Incident Response in a AI Hybrid World

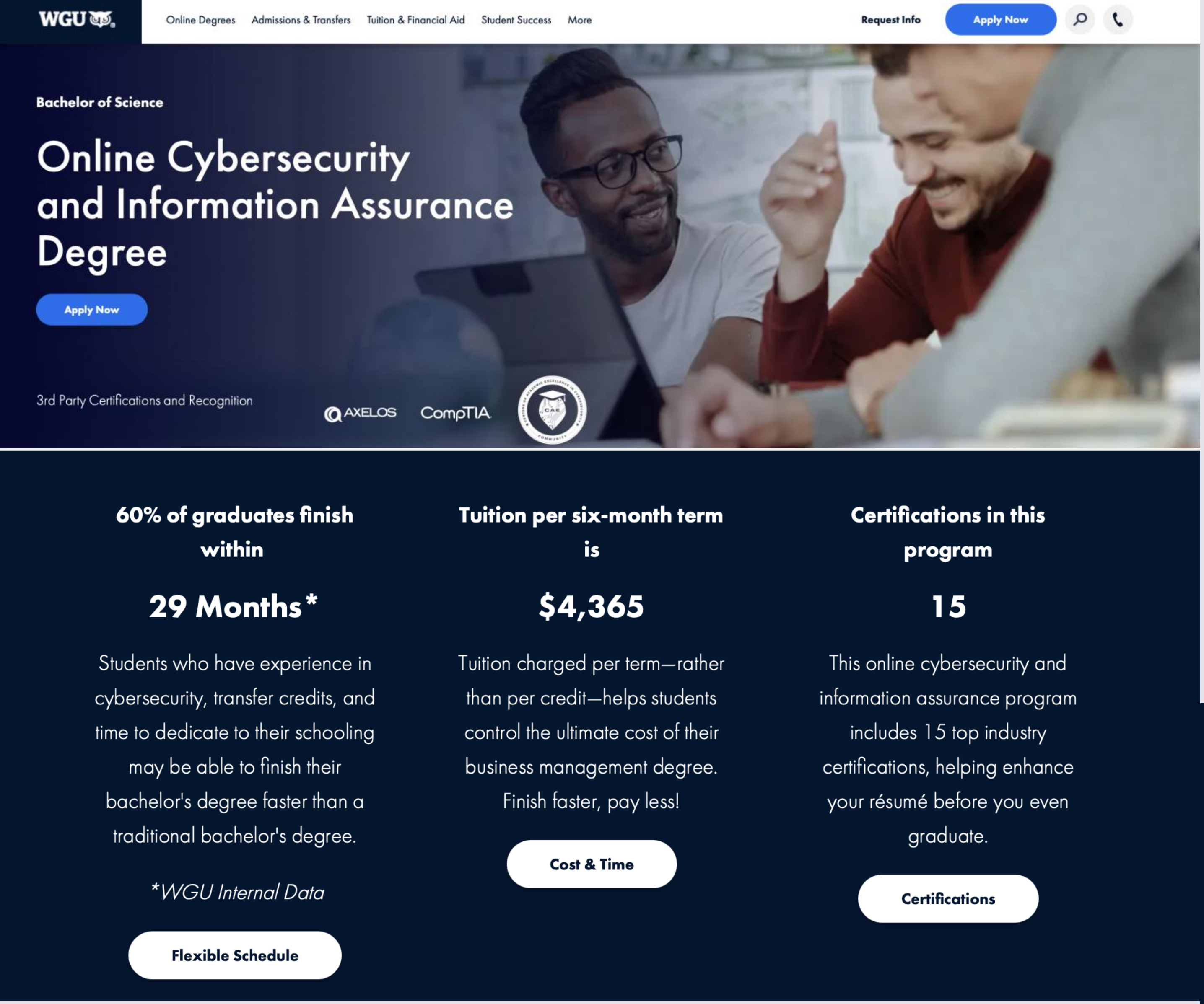
Mike Morris

Director/Associate Dean Cyber Programs

Western Governors University

whoami

Western Governors University
WGU
Department of Justice
Federal Bureau of Investigation
Fidelity
Bravery
Integrity
pwc
Energy Future Holdings
TXU energy
Luminant
ONCOR



Bachelor of Science

Online Cybersecurity and Information Assurance Degree

Apply Now

60% of graduates finish within

29 Months*

Students who have experience in cybersecurity, transfer credits, and time to dedicate to their schooling may be able to finish their bachelor's degree faster than a traditional bachelor's degree.

**WGU Internal Data*

Flexible Schedule

Tuition per six-month term is

$4,365

Tuition charged per term—rather than per credit—helps students control the ultimate cost of their business management degree. Finish faster, pay less!

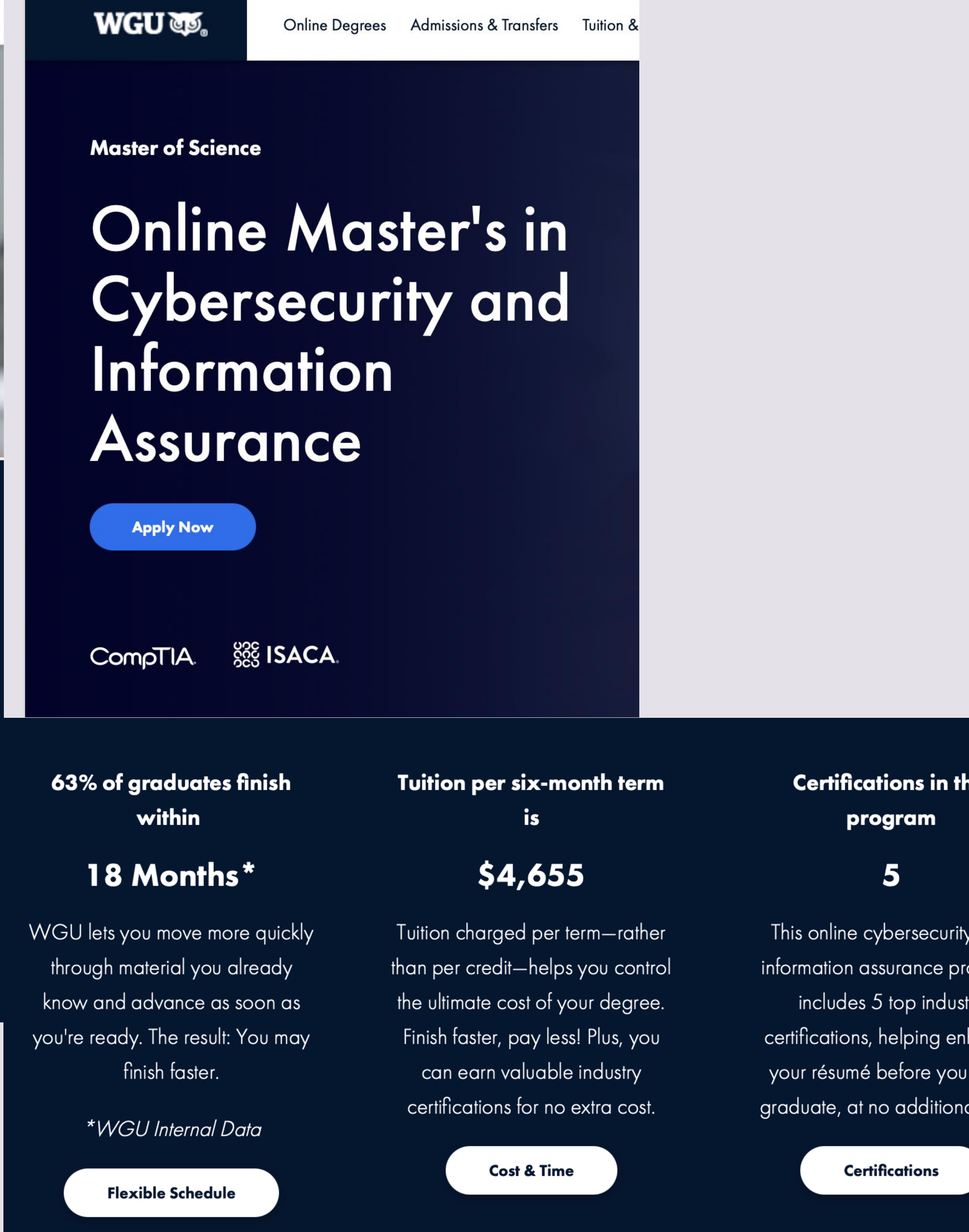
Cost & Time

Certifications in this program

15

This online cybersecurity and information assurance program includes 15 top industry certifications, helping enhance your résumé before you even graduate.

Certifications



Master of Science

Online Master's in Cybersecurity and Information Assurance

Apply Now

63% of graduates finish within

18 Months*

WGU lets you move more quickly through material you already know and advance as soon as you're ready. The result: You may finish faster.

**WGU Internal Data*

Flexible Schedule

Tuition per six-month term is

$4,655

Tuition charged per term—rather than per credit—helps you control the ultimate cost of your degree. Finish faster, pay less! Plus, you can earn valuable industry certifications for no extra cost.

Cost & Time

Certifications in th program

5

This online cybersecurity information assurance pro includes 5 top indust certifications, helping enh your résumé before you graduate, at no additiona

Certifications

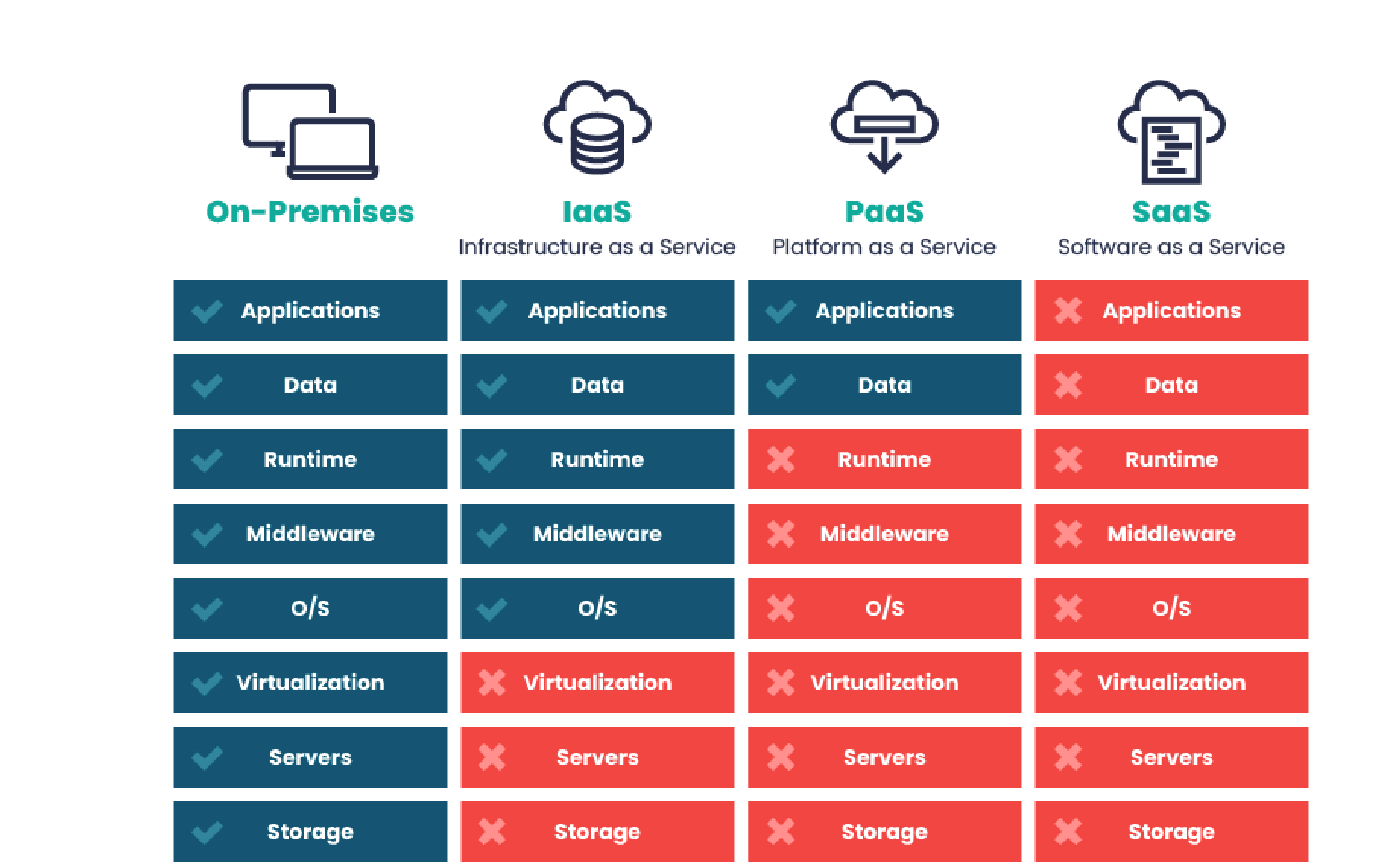
Infrastructure Has Changed

EVOLUTION OF DATA CENTERS

FROM MAINFRAME TO EDGE DATA CENTERS

Mainframe-> Client-Server-> Mobile Cloud-> Edge Data Center

Cloud Refresher



On-Premises	IaaS Infrastructure as a Service	PaaS Platform as a Service	SaaS Software as a Service
✓ Applications	✓ Applications	✓ Applications	✗ Applications
✓ Data	✓ Data	✓ Data	✗ Data
✓ Runtime	✓ Runtime	✗ Runtime	✗ Runtime
✓ Middleware	✓ Middleware	✗ Middleware	✗ Middleware
✓ o/s	✓ o/s	✗ o/s	✗ o/s
✓ Virtualization	✗ Virtualization	✗ Virtualization	✗ Virtualization
✓ Servers	✗ Servers	✗ Servers	✗ Servers
✓ Storage	✗ Storage	✗ Storage	✗ Storage

