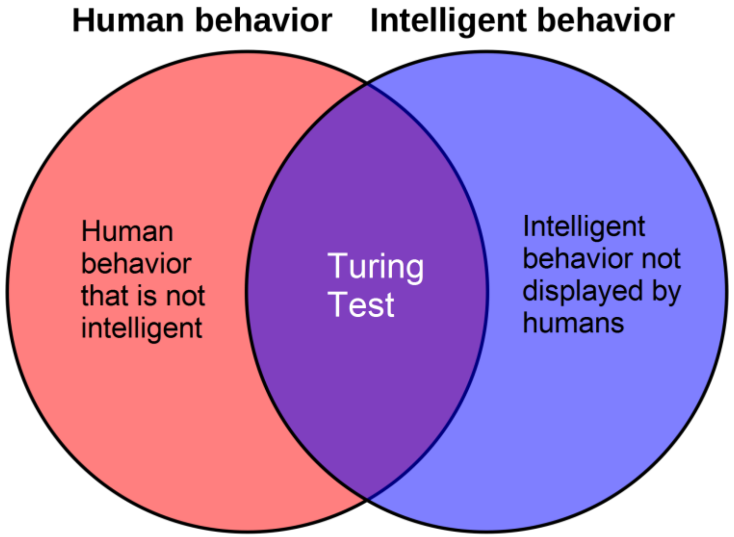
# Artificial Intelligence Taxonomy Overview

# Artificial Intelligence: Overview

**The term artificial intelligence (AI) is challenging to define due to constantly evolving technology.**

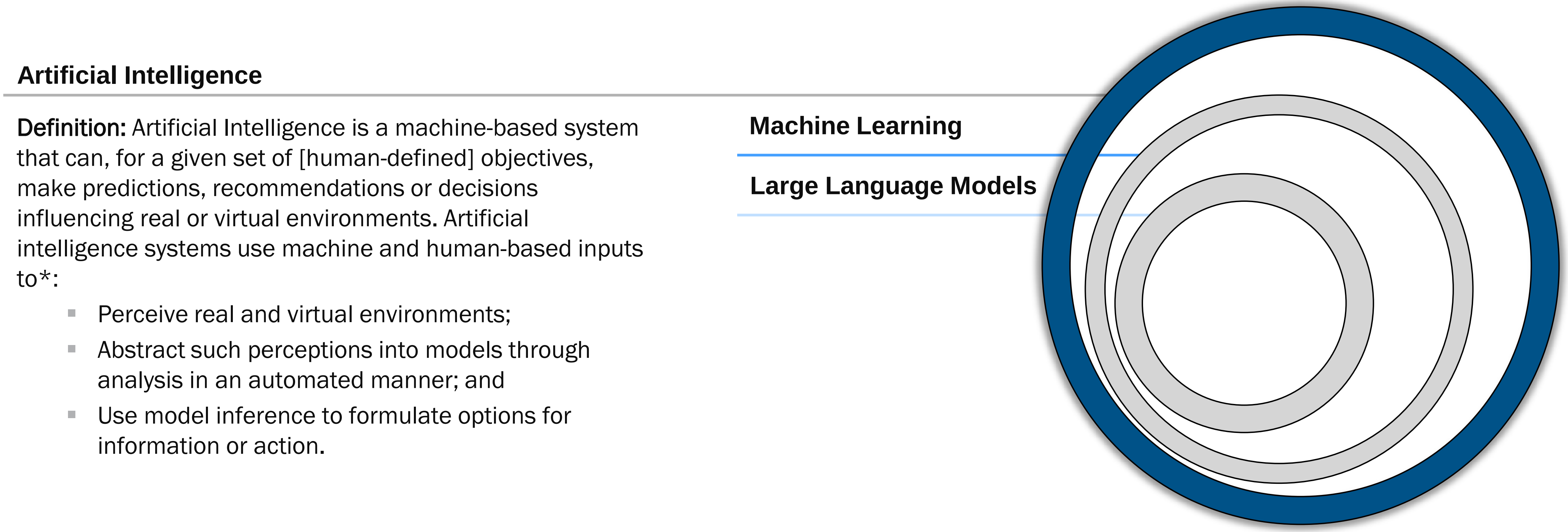
- Conceptualized in science fiction literature
- Can we use machines to automate tasks that previously required human labor?
- Can we create a human-like intelligence?
  - Chatbots can already pass the Turing test


