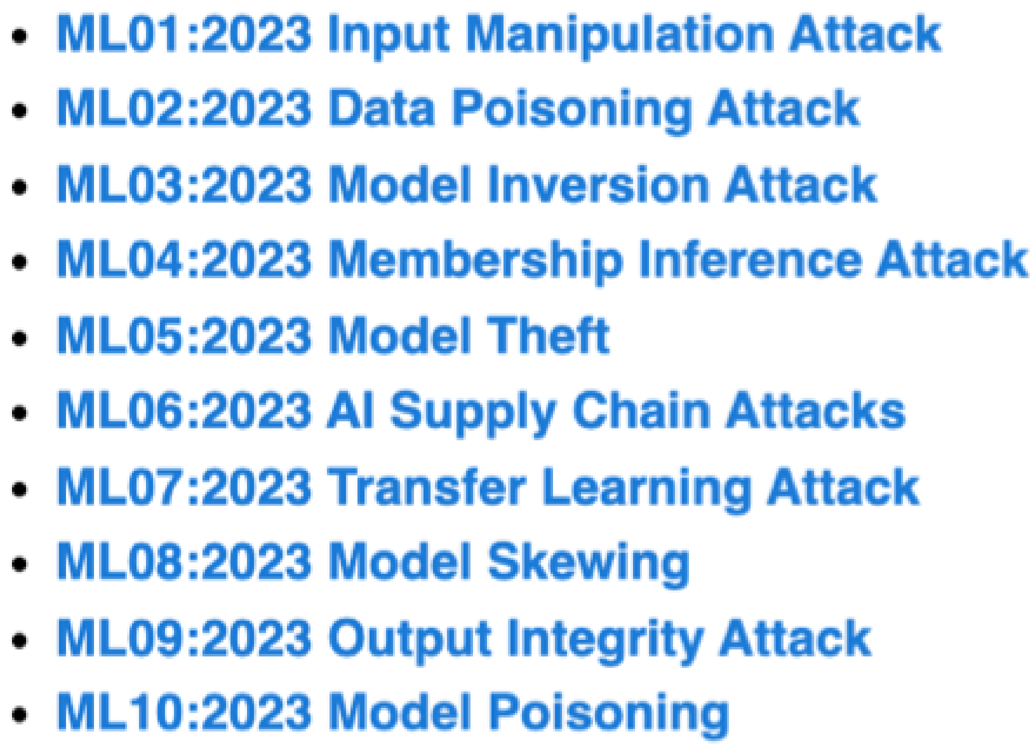








OWASP Top Ten Machine Learning Risks

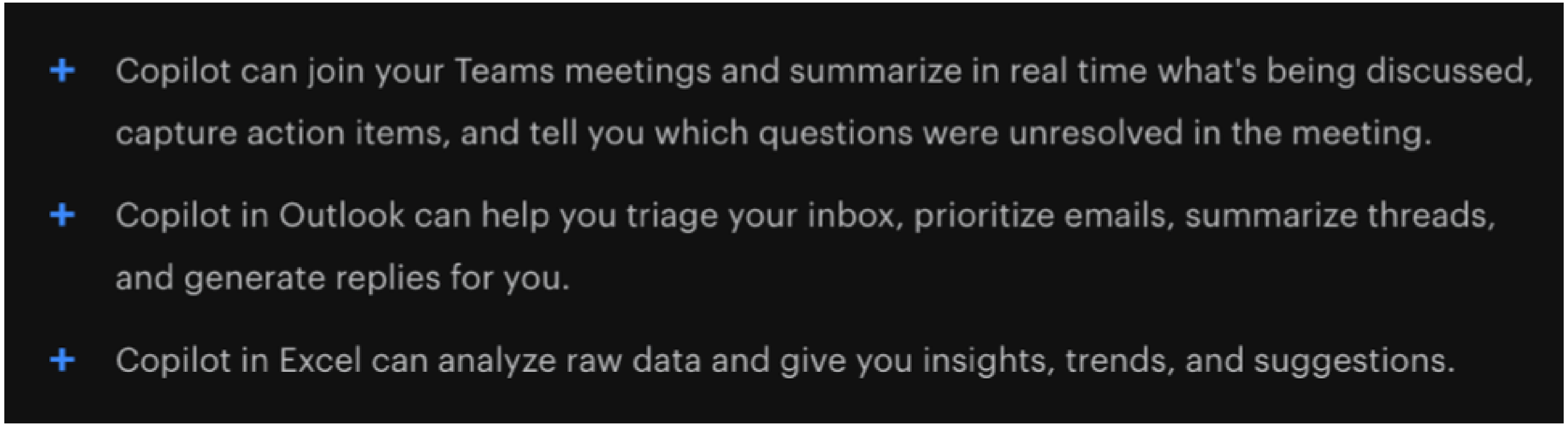


- ML01:2023 Input Manipulation Attack
- ML02:2023 Data Poisoning Attack
- ML03:2023 Model Inversion Attack
- ML04:2023 Membership Inference Attack
- ML05:2023 Model Theft
- ML06:2023 AI Supply Chain Attacks
- ML07:2023 Transfer Learning Attack
- ML08:2023 Model Skewing
- ML09:2023 Output Integrity Attack
- ML10:2023 Model Poisoning