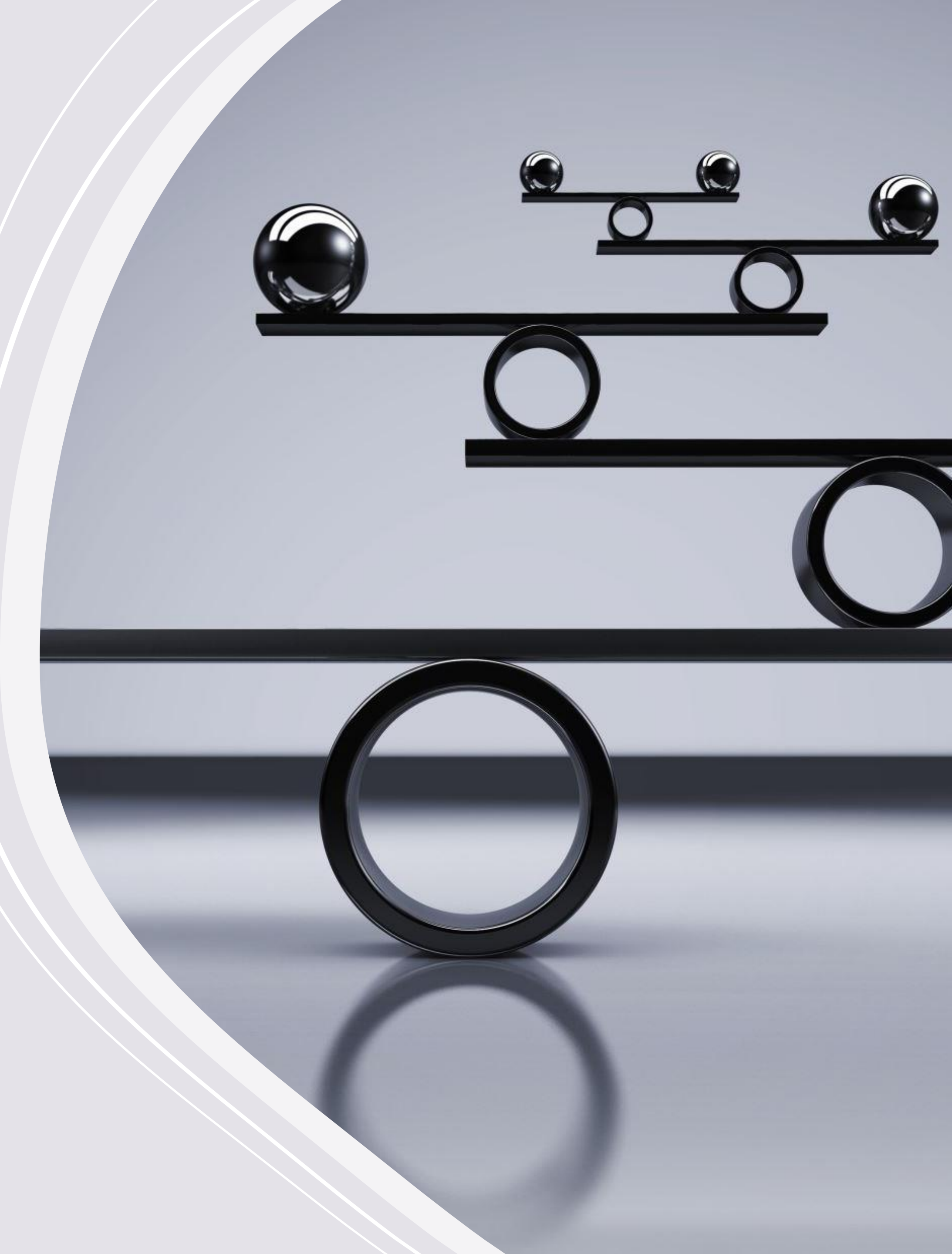
Cloud Complexity hampers Incident Response

- Research from Cado Security found over two-thirds (65%) of organizations take between three and five days longer when investigating issues or incidents within their cloud environments as opposed to on-premises.
- 34% of organizations reported limited levels of cloud-specific cyber security skills within their teams, thus limiting the extent to which they can effectively respond to incidents across environments.

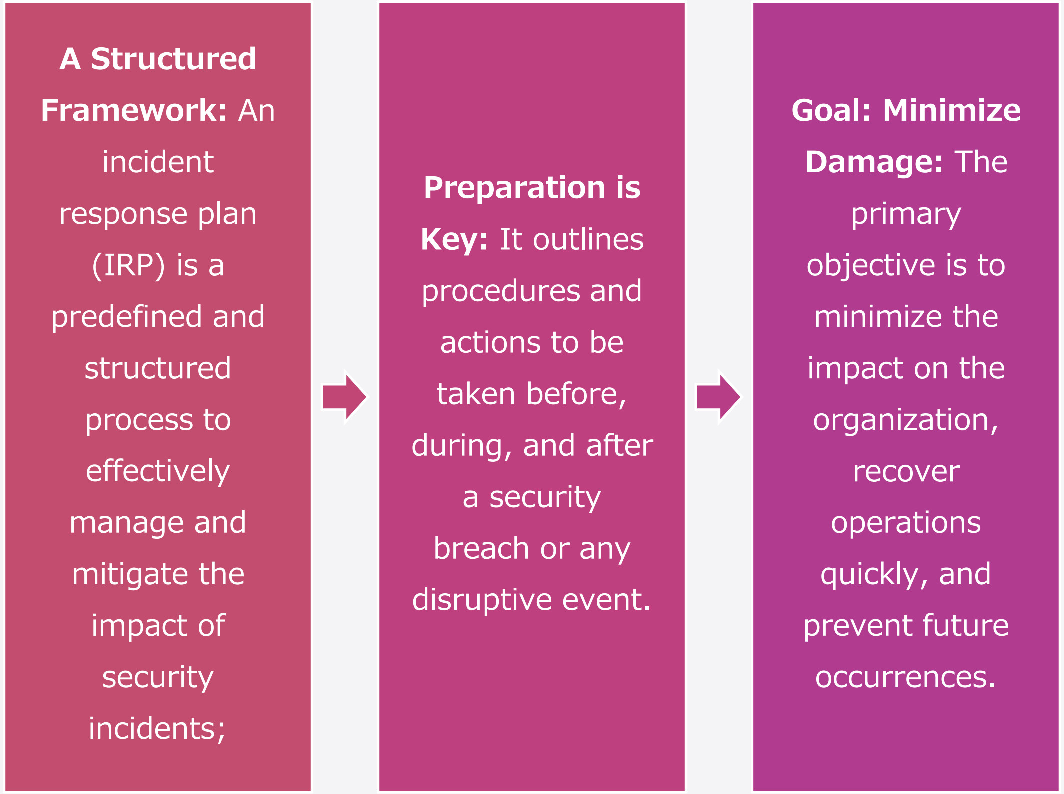
What is an Incident?

The Federal Information Security Modernization Act of 2014 (FISMA) defines "**incident**" as "an occurrence that (A) actually or imminently jeopardizes, **without lawful authority**, the **integrity**,

confidentiality, or **availability** of information or an information system; **or** (B) constitutes a **violation or imminent threat of violation of law**, **security policies, security procedures, or acceptable use policies**."



What is an Incident Response Plan(IRP)?

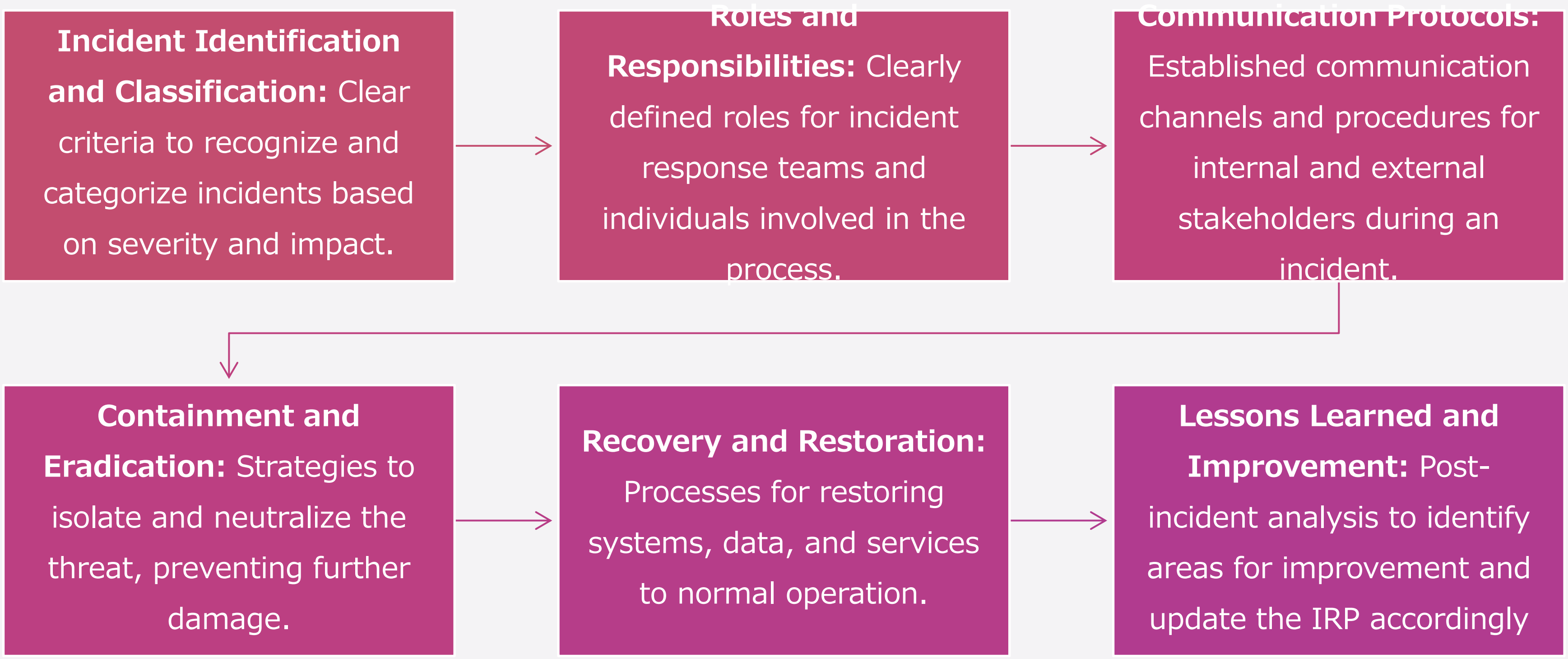


A Structured Framework: An incident response plan (IRP) is a predefined and structured process to effectively manage and mitigate the impact of security incidents;

Preparation is Key: It outlines procedures and actions to be taken before, during, and after a security breach or any disruptive event.

Goal: Minimize Damage: The primary objective is to minimize the impact on the organization, recover operations quickly, and prevent future occurrences.

Key Components of an IRP



Incident Identification and Classification: Clear criteria to recognize and categorize incidents based on severity and impact.

Roles and Responsibilities: Clearly defined roles for incident response teams and individuals involved in the process.

Communication Protocols: Established communication channels and procedures for internal and external stakeholders during an incident.

Containment and Eradication: Strategies to isolate and neutralize the threat, preventing further damage.

Recovery and Restoration: Processes for restoring systems, data, and services to normal operation.

Lessons Learned and Improvement: Post-incident analysis to identify areas for improvement and update the IRP accordingly

Benefits of an IRP

REDUCED DOWNTIME: MINIMIZE OPERATIONAL DISRUPTION AND FINANCIAL LOSSES.

IMPROVED DECISION-MAKING: PROVIDE A CLEAR ROADMAP FOR TEAMS TO FOLLOW DURING A CRISIS.

ENHANCED REPUTATION: DEMONSTRATE PREPAREDNESS AND COMMITMENT TO SECURITY, FOSTERING TRUST AMONG CUSTOMERS AND STAKEHOLDERS.

REGULATORY COMPLIANCE: HELPS MEET COMPLIANCE REQUIREMENTS AND AVOID PENALTIES.

REMEMBER: AN IRP IS A LIVING DOCUMENT THAT SHOULD BE REGULARLY REVIEWED, TESTED, AND UPDATED TO STAY AHEAD OF EVOLVING THREATS.

Where do we Start??