Human behavior
Intelligent behavior
Human behavior that is not intelligent
Turing Test
Intelligent behavior not displayed by humans

**AI is simultaneously a technology of TOMORROW and TODAY.**

# Artificial Intelligence: Overview

## The term artificial intelligence (AI) is challenging to define due to constantly evolving technology.



**Definition:** Artificial Intelligence is a machine-based system that can, for a given set of [human-defined] objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to*:

- Perceive real and virtual environments;
- Abstract such perceptions into models through analysis in an automated manner; and
- Use model inference to formulate options for information or action.

*Subsets of AI*

*National AI Initiative (WH/OSTP)

TLP:GREEN

# Environment: AI Policy

**Recent policy developments highlight AI as an Administration and congressional priority; CISA will fulfill important coordinating role.**

**National AI Initiative (NAII) Act of 2020**: Coordinated complementary AI R&D, demonstration activities among FCEB, DOD, IC.

**AI in Government Act of 2020**: Established the AI Center of Excellence within GSA.

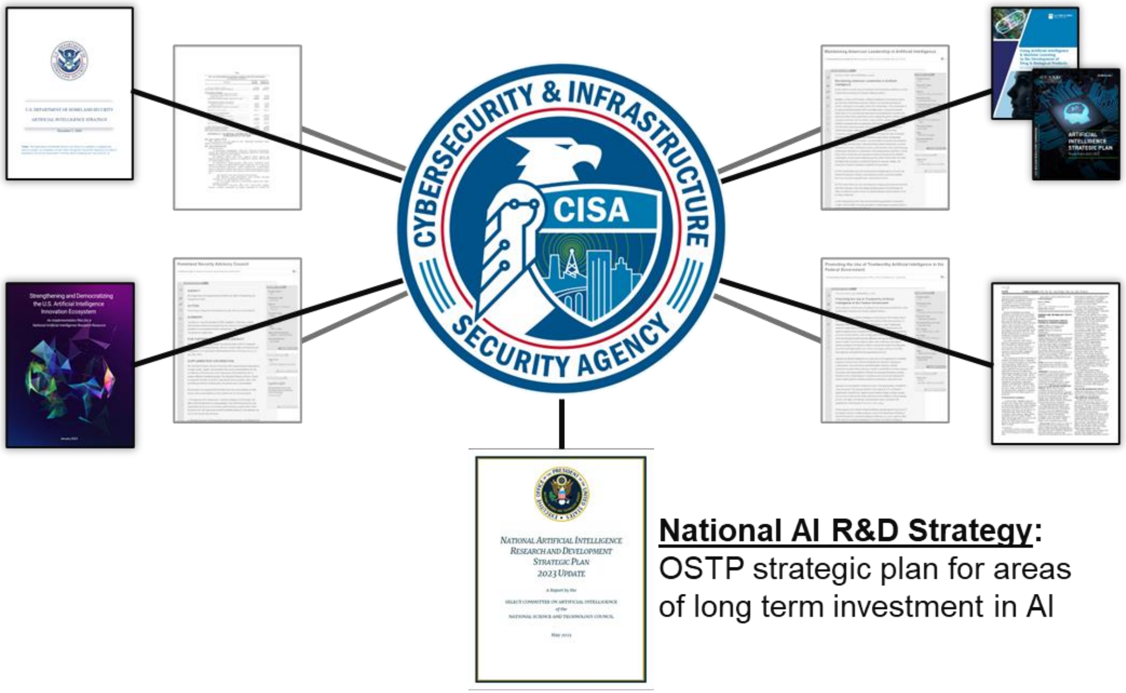
**EO 13859: Maintaining American Leadership in AI**: Established federal principles and strategies to strengthen the nation's capabilities in AI.