- https://owasp.org/www-project-machine-learning-security-top-10/

W
X
T
P
O

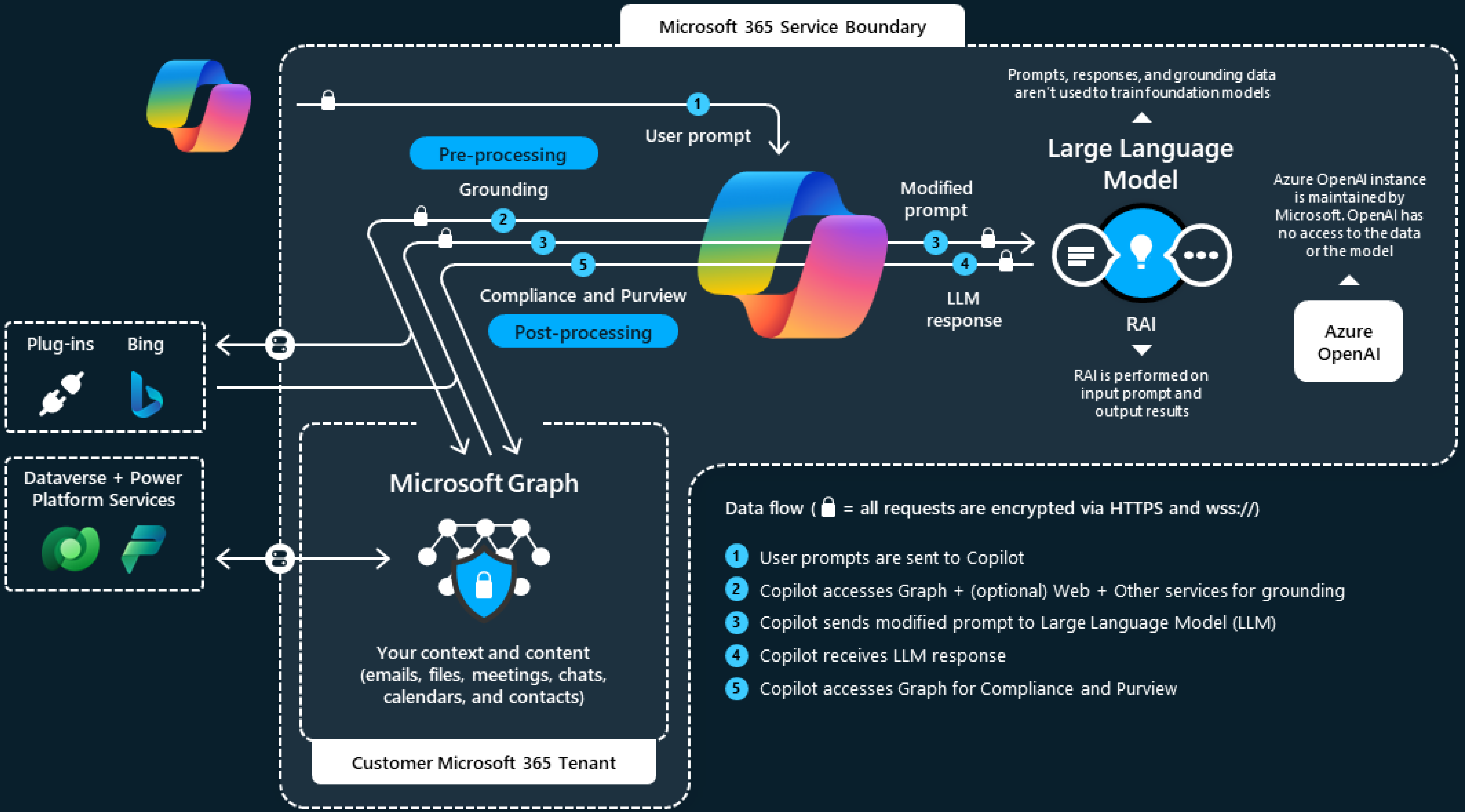
Microsoft 365 Copilot Use Cases



+ Copilot can join your Teams meetings and summarize in real time what's being discussed, capture action items, and tell you which questions were unresolved in the meeting.
+ Copilot in Outlook can help you triage your inbox, prioritize emails, summarize threads, and generate replies for you.
+ Copilot in Excel can analyze raw data and give you insights, trends, and suggestions.

- Writes documents for you
  - Based on data found in your Email, documents, spreadsheets, and other files you have access to
  - In the Microsoft365 cloud
  - **Based on your Microsoft365 permissions**

Microsoft Copilot for Microsoft 365 architecture


Microsoft 365 Service Boundary
1
User prompt
Pre-processing
Grounding
2
3
5
Compliance and Purview
Post-processing
Modified prompt
3
4
LLM response
Prompts, responses, and grounding data aren't used to train foundation models
Large Language Model
RAI
RAI is performed on input prompt and output results
Azure OpenAI instance is maintained by Microsoft. OpenAI has no access to the data or the model
Azure OpenAI
Plug-ins
Bing
Dataverse + Power Platform Services
Microsoft Graph
Your context and content (emails, files, meetings, chats, calendars, and contacts)
Customer Microsoft 365 Tenant

Data flow (🔒 = all requests are encrypted via HTTPS and wss://)

1. User prompts are sent to Copilot
2. Copilot accesses Graph + (optional) Web + Other services for grounding
3. Copilot sends modified prompt to Large Language Model (LLM)
4. Copilot receives LLM response
5. Copilot accesses Graph for Compliance and Purview

Data, Privacy, and Security for Microsoft Copilot for Microsoft 365

The information in this article is intended to help provide answers to the following questions:

How does Microsoft Copilot for Microsoft 365 use your proprietary organizational data?

How does Microsoft Copilot for Microsoft 365 protect organizational information and data?

What data is stored about user interactions with Microsoft Copilot for Microsoft 365?

What data residency commitments does Microsoft Copilot make?

What extensibility options are available for Microsoft Copilot for Microsoft 365

How does Microsoft Copilot for Microsoft 365 meet regulatory compliance requirements?

Do controls for connected experiences in Microsoft 365 Apps apply to Microsoft Copilot for Microsoft 365?

Can I trust the content that Microsoft Copilot for Microsoft 365 creates? Who owns that content?

What are Microsoft's commitments to using AI responsibly?