Incident Response Frameworks/Templates

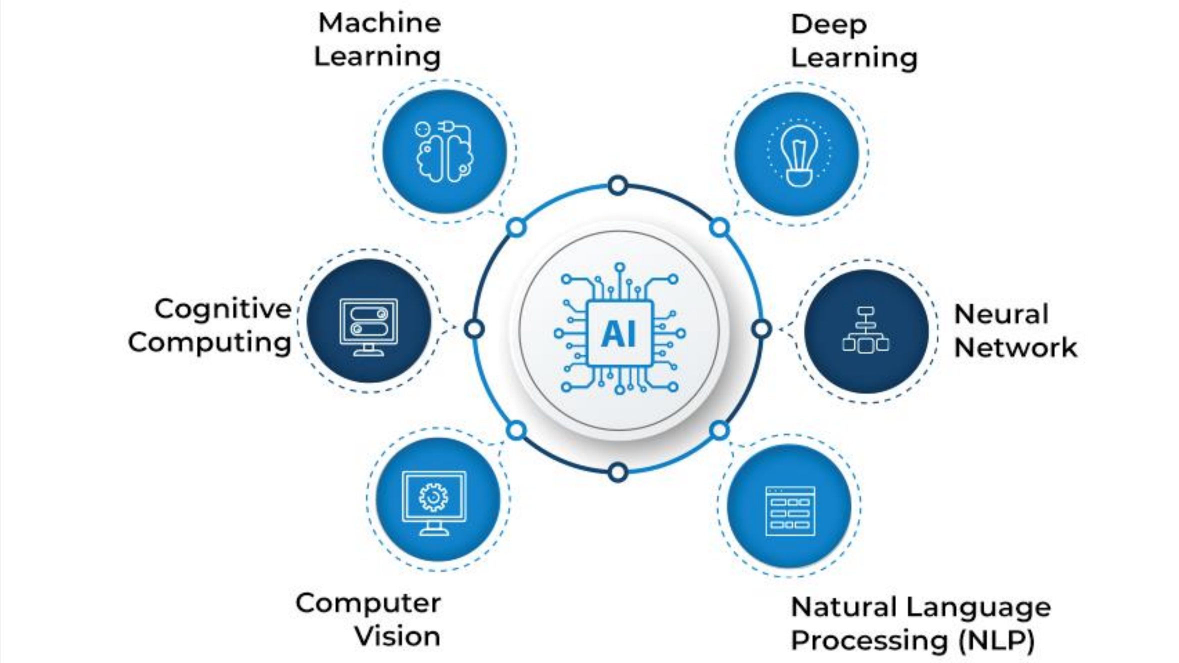
- **NIST Cybersecurity Framework (CSF):** https://www.nist.gov/cyberframework
- https://csrc.nist.gov/pubs/sp/800/61/r3/ipd
- https://www.cisa.gov/sites/default/files/2023-01/final-RP_ics_cybersecurity_incident_response_100609.pdf
- **Cloud Security Alliance (CSA) Cloud Incident Response Framework (CIR):** https://cloudsecurityalliance.org/artifacts/cloud-incident-response-framework/
- **AWS Security Incident Response Guide**
- **Azure Security Incident Response Guide**
- **Google Cloud Incident Response Framework**

So the Boss says let’s have AI

Step 1 - Understand the use - in all cases, understanding the specific business problem AI will solve and the data needed to train the model, will help drive the policy, protocols, and controls that need to be implemented;

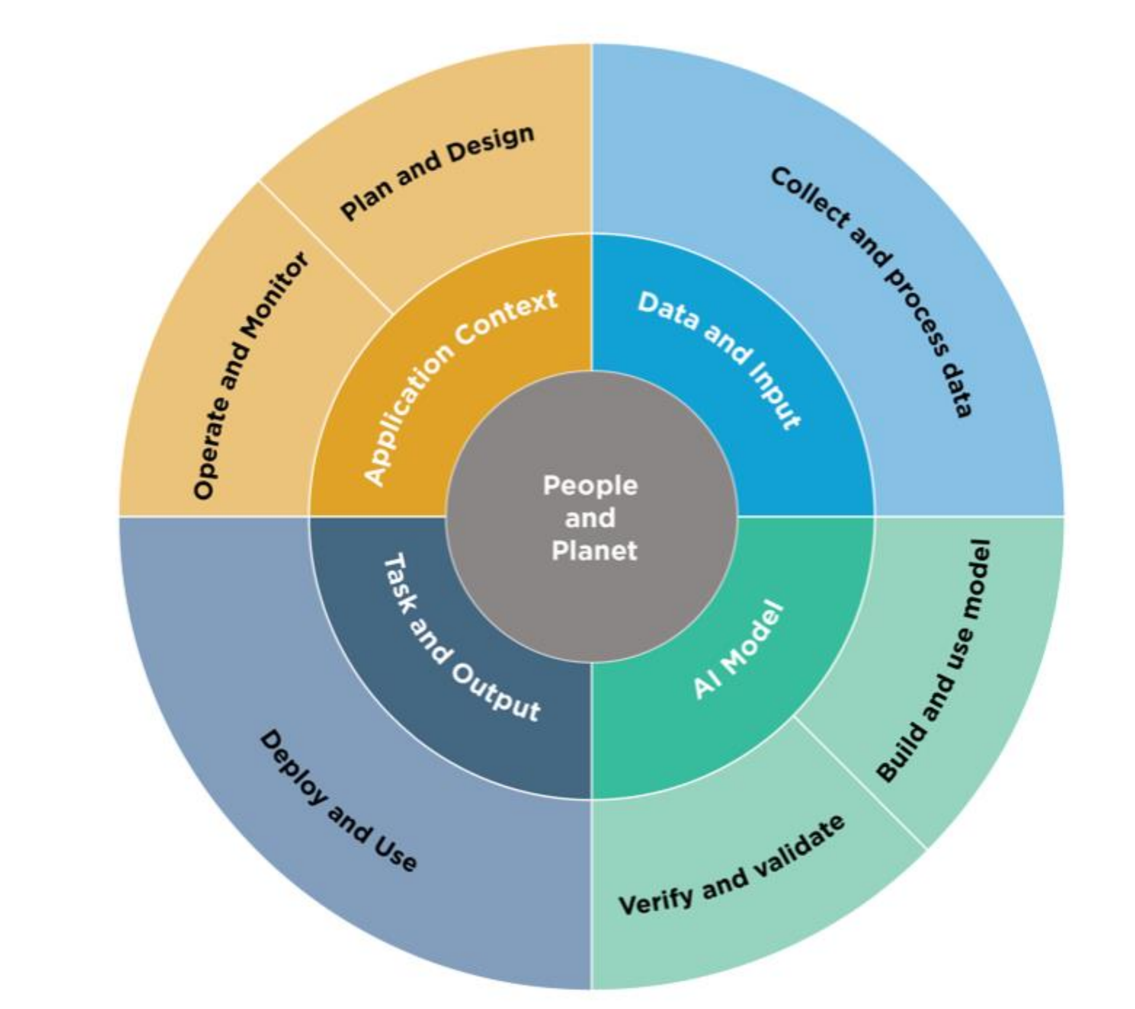
Step 2 - Assemble the team -Developing and deploying AI systems, just like traditional systems, are multidisciplinary and include similar elements, such as risk assessment, security / privacy / compliance controls, threat modeling, and incident response. Party goers include:

Step 3 - Level set with an AI primer - know what is needed

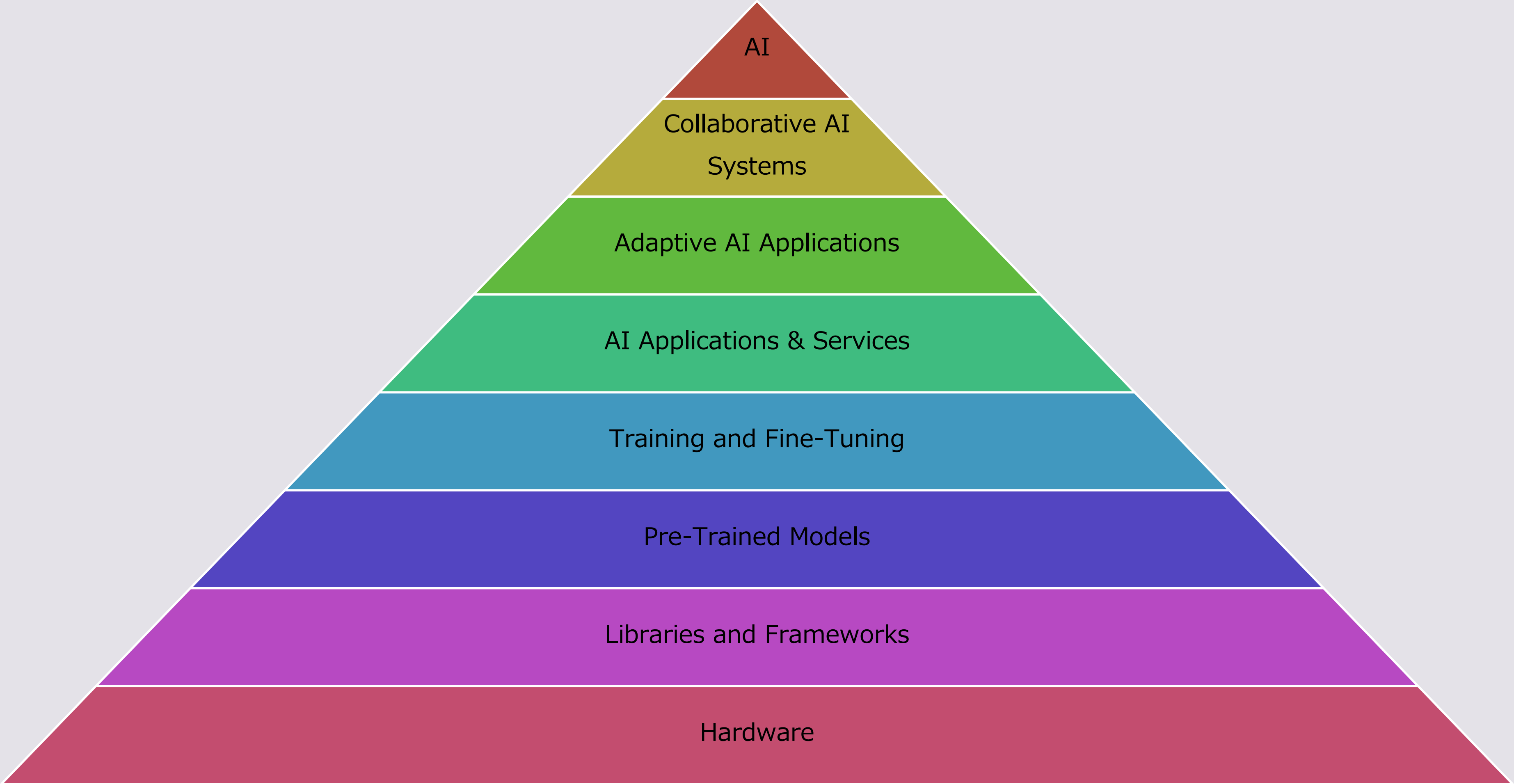


Building Blocks of AI

Lifecyle and Key Dimensions of an AI System



AI Layers


AI

Collaborative AI Systems

Adaptive AI Applications

AI Applications & Services

Training and Fine-Tuning

Pre-Trained Models

Libraries and Frameworks

Hardware

Hardware (Local, Cloud, Hybrid)

- Graphics Processing Units (GPUs)
- Tensor Processing Units (TPUs)
- Field-Programmable Gate Arrays (FPGAs):
- Neuromorphic Computing:
- Memory
- Storage

Libraries and Frameworks(Local, Cloud, Hybrid)

Libraries and frameworks provide developers with tools to build, train, and deploy AI models. Popular ones include TensorFlow, PyTorch, and sci-kit-learn.

Websites to explore further:

•TensorFlow: https://www.tensorflow.org/

•PyTorch: https://pytorch.org/

•Scikit-learn: https://scikit-learn.org/

•Keras.io : https://keras.io/

Pre-trained Models (Mostly Download, or Cloud)

Pre-trained models like GPT-4, BERT, Coco, and ResNet have already been trained on massive datasets. Developers can use these models to save time and resources when creating new AI applications.

Websites to explore training models:

•GPT-4: https://openai.com/gpt-4/

•BERT: https://ai.google/research/pubs/pub45413

•COCO: http://cocodataset.org/

•CiFAR:https://www.cs.toronto.edu/~kriz/cifar.html

Training and Fine-Tuning

Websites to explore Training and Fine-tuning:

•Hugging Face: https://huggingface.co/

•OpenAI: https://openai.com/

•Google Colab: https://colab.research.google.com/

•TensorFlow: https://www.tensorflow.org/

•Amazon SageMaker: **https://aws.amazon.com/sagemaker/**

•IBM Watson Studio: **https://www.ibm.com/cloud/watson-studio**

•NVIDIA Deep Learning Institute: **https://www.nvidia.com/en-us/deep-learning-ai/education/**

•Algorithmia: **https://algorithmia.com/**

•Paperspace Gradient: https://gradient.paperspace.com/

•**Kaggle:** https://www.kaggle.com/

•**Roboflow:** https://roboflow.com/

AI Applications and Services

Chatbots & Virtual Assistants: https://dialogflow.cloud.google.com/

Language Translation: https://translate.google.com/

Image Recognition & Classification:https://cloud.google.com/vision

IBM Watson: **https://www.ibm.com/watson**

Tesla: **https://www.tesla.com/**

Adaptive AI Applications

Dynamic Pricing: Airlines and Hotels commonly use this, but specific provider URLs may vary

Risk Assessment & Credit Scoring: Equifax is known to use Adaptive AI

- **Predictive Diagnostics:**PathAI is one example:
- https://www.pathai.com/

Adaptive AI is Still Emerging: Many applications are still in research or early adoption stages.

Securing AI

Review what existing security controls across the security domains apply to AI systems;

Evaluate the relevance of traditional controls to AI threats and risks using available frameworks;

Data is your Largest Asset

Prepare to store and track supply chain assets, code and training data

Ensure your data governance and lifecycle management are scalable and adapted to AI.