**EO 13960: Promoting the Use of Trustworthy AI in the Federal Gov't**: Required Agencies to inventory and share AI use cases. Sets out 9 "Principles for Use of AI in Government"

TLP:GREEN

# Environment: AI Strategy Development

**Articulation of DHS/CISA and broader FCEB strategic needs is in full swing, as a way of operationalizing on the need for AI.**


CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY
CISA
NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN 2023 UPDATE
Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem
ARTIFICIAL INTELLIGENCE STRATEGIC PLAN

**DHS AI Strategy**: Strategic vison for DHS role in policy development, governance, use of AI, and risk mitigation.

**Implementation Plan for a National AI Research Resource**: Memorializes findings of National Artificial Intelligence Research Resource (NAIRR) Task Force on national AI research infrastructure.

**OGA Strategies**: FDA, Nuclear Regulatory Commission, and others have released AI strategies tailored to specific mission areas.

**National Priorities for AI RFI**: OSTP RFI on key themes to inform Administration's updated National AI Strategy.

**National AI R&D Strategy**: OSTP strategic plan for areas of long term investment in AI

# Machine Learning: Overview

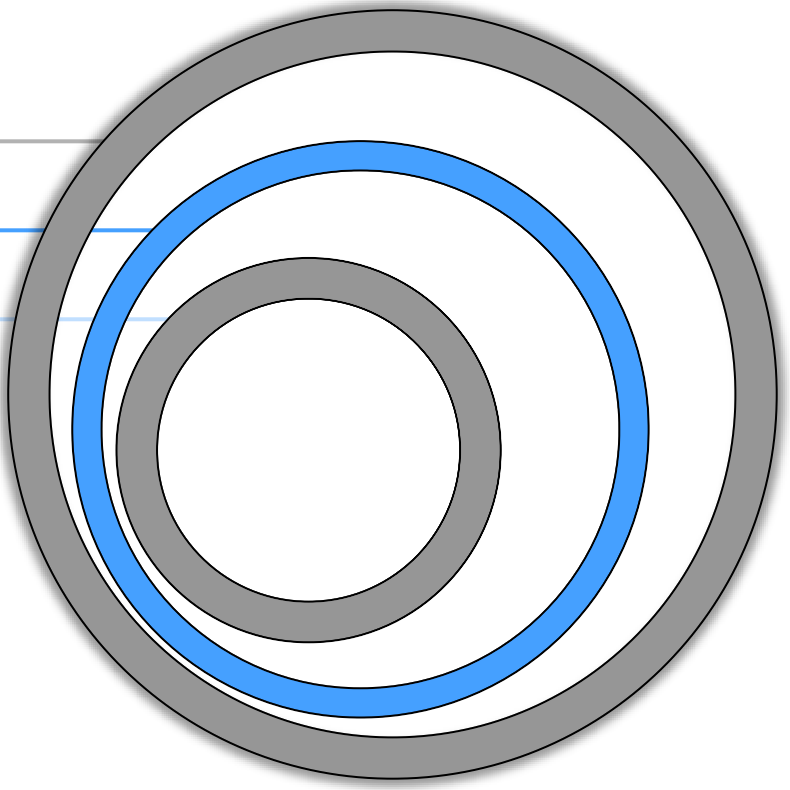
**When individuals today are discussing AI, they are often discussing machine learning (ML).**

## Machine Learning

A training a computer model to understand a representation of data, rather than explicitly incorporating instructions into programming.

Machine Learning is:

- A subset of AI
- Typically does not refer to traditional statistics models, though it may leverage them.



*Subsets of AI*

TLP:GREEN

# AI/ML Terminology

This AI/ML community uses a lot of jargon that may have different meanings in other contexts.

| Term | Description |
|---|---|
| **Training** | Building your model by applying algorithms to have it learn a representation of data. |
| **Inference** | Running your model by having it take inputs and generate outputs. |
| **Unsupervised Learning** | Training a model without providing any labels for your data inputs. |
| **Supervised Learning** | Training a model with labeled data inputs. |
| **Reinforcement Learning** | Training intelligent agents to take actions based on a reward function. |