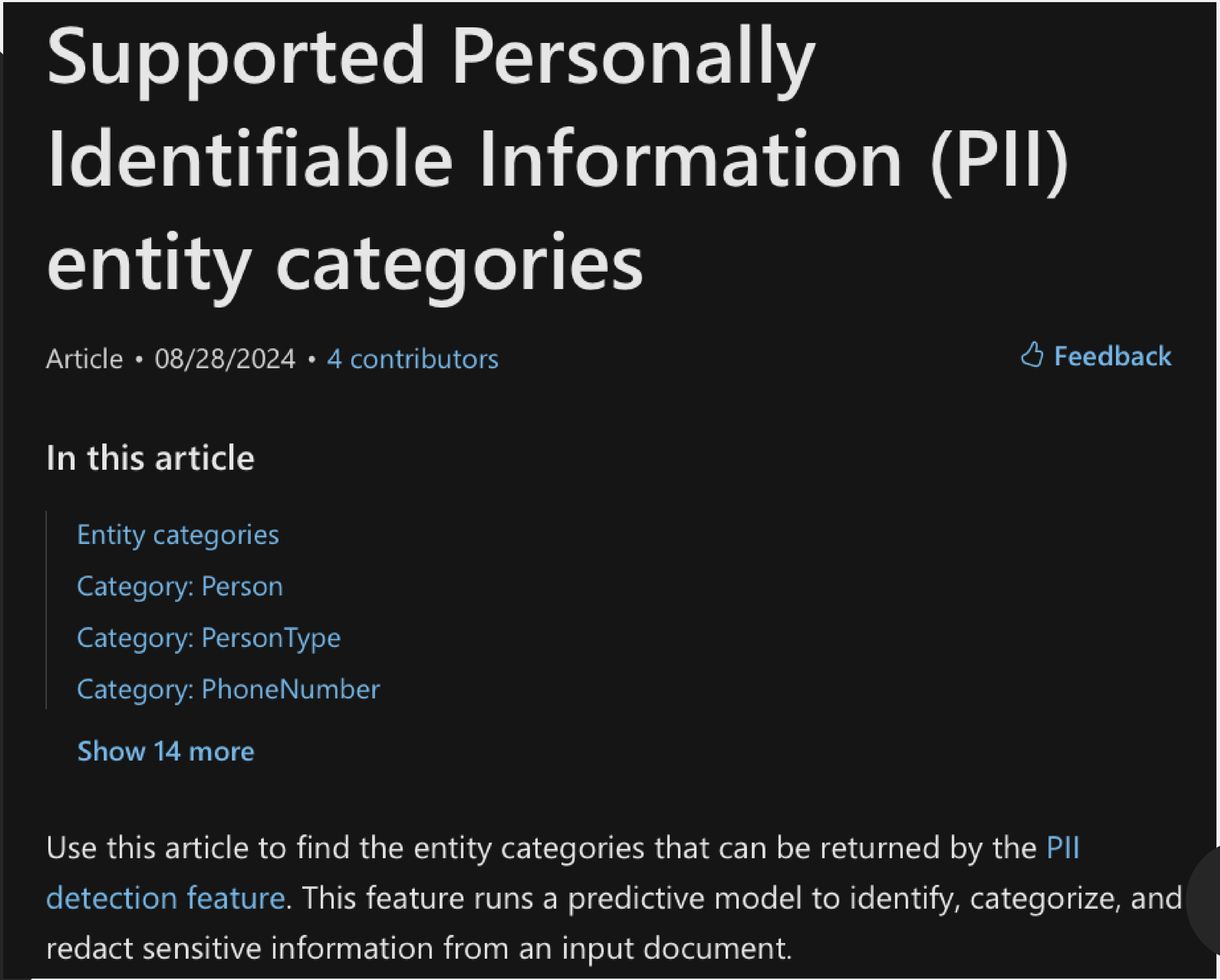
This data is processed and stored in alignment with contractual commitments with your organization’s other content in Microsoft 365. The data is encrypted while it's stored and isn't used to train foundation LLMs, including those used by Microsoft Copilot for Microsoft 365.