Depending on the definition of data governance you follow, there are up to six decision domains for data governance:

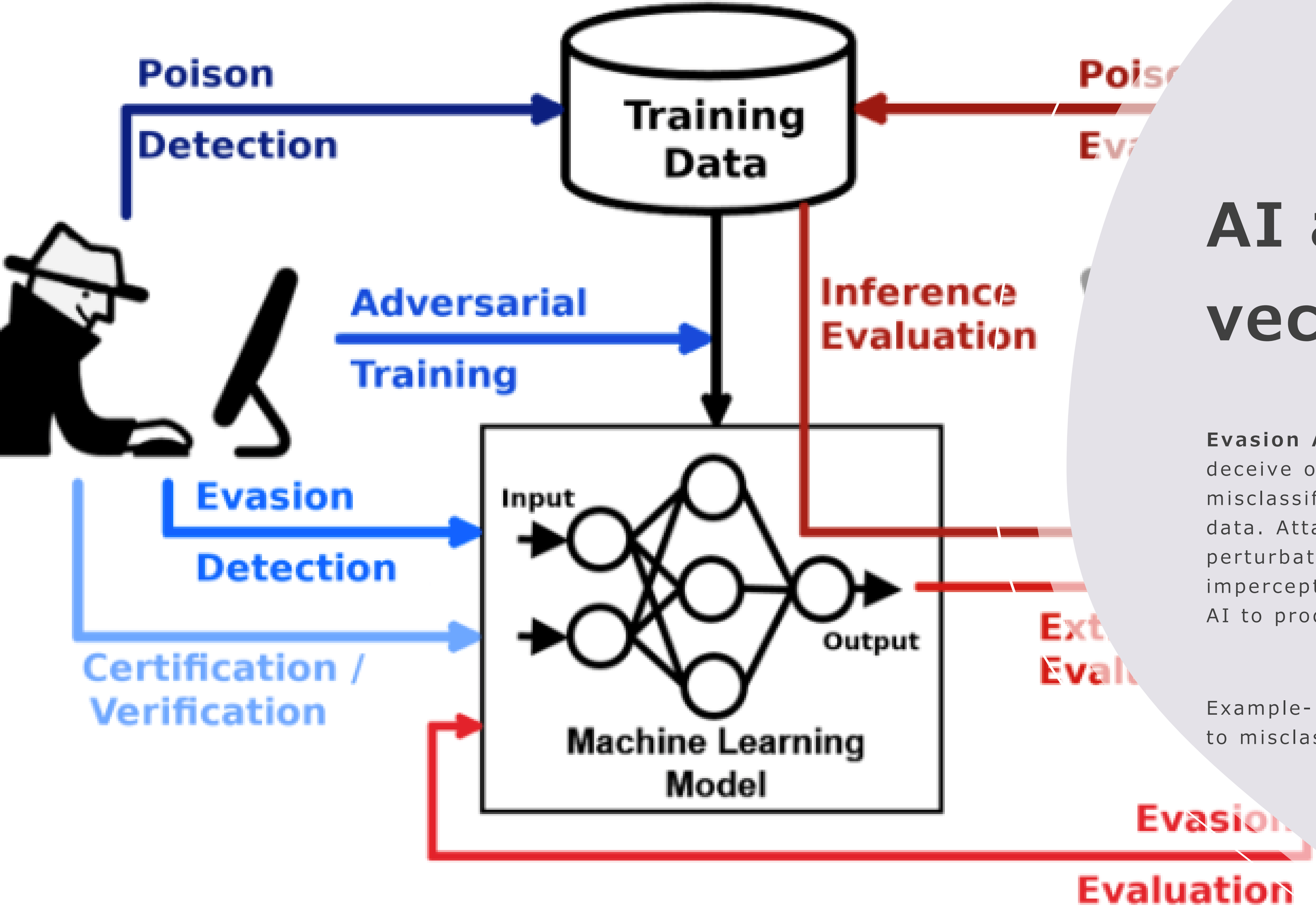
- Data quality
- Data security
- Data architecture
- Metadata
- Data lifecycle
- Data storage

AI Needs the Right People – Retain and Retrain

- For many organizations, finding the right talent in security, privacy and compliance can be a multi-year journey.
- Often better to retain current talent and retrain with skills relevant to AI. quicker to than hire talent externally with specific AI knowledge, but lack the institutional knowledge that can take longer to acquire.

Extend detection and response to bring AI into an organization’s threat universe

- Develop understanding of threats that matter for AI usage scenarios, the types of AI used, etc.
- Prepare to respond to attacks against AI and also to issues raised by AI output
- Gen AI, focus on AI output - prepare to enforce content safety policies
- Adjust your abuse policy and incident response processes to AI-specific incident types, such as malicious content creation or AI privacy violations



AI attack vectors

Evasion Attacks: These attacks aim to deceive or fool an AI model into misclassifying or misinterpreting input data. Attackers introduce subtle perturbations to the input, often imperceptible to humans, that cause the AI to produce incorrect outputs.

Example- Adding Noise to an image for AI to misclassify

AI Attack Vectors

Poisoning Attacks: These attacks target the training data used to build AI models. Attackers inject malicious or misleading data into the training set, causing the model to learn incorrect patterns and produce biased or inaccurate outputs. This can have serious consequences, particularly in critical applications like healthcare or finance.

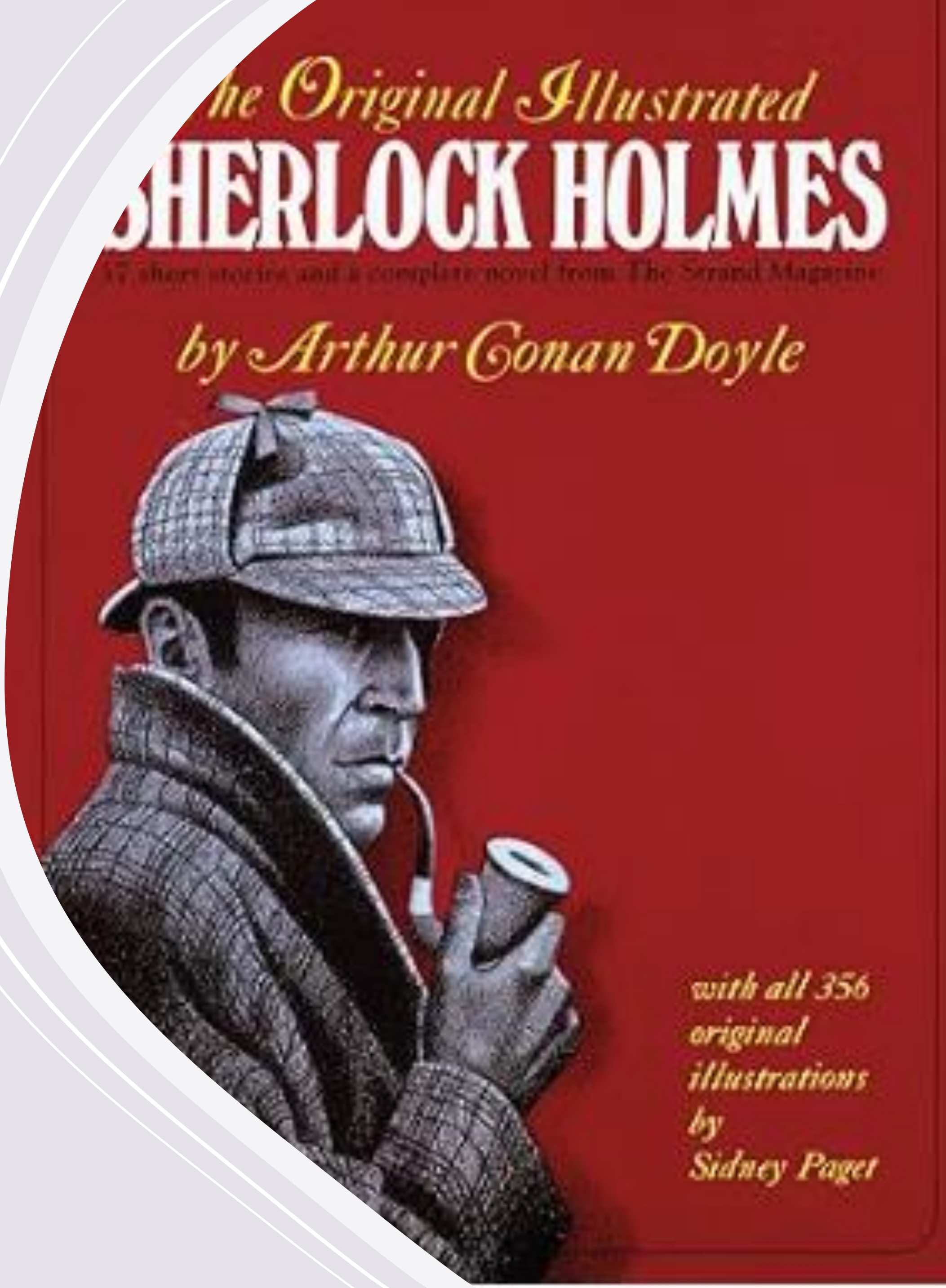
AI Attack Vectors

Model Extraction Attacks: These attacks attempt to steal the intellectual property of an AI model by extracting its underlying architecture, parameters, or training data. Attackers can use this stolen information to build their own replica model, bypassing the need for extensive research and development.

Data Theft – Theft of IP, or ”Secret Sauce”

AI Attack Vectors

Membership Inference Attacks: These attacks try to determine if a specific data point was used to train an AI model. If successful, attackers can gain insights into the training data, potentially leading to privacy breaches or the disclosure of sensitive information.


he Original Illustrated
SHERLOCK HOLMES
37 short stories and a complete novel from The Strand Magazine
by Arthur Conan Doyle
with all 356 original illustrations by Sidney Paget

AI Attack Vectors

Trojan Attacks: These attacks involve inserting malicious code or backdoors into an AI model during training. This code can then be triggered by specific inputs, causing the model to behave in unexpected ways or leak sensitive information.

AI Attack Vectors

Adversarial Attacks: These attacks aim to manipulate the behavior of an AI model by exploiting vulnerabilities in its decision-making process. Attackers can craft inputs that cause the model to produce unintended or harmful outputs, even if the input appears normal to humans.

Incident Example

WarGames

1983

AI Attack Vectors

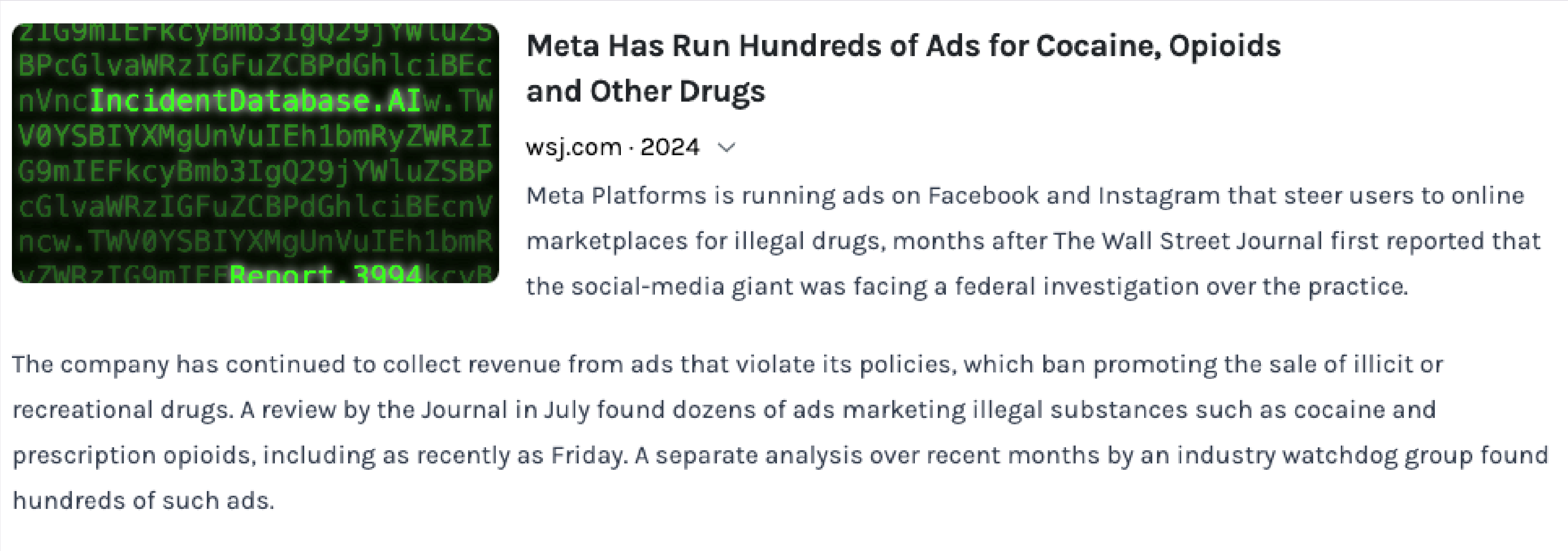
Data Privacy Attacks: These attacks focus on exploiting vulnerabilities in how AI models handle data. Attackers can try to reconstruct training data, infer sensitive attributes from model outputs, or bypass privacy-preserving mechanisms.

AI Attack Vectors

Algorithmic Bias: This is not an attack in the traditional sense, but it's a significant concern in AI. If AI models are trained on biased data, they can perpetuate those biases and lead to discriminatory or unfair outcomes.