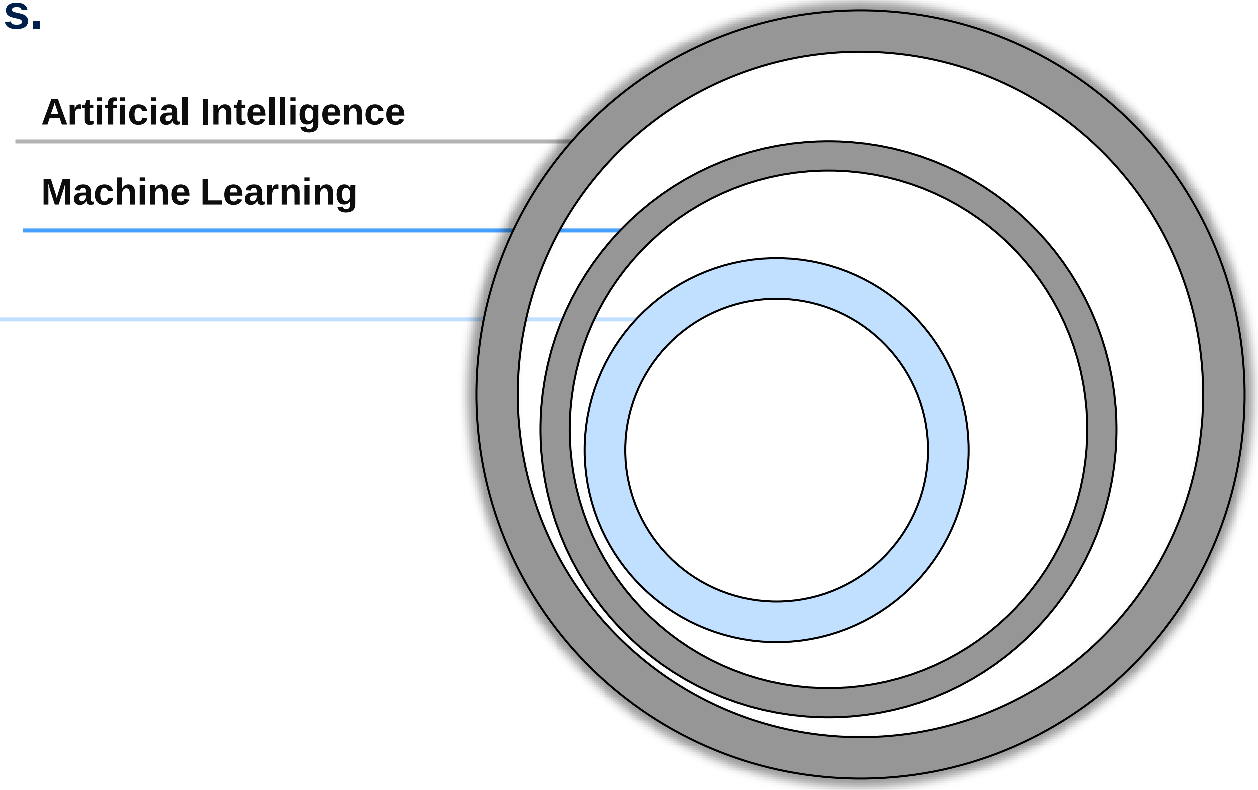
# Types of AI/ML

Most modern AI systems use a combination of AI/ML approaches. These are just a few common categories of ML.

| Term | Description |
|---|---|
| **Regression** | Statistical process for predicting the value of dependent variables based on one or more independent variables or data values. |
| **Classification** | Predicting a class label for based on one or more independent variables or data values. |
| **Deep Learning** | This is a model architecture that has multiple layers, most commonly neural networks with at least one intermediate layer. |
| **Ensemble Learning** | Using a set of algorithms and/or training a set of models to provide better performance than each of them would provide individually. |

# Large Language Models: Overview

**Large language models (LLMs) are the key to human-AI interaction as their text-based prompts provide human interoperability to other models.**



Artificial Intelligence

Machine Learning

### Large Language Models

**LLMs:** A class of language models that use deep learning algorithms and are trained on extremely large textual datasets that can be multiple terabytes in size*.

- Successful LLMs require **fine-tuning.** ChatGPT was fine tuned using a large number of human interactions.
- The **growth curve for LLMs has been exponential** and far exceeds previous technology adoption curves. Change management across USG will be a challenge due to this growth curve.