Data stored about user interactions with Microsoft Copilot for Microsoft 365

What the process looks like

We retain certain data from your interactions with us, but we take steps to reduce the amount of personal information in our training datasets before they are used to improve and train our models. This data helps us better understand user needs and preferences, allowing our model to become more efficient over time.

How your data is used to improve model performance: OpenAI



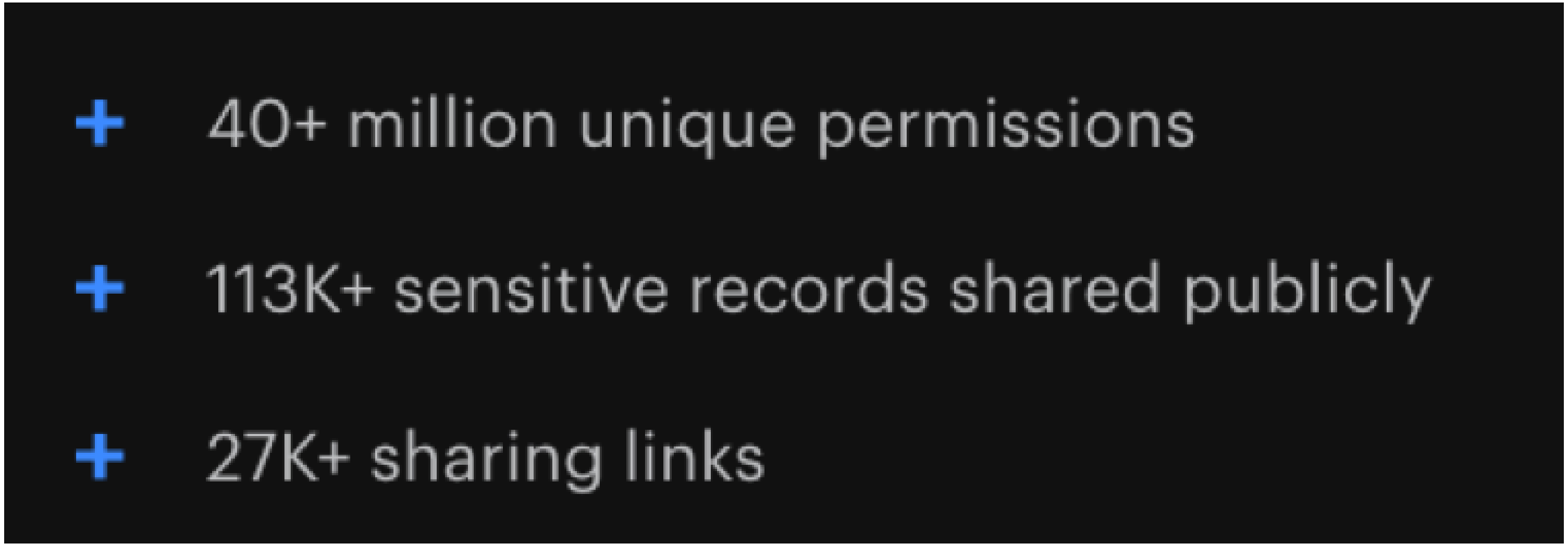
Supported Personally Identifiable Information (PII) entity categories