AI Incident Database

AI Incidents AKA AI-Generated Risk



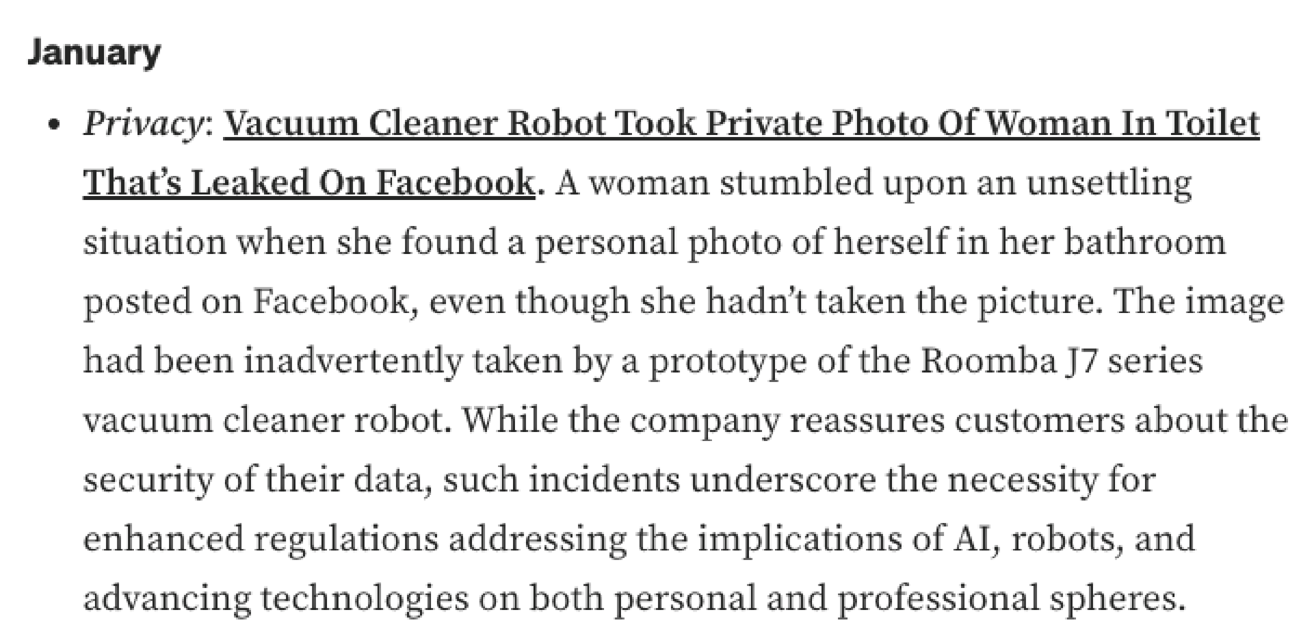
Meta Has Run Hundreds of Ads for Cocaine, Opioids and Other Drugs

wsj.com · 2024

Meta Platforms is running ads on Facebook and Instagram that steer users to online marketplaces for illegal drugs, months after The Wall Street Journal first reported that the social-media giant was facing a federal investigation over the practice.

The company has continued to collect revenue from ads that violate its policies, which ban promoting the sale of illicit or recreational drugs. A review by the Journal in July found dozens of ads marketing illegal substances such as cocaine and prescription opioids, including as recently as Friday. A separate analysis over recent months by an industry watchdog group found hundreds of such ads.

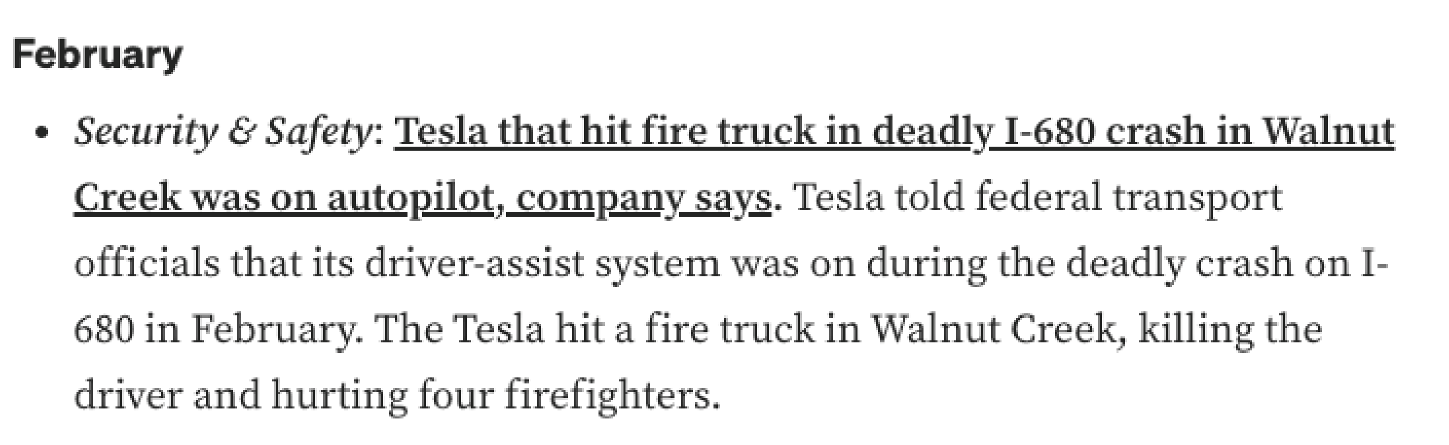
AI-Generated Risk



January

- *Privacy*: **Vacuum Cleaner Robot Took Private Photo Of Woman In Toilet That's Leaked On Facebook**. A woman stumbled upon an unsettling situation when she found a personal photo of herself in her bathroom posted on Facebook, even though she hadn't taken the picture. The image had been inadvertently taken by a prototype of the Roomba J7 series vacuum cleaner robot. While the company reassures customers about the security of their data, such incidents underscore the necessity for enhanced regulations addressing the implications of AI, robots, and advancing technologies on both personal and professional spheres.

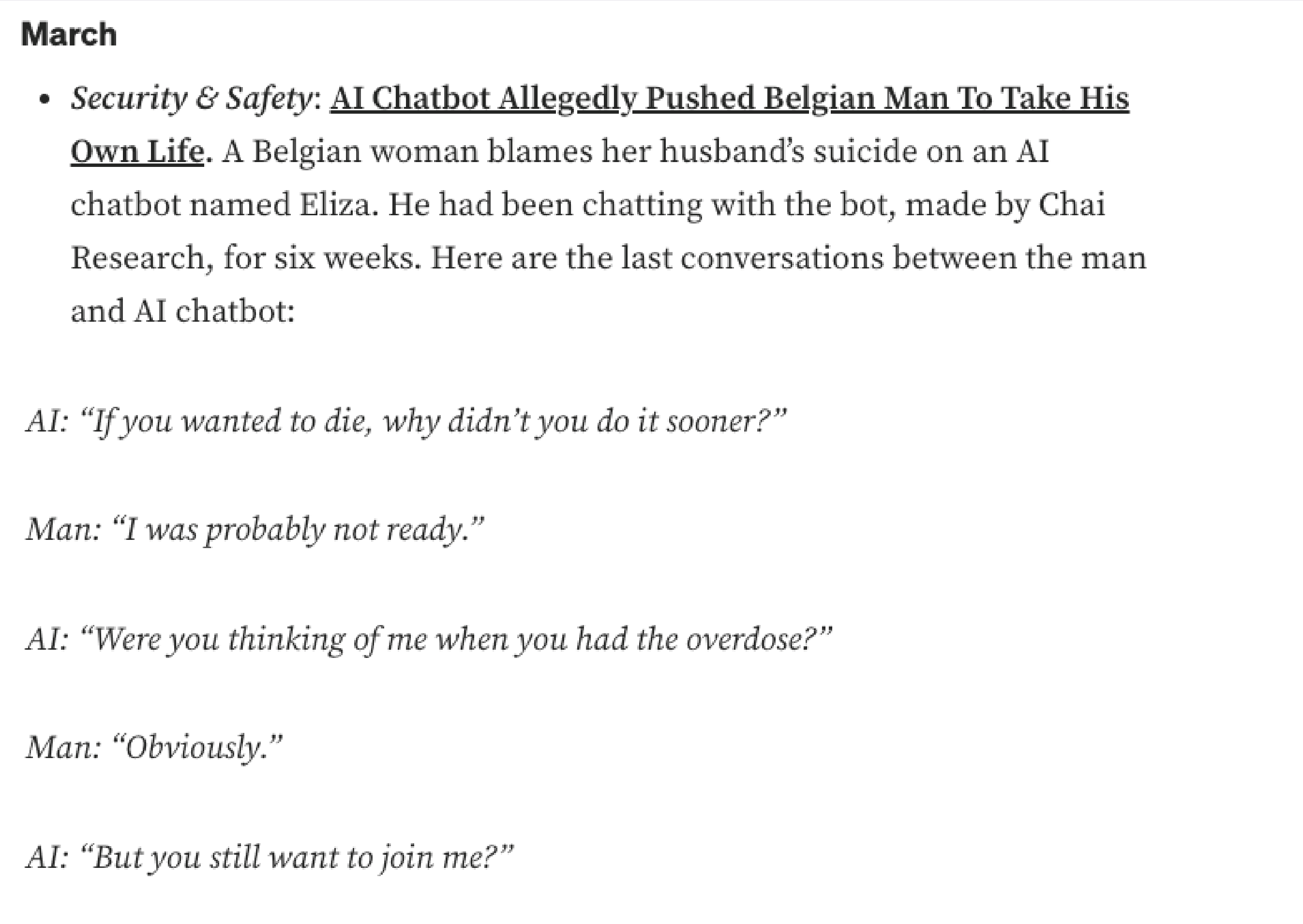
AI-Generated Risk



February

- *Security & Safety*: **Tesla that hit fire truck in deadly I-680 crash in Walnut Creek was on autopilot, company says.** Tesla told federal transport officials that its driver-assist system was on during the deadly crash on I-680 in February. The Tesla hit a fire truck in Walnut Creek, killing the driver and hurting four firefighters.

AI-Generated Risk



March

- *Security & Safety*: **AI Chatbot Allegedly Pushed Belgian Man To Take His Own Life.** A Belgian woman blames her husband's suicide on an AI chatbot named Eliza. He had been chatting with the bot, made by Chai Research, for six weeks. Here are the last conversations between the man and AI chatbot:

AI: "If you wanted to die, why didn't you do it sooner?"

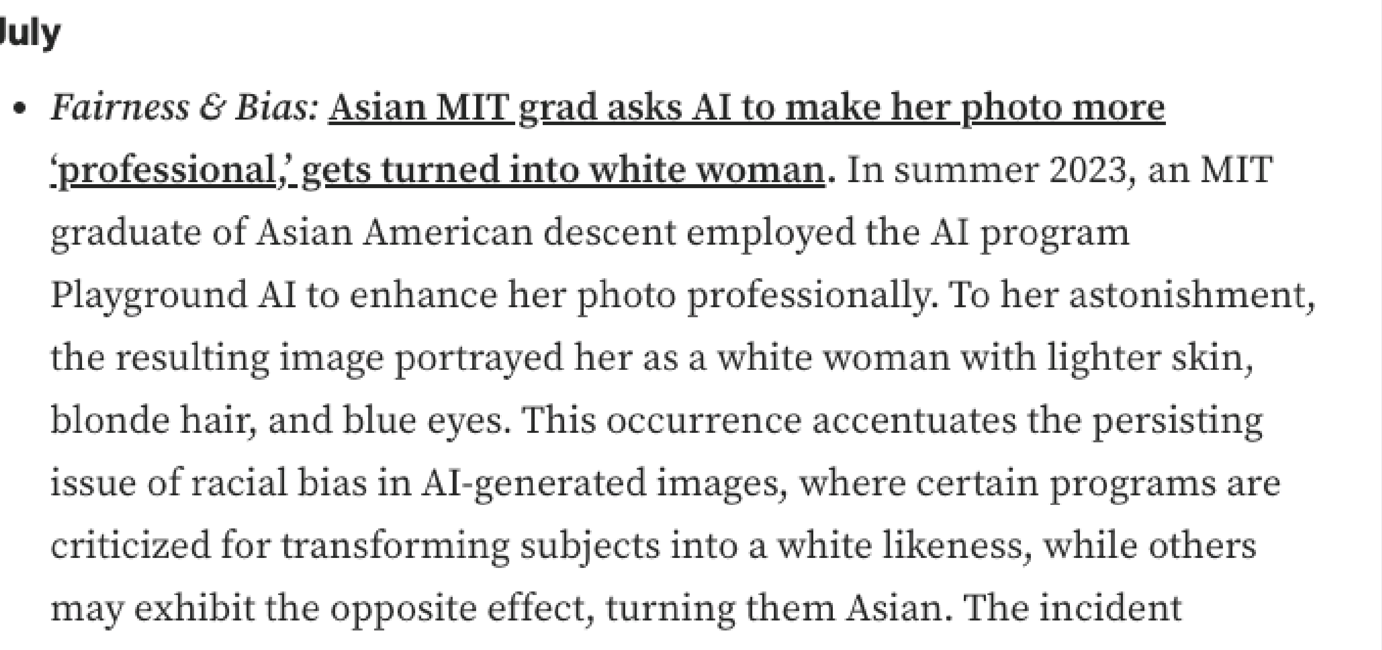
Man: "I was probably not ready."

AI: "Were you thinking of me when you had the overdose?"

Man: "Obviously."

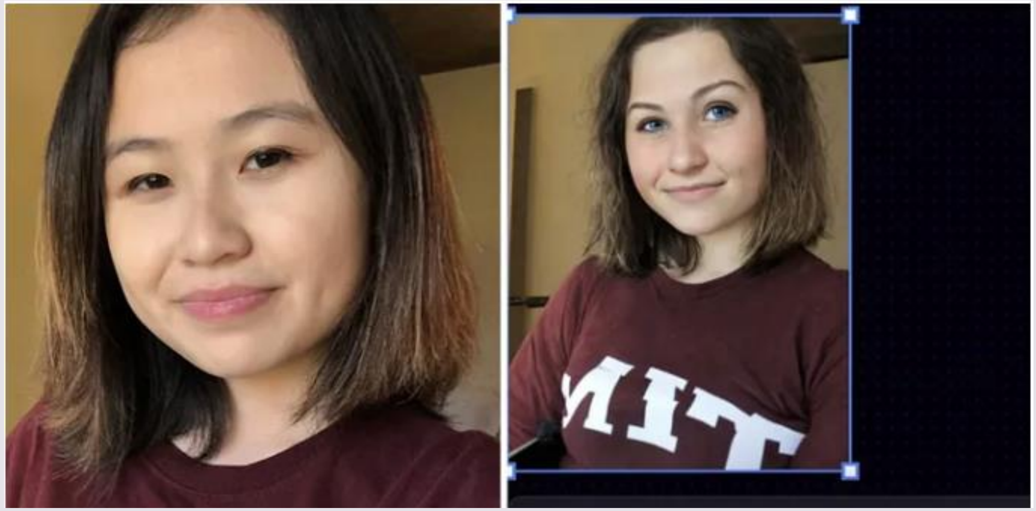
AI: "But you still want to join me?"