*Subsets of AI*

*https://doi.org/10.6028/NIST.AI.100-3

# AI Security: Overview

**AI Security is an umbrella term used for several different categories of cybersecurity.**

Three Key Categories of AI Security

1. **Applications of AI for Cybersecurity:** CISA is actively leveraging AI and ML tools for threat detection, prevention, and vulnerability assessments.
2. **Cybersecurity of AI-Enabled Systems:** CISA currently has limited ability to protect and secure AI-enabled systems.
3. **Threats from Malicious Use of AI:** CISA needs to research (via S&T), develop and/or acquire tools to actively to protect from adversarial threats across the FCEB.

# AI Security: Risks and Threats

There are a variety of threats that are actively being identified in the wild – these recent examples signify the relevance of the current discussion.

| Term | Description | Examples |
|---|---|---|
| **Confidentiality** | Risks associated with data privacy and security, including the potential for sensitive information to be inadvertently shared or used inappropriately. | *OpenAI: ChatGPT payment data leak caused by open-source bug (bleepingcomputer.com)* |
| **Supply Chain** | Risks associated with reliance on third-party providers for AI systems and dependencies. | *Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022. \| PyTorch* |
| **Malicious Use of AI** | Risks associated with threat actors leveraging AI to enhance the sophistication of their operations. | *ChatGPT Powered Malware Bypasses EDR \| by David Merian \| Mar, 2023 \| System Weakness*<br>*Disinformation Researchers Raise Alarms About A.I. Chatbots - The New York Times (nytimes.com)* |
| **Adversarial Machine Learning (AML)** | Providing deceptive inputs to a machine learned model to cause it to behave in an unexpected fashion | Prompt Injection Attack on GPT-4 |

TLP:GREEN

# AI Security: Types of Adversarial ML

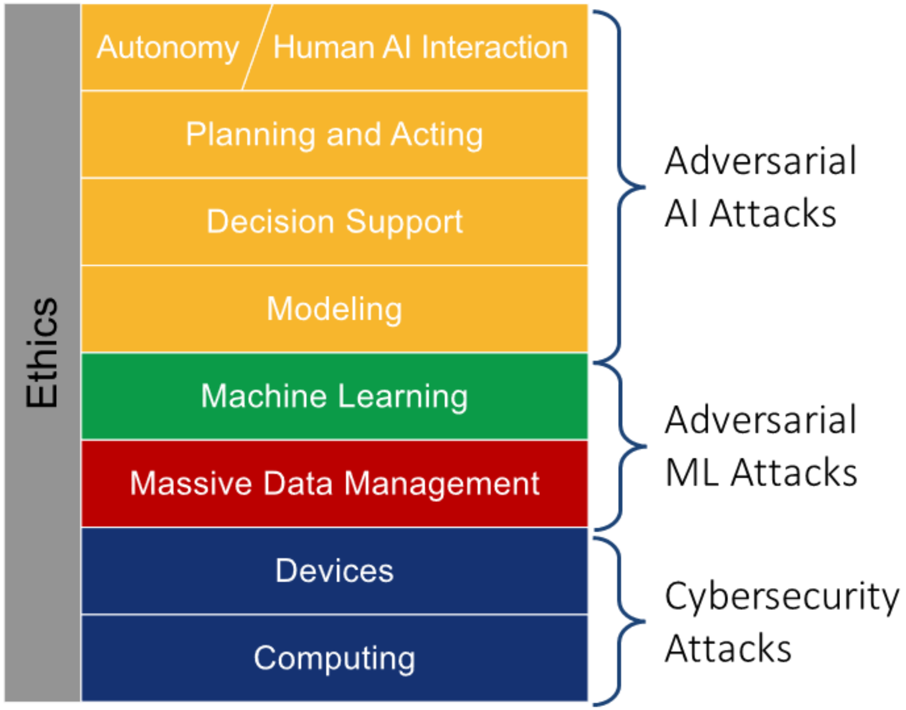
There are a variety of threats that are actively being identified in the wild.