Article • 08/28/2024 • 4 contributors

In this article

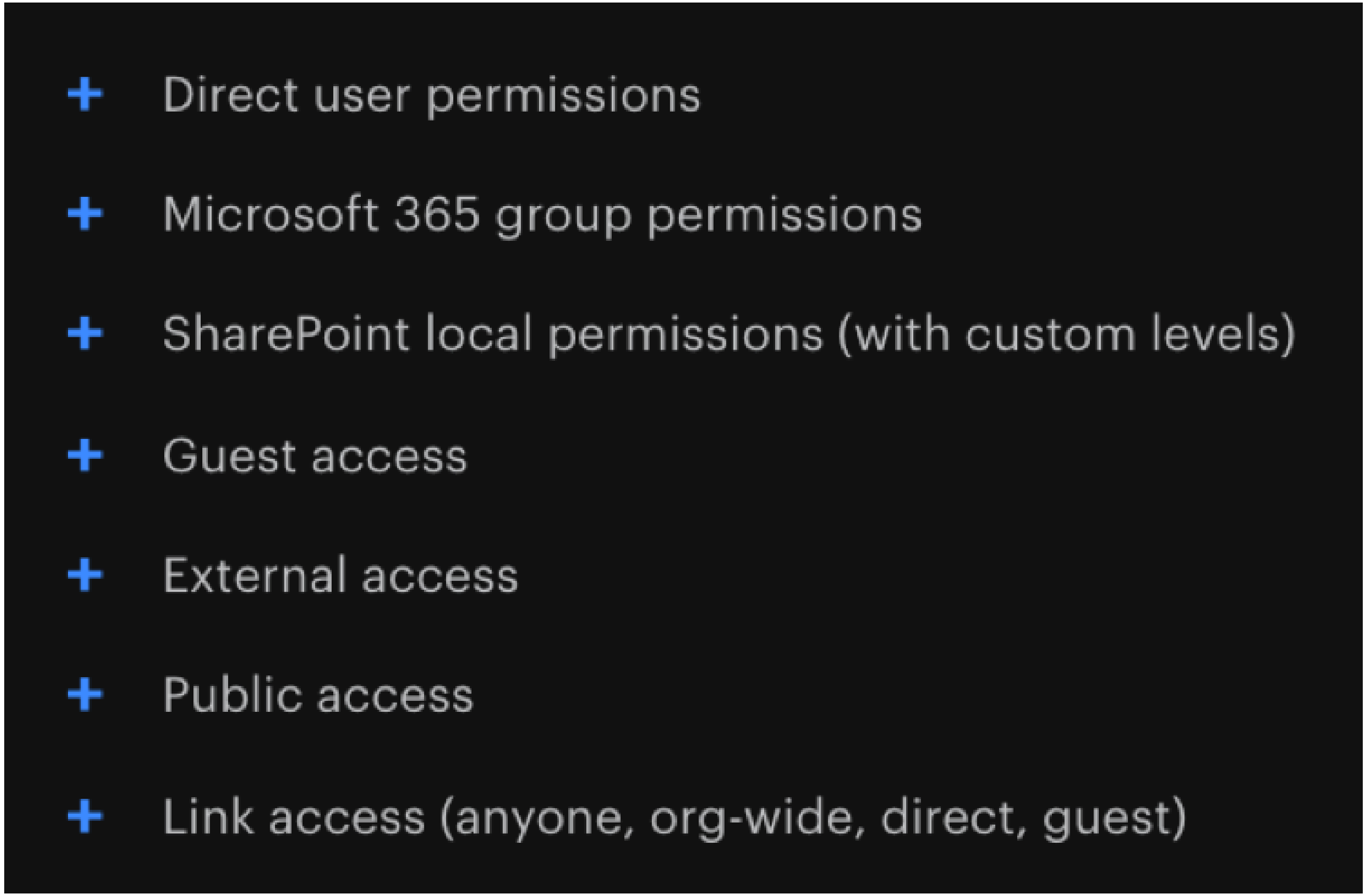
Use this article to find the entity categories that can be returned by the PII detection feature. This feature runs a predictive model to identify, categorize, and redact sensitive information from an input document.

The Average M365 Tenant has



+ 40+ million unique permissions
+ 113K+ sensitive records shared publicly
+ 27K+ sharing links

Why Does This Happen?



- Direct user permissions
- Microsoft 365 group permissions
- SharePoint local permissions (with custom levels)
- Guest access
- External access
- Public access
- Link access (anyone, org-wide, direct, guest)