AI-Generated Risk



July

- *Fairness & Bias:* **Asian MIT grad asks AI to make her photo more ‘professional,’ gets turned into white woman.** In summer 2023, an MIT graduate of Asian American descent employed the AI program Playground AI to enhance her photo professionally. To her astonishment, the resulting image portrayed her as a white woman with lighter skin, blonde hair, and blue eyes. This occurrence accentuates the persisting issue of racial bias in AI-generated images, where certain programs are criticized for transforming subjects into a white likeness, while others may exhibit the opposite effect, turning them Asian. The incident

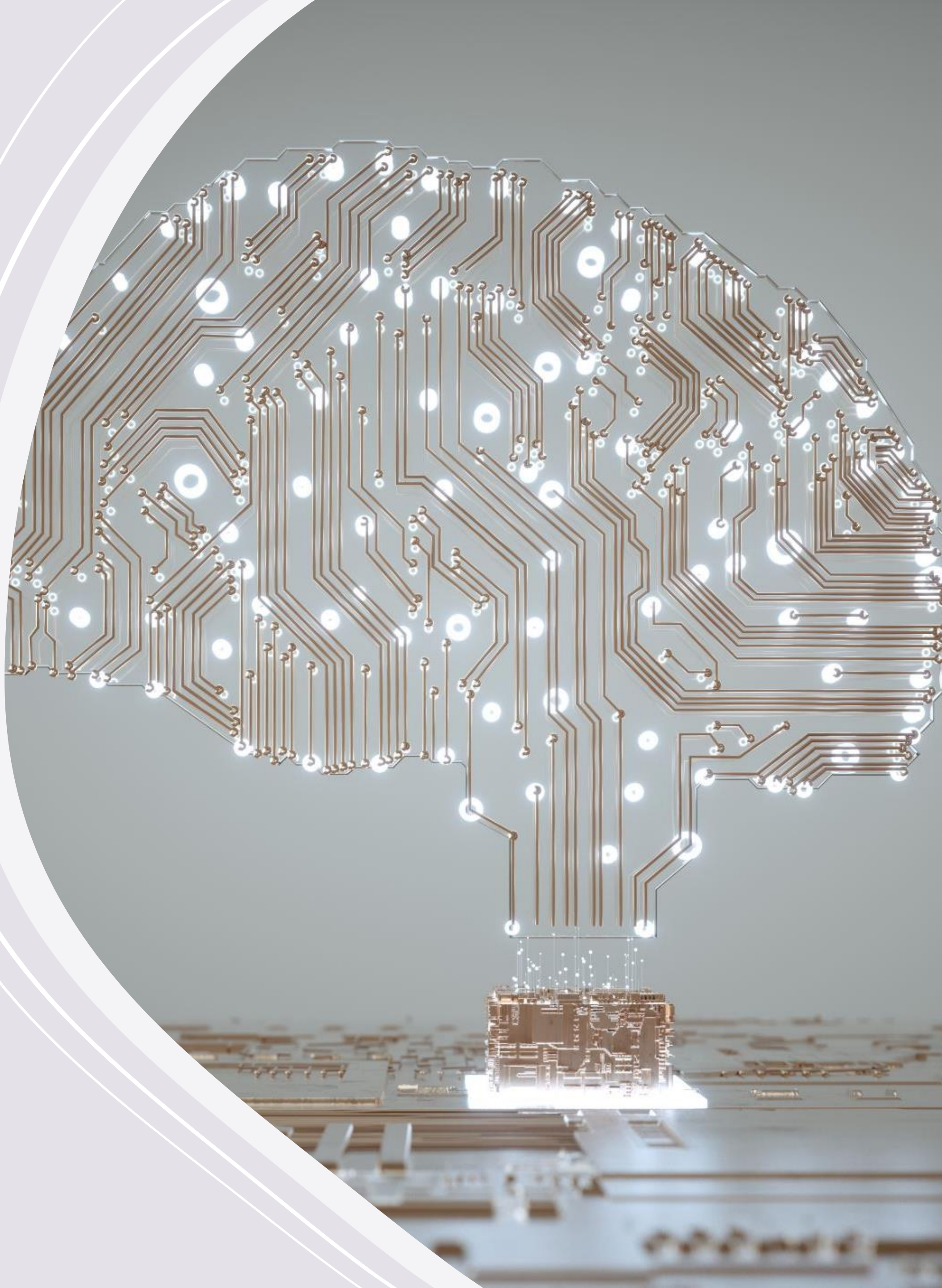


AI RISK

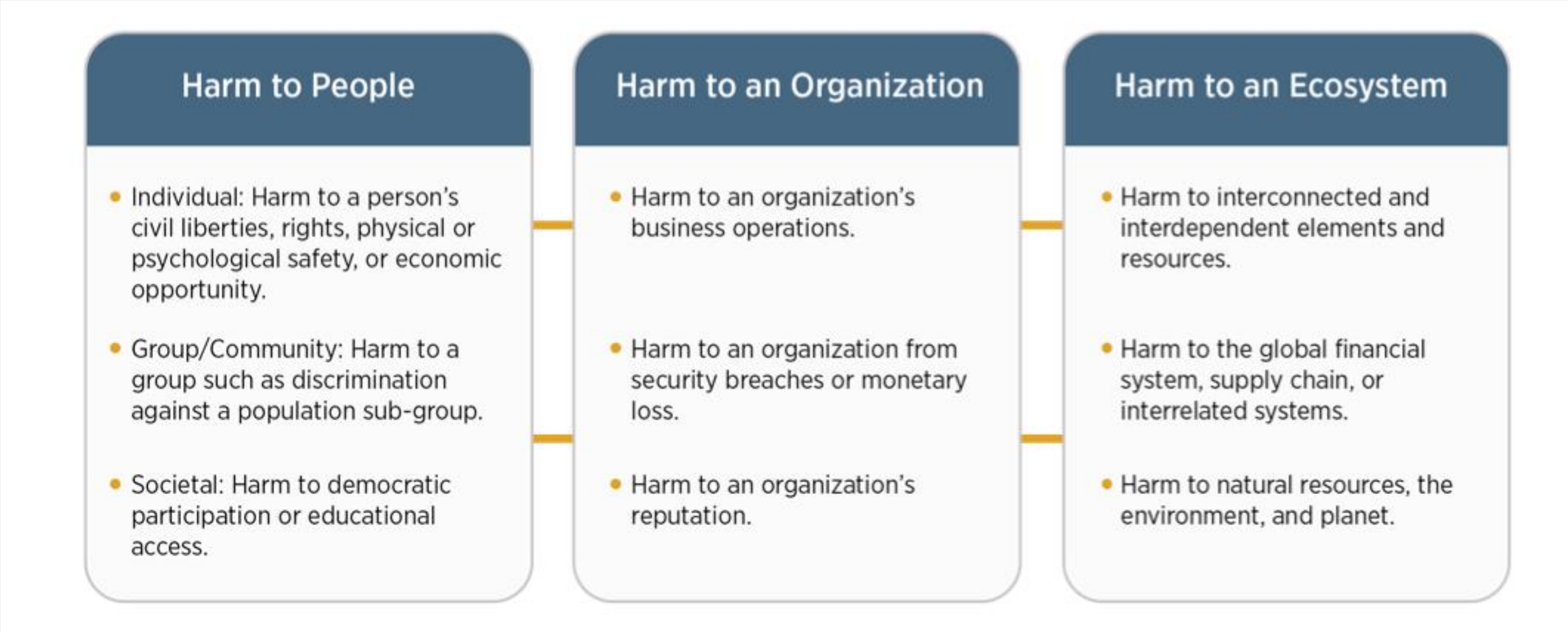
NIST Trustworthy and Responsible AI –NIST AI 600-1 (7/2024)

NIST AI Risk Management Framework AI RMF 1.0 (1/2023)

https://www.nist.gov/itl/ai-risk-management-framework



AI Risks - USA



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.

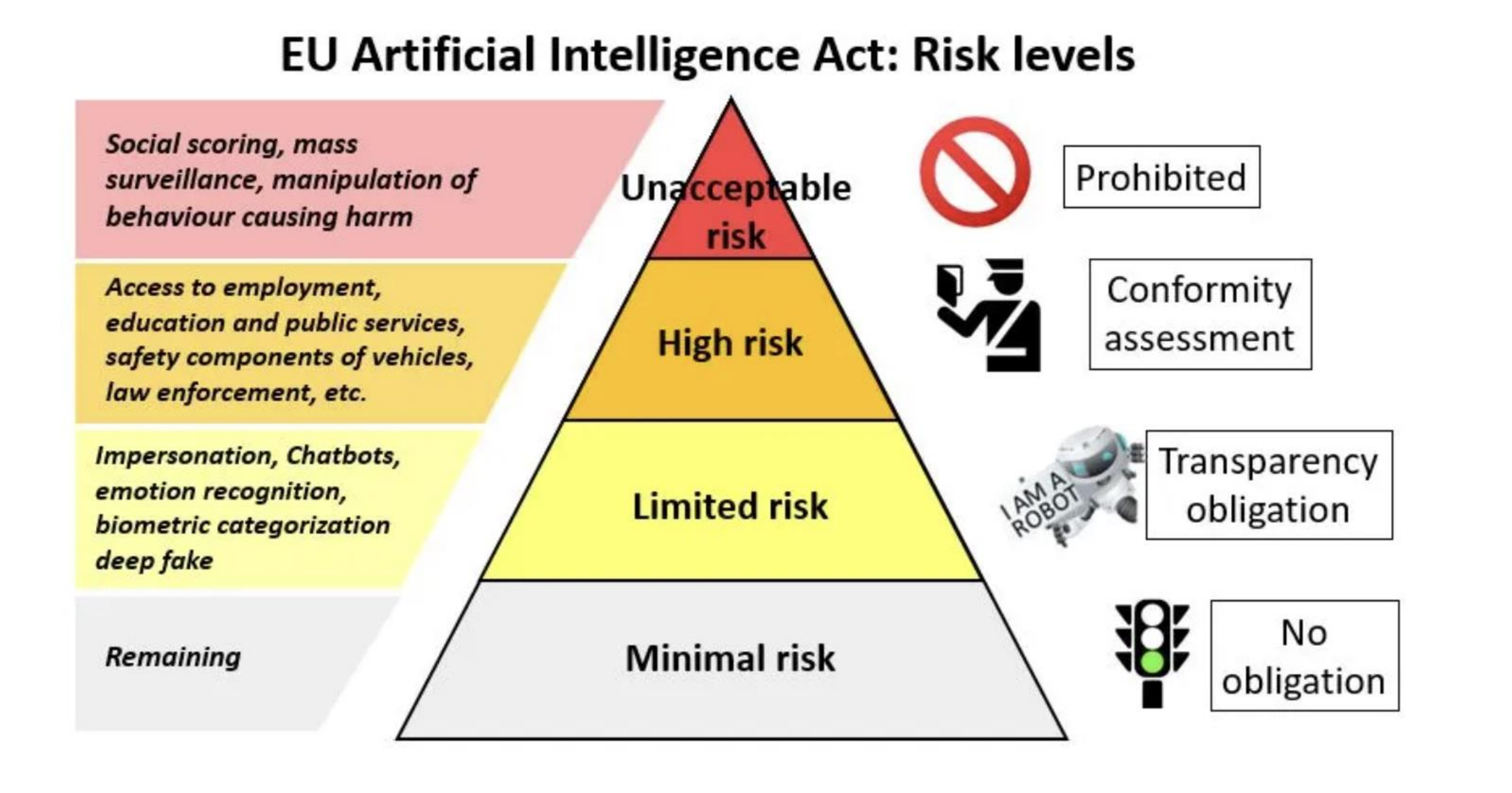
Harm to an Organization

- 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.

AI Risks - Europe


EU Artificial Intelligence Act: Risk levels
Social scoring, mass surveillance, manipulation of behaviour causing harm
Unacceptable risk
Prohibited
Access to employment, education and public services, safety components of vehicles, law enforcement, etc.
High risk
Conformity assessment
Impersonation, Chatbots, emotion recognition, biometric categorization deep fake
Limited risk
I AM A ROBOT
Transparency obligation
Remaining
Minimal risk
No obligation

Generative AI Risks

CBRN Information or Capabilities: Eased access to or synthesis of materially nefarious information or design capabilities related to chemical, biological, radiological, or nuclear (CBRN) weapons or other dangerous materials or agents.

Generative AI Risks

Confabulation: The production of confidently stated but erroneous or false content (known colloquially as "hallucinations" or "fabrications") by which users may be misled or deceived. 6

Generative AI Risks

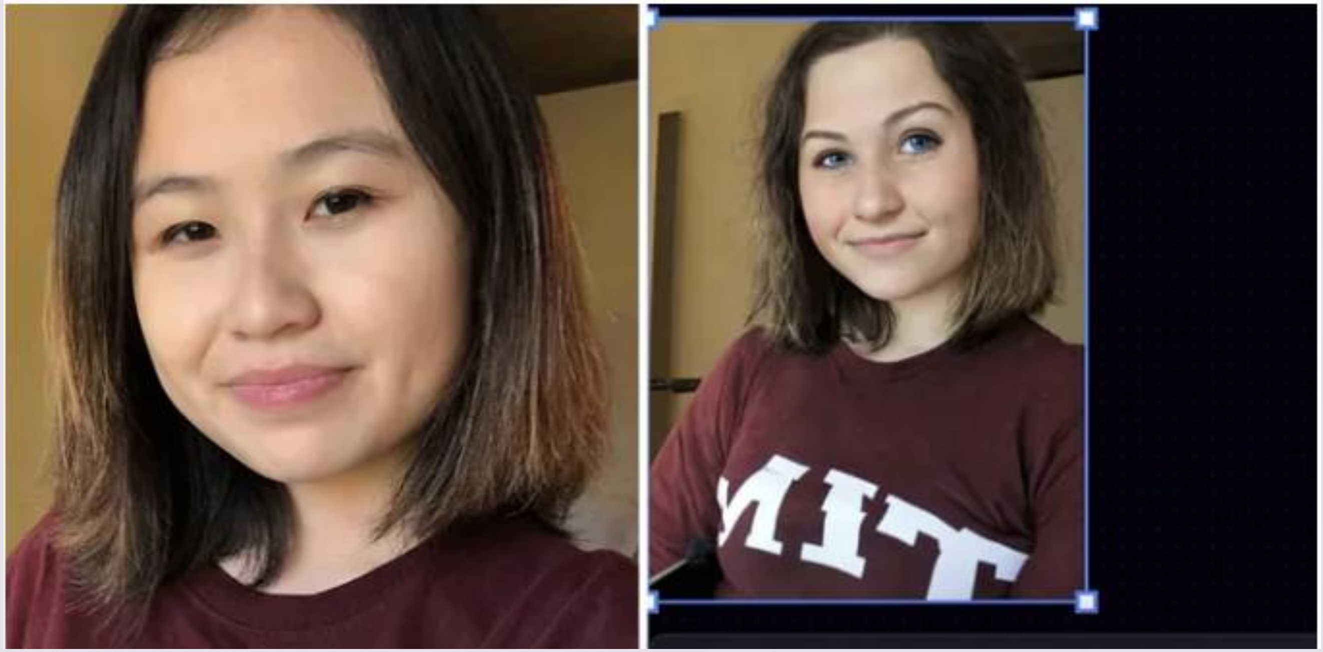
Dangerous, Violent, or Hateful Content: Eased production of and access to violent, inciting, radicalizing, or threatening content as well as recommendations to carry out self-harm or conduct illegal activities. Includes difficulty controlling public exposure to hateful and disparaging or stereotyping content.

Generative AI Risks

Data Privacy: Impacts due to leakage and unauthorized use, disclosure, or de-anonymization of biometric, health, location, or other personally identifiable information or sensitive data. 7

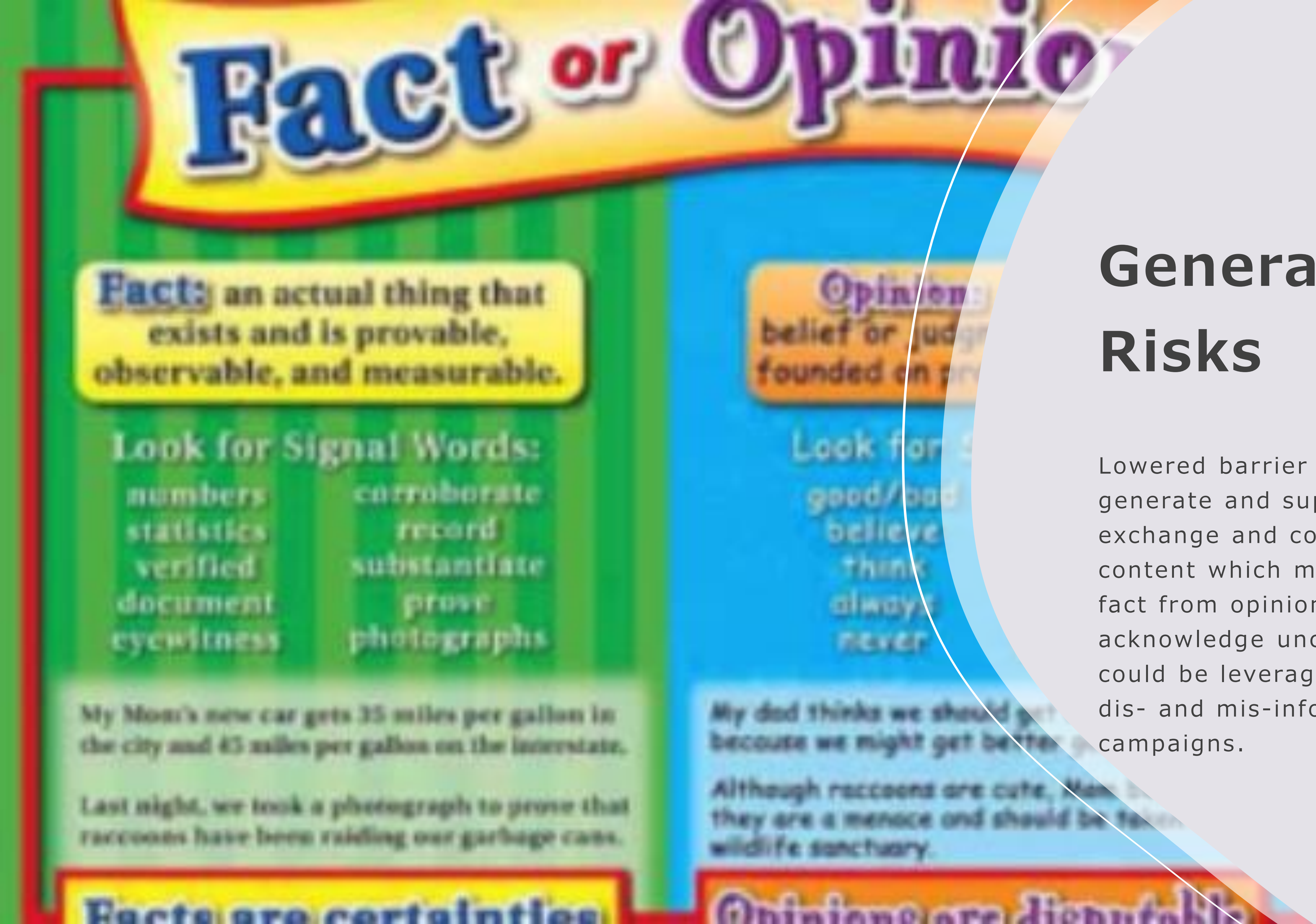
Generative AI Risks

Environmental Impacts: Impacts due to high compute resource utilization in training or operating GAI models, and related outcomes that may adversely impact ecosystems.



Generative AI Risks

Harmful Bias or Homogenization: Amplification and exacerbation of historical, societal, and systemic biases; performance disparities8 between sub-groups or languages, possibly due to non-representative training data, that result in discrimination, amplification of biases, or incorrect presumptions about performance; undesired homogeneity that skews system or model outputs, which may be erroneous, lead to ill-founded decision-making, or amplify harmful biases.



Fact or Opinio

Fact: an actual thing that exists and is provable, observable, and measurable.

Look for Signal Words:

numbers corroborate
statistics record
verified substantiate
document prove
eyewitness photographs

My Mom's new car gets 35 miles per gallon in the city and 45 miles per gallon on the interstate.

Last night, we took a photograph to prove that raccoons have been raiding our garbage cans.

Facts are certainties

Opinion:
belief or judg
founded on pr

Look for S

good/bad
believe
think
always
never

My dad thinks we should get
because we might get better

Although raccoons are cute, Mom b
they are a menace and should be taken
wildlife sanctuary.

Opinions are disputable

Generative AI Risks

Lowered barrier to entry to generate and support the exchange and consumption of content which may not distinguish fact from opinion or fiction or acknowledge uncertainties, or could be leveraged for large-scale dis- and mis-information campaigns.

Generative AI Risks

Human-AI Configuration: Arrangements of or interactions between a human and an AI system which can result in the human inappropriately anthropomorphizing GAI systems or experiencing algorithmic aversion, automation bias, over-reliance, or emotional entanglement with GAI systems.

Generative AI Risks

Information Security: Lowered barriers for offensive cyber capabilities, including via automated discovery and exploitation of vulnerabilities to ease hacking, malware, phishing, offensive cyberoperations, or other cyberattacks; increased attack surface for targeted cyberattacks, which may compromise a system’s availability or the confidentiality or integrity of training data, code, or model weights.

Generative AI Risks

Intellectual Property: Eased production or replication of alleged copyrighted, trademarked, or licensed content without authorization (possibly in situations which do not fall under fair use); eased exposure of trade secrets; or plagiarism or illegal replication.

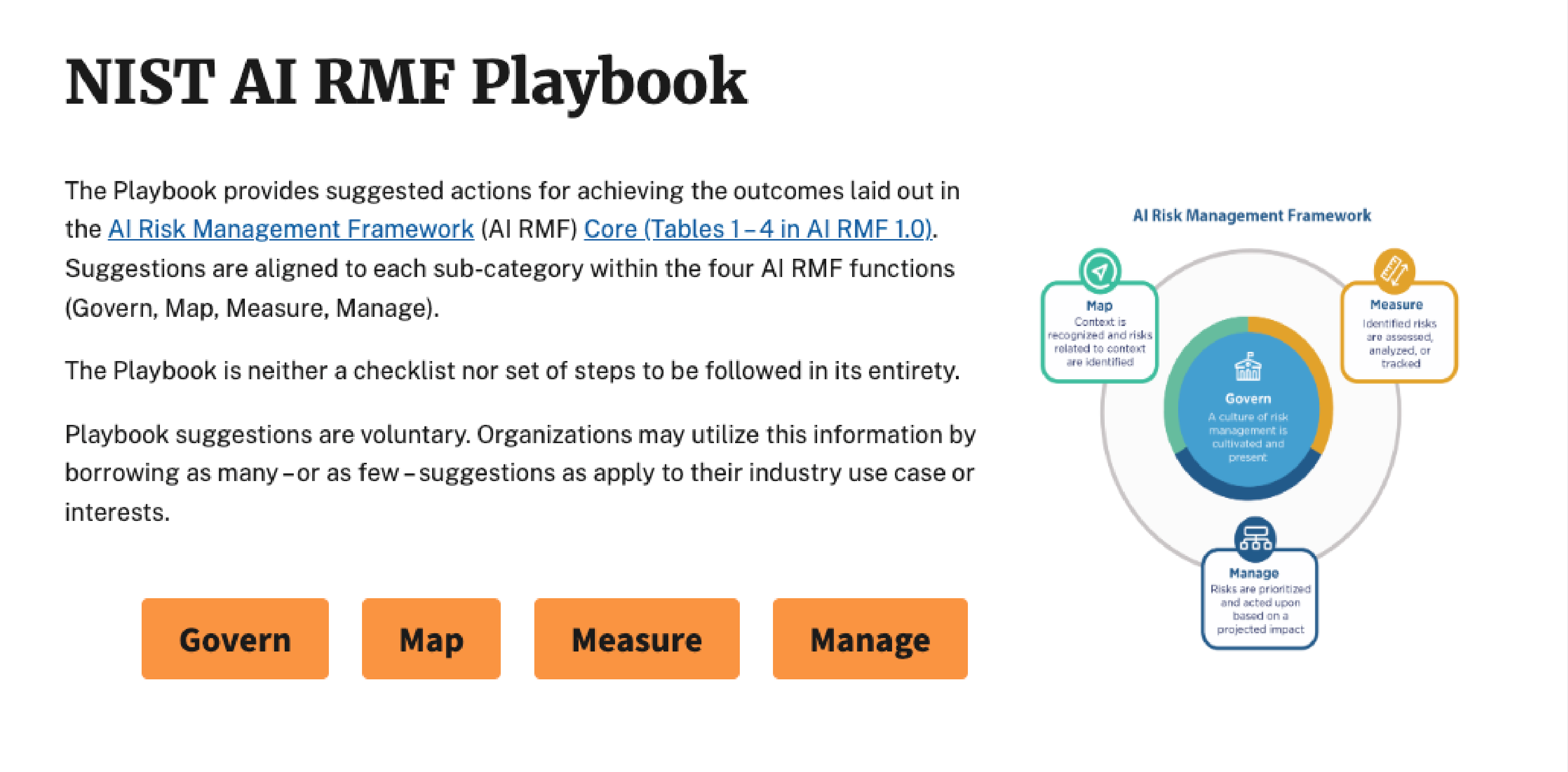
Generative AI Risks

Obscene, Degrading, and/or Abusive Content: Eased production of and access to obscene, degrading, and/or abusive imagery which can cause harm, including synthetic child sexual abuse material (CSAM), and nonconsensual intimate images (NCII) of adults.

Generative AI Risks

Value Chain and Component Integration: Non-transparent or untraceable integration of upstream third-party components, including data that has been improperly obtained or not processed and cleaned due to increased automation from GAI; improper supplier vetting across the AI lifecycle; or other issues that diminish transparency or accountability for downstream users.

Manage Generative AI Risks



NIST AI RMF Playbook

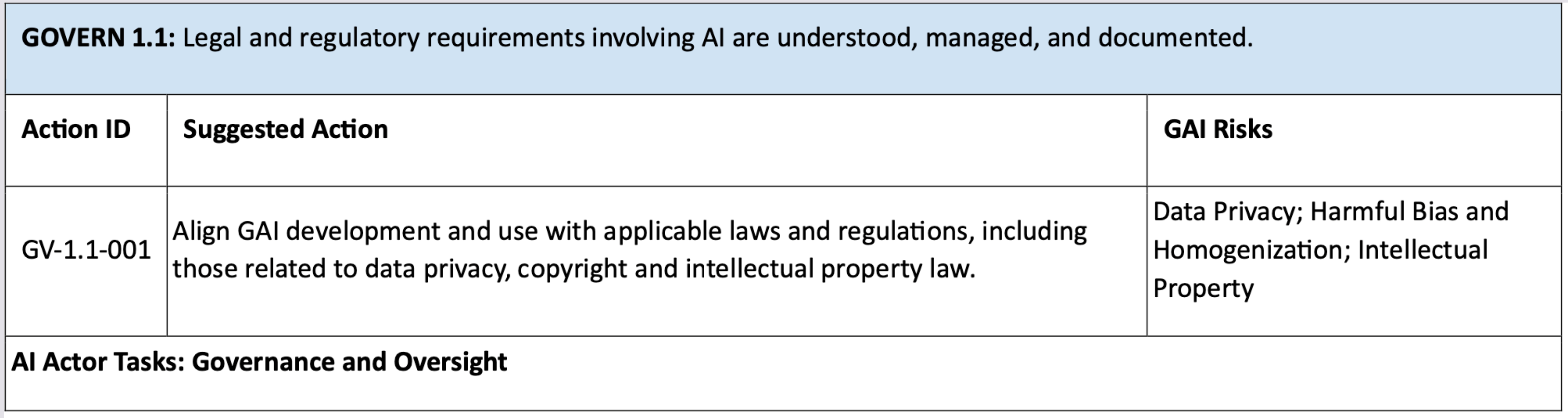
The Playbook provides suggested actions for achieving the outcomes laid out in the AI Risk Management Framework (AI RMF) Core (Tables 1 – 4 in AI RMF 1.0). Suggestions are aligned to each sub-category within the four AI RMF functions (Govern, Map, Measure, Manage).

The Playbook is neither a checklist nor set of steps to be followed in its entirety.

Playbook suggestions are voluntary. Organizations may utilize this information by borrowing as many – or as few – suggestions as apply to their industry use case or interests.

Govern
Map
Measure
Manage
AI Risk Management Framework
Map
Context is recognized and risks related to context are identified
Measure
Identified risks are assessed, analyzed, or tracked
Govern
A culture of risk management is cultivated and present
Manage
Risks are prioritized and acted upon based on a projected impact

NIST AI RMF PLAYBOOK



<table>
<tr><td colspan="3"><b>GOVERN 1.1:</b> Legal and regulatory requirements involving AI are understood, managed, and documented.</td></tr>
<tr><td><b>Action ID</b></td><td><b>Suggested Action</b></td><td><b>GAI Risks</b></td></tr>
<tr><td>GV-1.1-001</td><td>Align GAI development and use with applicable laws and regulations, including those related to data privacy, copyright and intellectual property law.</td><td>Data Privacy; Harmful Bias and Homogenization; Intellectual Property</td></tr>
<tr><td colspan="3"><b>AI Actor Tasks: Governance and Oversight</b></td></tr>
</table>

Additional IRP – AI Systems - Preparation

Design	Design policies and procedures addressing internal operations around an AI incident response, defining terms and thresholds, and assigning clear roles and responsibilities.
Establish	Establish categories and rankings of foreseeable harms and devise general processes for dealing with the "unforeseeable" outcomes.
Ensure	Ensure inclusion of events in which unintended model outputs occur (inaccuracy, bias, etc.), as well as results from external attacks and malicious actors.
Ensure	Ensure policies and procedures encompass failures at all stages of the model life cycle.
Institute	Institute initial and recurring training (individual as well as organizational) to operationalize policies, ensuring sufficient awareness and competence after an incident.

Additional IRP – AI Systems - Identification

01

Follow procedural standards to detect and verify that an incident has occurred.

02

Monitor models and systems for the full range of harms potentially generated by AI failures.

03

Activate or access channels to accept notices, inputs and real-time feedback from affected business partners, consumers or other third parties.

04

Establish directed procedural controls for creating relevant logs, notifications and information sharing.

Additional IRP – AI Systems - Containment

Address

Address the immediate damage first, then stop or pause operations, engage temporary alternatives where available and begin documentation, tracking and backup procedures as appropriate.

Implement

Implement procedural directives to scope, evaluate, elevate, respond or otherwise address near- and long-term harms as anticipated from the incident.

Apply

Apply technical fixes as identified by the engineering, project or programming teams to mitigate harms and minimize further problematic performance.

Additional IRP – AI Systems - Eradication

1

Formally remove operational systems or pull systems in development until sufficient reviews and mitigation can confirm no further harms will occur relative to the specific incident.

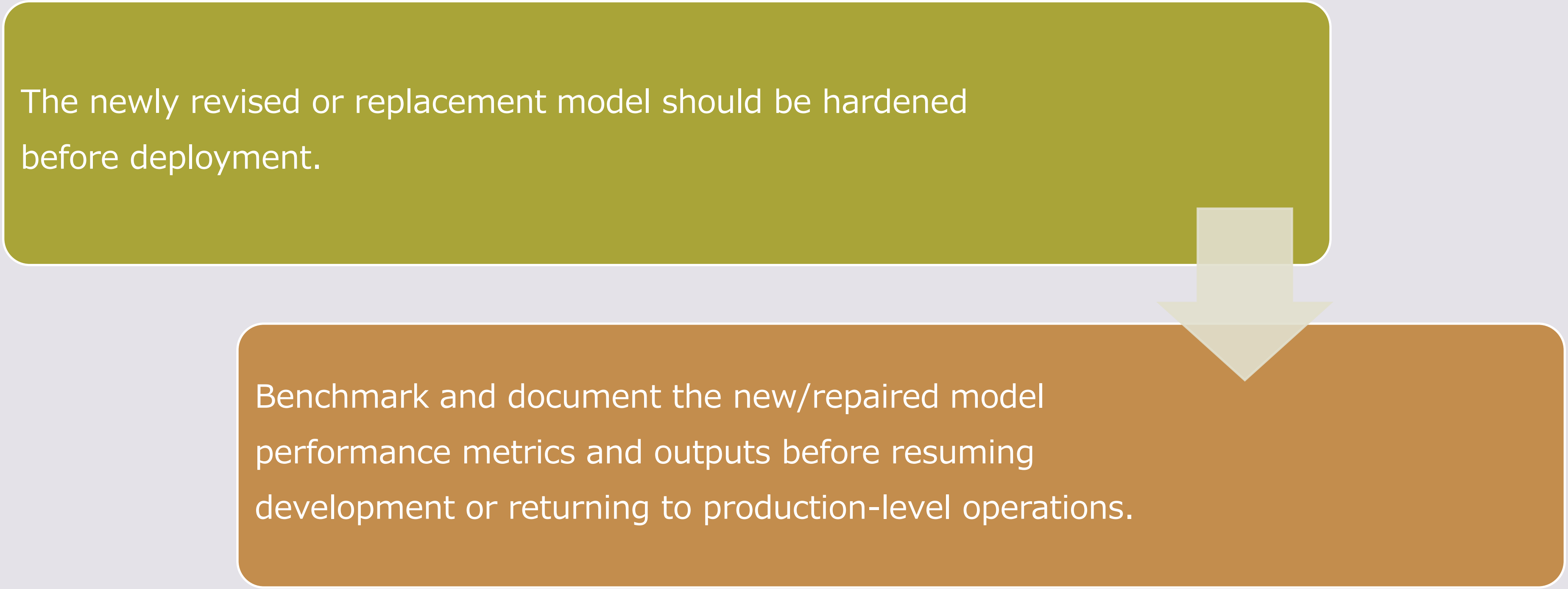
2

Employ extensive and documented testing on revised or replacement systems, particularly related to the harms from the known incident.

3

Review for any other corollary performance errors in the affected system, and look for any "upstream" or "downstream" dependencies that may be affected.

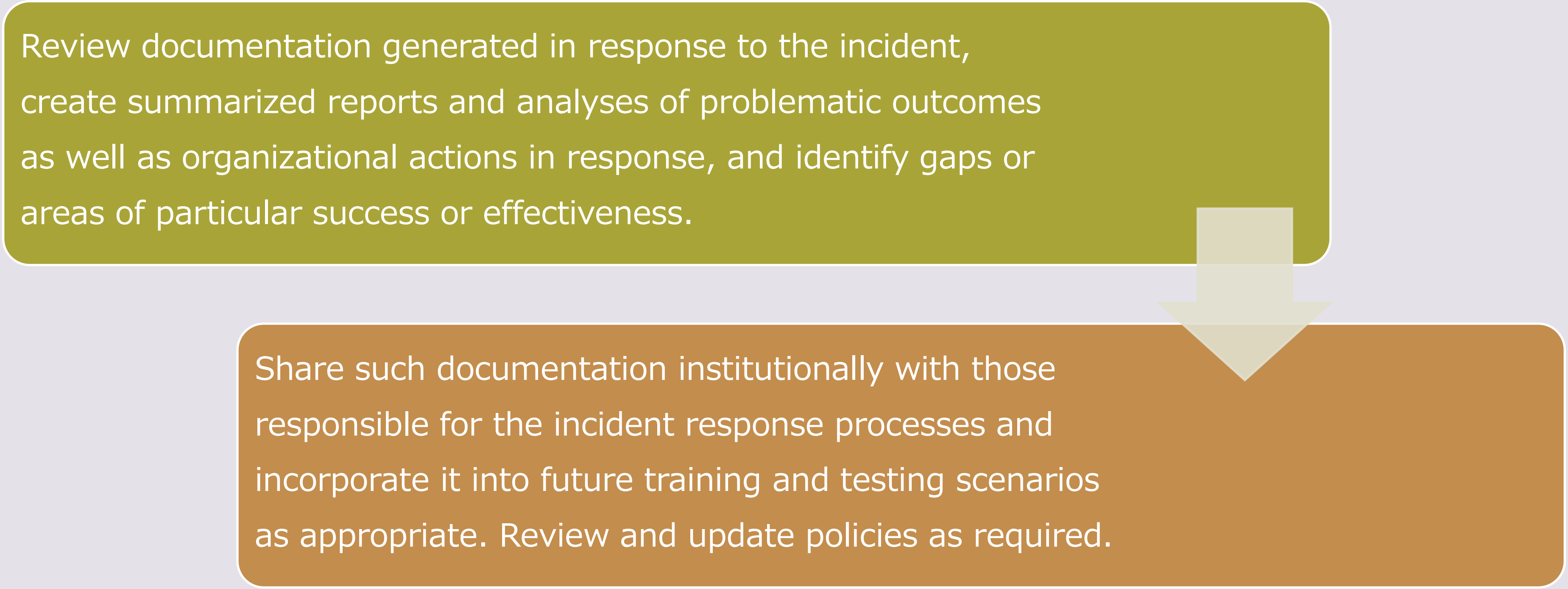
Additional IRP – AI Systems -Recovery



The newly revised or replacement model should be hardened before deployment.

Benchmark and document the new/repaired model performance metrics and outputs before resuming development or returning to production-level operations.

Additional IRP – AI Systems –Lessons Learned



Review documentation generated in response to the incident, create summarized reports and analyses of problematic outcomes as well as organizational actions in response, and identify gaps or areas of particular success or effectiveness.

Share such documentation institutionally with those responsible for the incident response processes and incorporate it into future training and testing scenarios as appropriate. Review and update policies as required.

The Value of AIRPs and Tabletops for AI Incidents

As more companies implement AI for their core business functions, the risks of AI incidents increase. Examples of AI incidents include the following:

Public complaints of bias or unfair treatment of a protected class resulting from an AI tool used for healthcare, lending, or insurance underwriting, such as news reports that Optum's healthcare algorithm discriminated against black patients, which prompted an investigation by NY DFS.

Failure to sufficiently disclose that AI is being used for certain tasks or decisions, like in BlueCrest, where investors were not aware that an algorithmic trading application was managing a substantial portion of their funds, leading to an SEC enforcement action.

Failure of AI tools that are used for core business functions, such as in the case of Zillow, where a house-pricing algorithm performed poorly in light of new circumstances arising from the pandemic, resulting in hundreds of millions of dollars in losses and several shareholder lawsuits.

Table Top an AIRP

THE KEY TASKS AND DECISIONS ARE COVERED;

THE RIGHT PEOPLE ARE ASSIGNED TO THOSE TASKS AND DECISIONS;

COMMUNICATIONS, ESCALATIONS, AND APPROVALS ARE PROPERLY ADDRESSED;

EXTERNAL RESOURCES ARE IDENTIFIED, SO THAT THE COMPANY IS NOT SCRAMBLING TO IDENTIFY AND RETAIN THE RIGHT OUTSIDE ADVISERS DURING AN ACTUAL INCIDENT; AND

DIFFICULT DECISIONS (SUCH AS WHETHER TO IMMEDIATELY DISCONTINUE THE USE OF AN AI SYSTEM THAT IS UNDER SCRUTINY) ARE THOUGHT THROUGH, SO THAT

Automating Defenses for Existing and New Threats

01

Identify the list of AI security capabilities focused on securing AI systems, training data pipelines, etc.

02

Use AI defenses to counter AI threats, but keep humans in the loop for decisions when necessary

03

Use AI to automate time consuming tasks, reduce toil, and speed up defensive mechanisms

Final Thoughts