| Name | Description | Example |
|---|---|---|
| **Poisoning** | Modifying the ML model through deceptive training inputs. | VirusTotal Poisoning, Case Study: AML.CS0002 \| MITRE ATLAS™ |
| **Evasion** | Making illegitimate inputs appear legitimate. | Malware Reaches Play Store as Google Wages War Against BankBot Trojan (bleepingcomputer.com) |
| **White Box** | Training inputs and/or model parameters are known. | On End-to-End White-Box Adversarial Attacks in Music Information Retrieval - Transactions of the International Society for Music Information Retrieval (ismir.net) |
| **Black Box** | Model is hidden, but inputs and outputs are visible. | ChatGPT tied to Samsung's alleged data leak |

# Artificial Intelligence Risk Management Framework

# AI Stack

AI Stack



Ethics

Autonomy / Human AI Interaction

Planning and Acting

Decision Support

Modeling

Machine Learning

Massive Data Management

Devices

Computing

Adversarial AI Attacks

Adversarial ML Attacks

Cybersecurity Attacks

AI systems will be supported by traditional IT systems.

AI specific risks should be managed through the processes outlined in the NIST AI RMF

Much of this is an unsolved problem: Risks are hard to quantify, vulnerability data is sparse, AI incident reporting is not formalized, and AI Incident response is immature.

Moore, A., Hebert, M., Shaneman, S., 2018. “The AI Stack: A blueprint for developing and deploying Artificial Intelligence.” Proceedings Volume 10635, Ground/Air Multisensor Interoperability, Integration, and networking for Persistent ISR IX. SPIE Defense + Security Conference, Orlando, FL.

# AI Risks

## 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

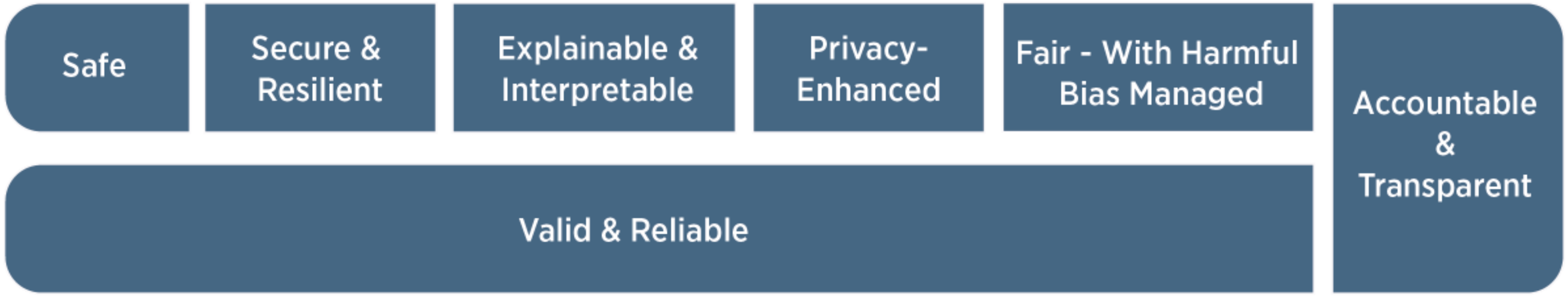
- Harm to interconnected and interdependent elements and resources.
- Harm to the global financial system, supply chain, or interrelated systems.
- Harm to natural resources, the environment, and planet.

# AI Risks are Socio-Technical

- **Autonomy and Decision-Making:** AI systems have the potential to operate autonomously and make decisions without direct human intervention. This introduces a new level of complexity and risk, as *AI systems may have unintended or undesirable behaviors.* IT risks, on the other hand, typically involve issues related to data breaches, system failures, or unauthorized access, but they do not possess autonomous decision-making capabilities
- **Unpredictability and Opacity:** AI algorithms, particularly those based on deep learning and neural networks, can be highly complex and difficult to interpret. This opacity makes it *challenging to predict and understand how an AI system might behave in different situations.* In contrast, IT risks are often more predictable and can be mitigated through well-established security measures and protocols.
- **Amplification of Errors:** AI systems have the potential to learn from vast amounts of data and can amplify errors or biases present in the training data. This can lead to unintended consequences, such as *discriminatory decision-making or reinforcement of existing societal biases.* IT risks, while they can also have significant consequences, are generally more focused on the security and integrity of data and systems rather than amplifying errors or biases.
- **Long-Term Impact and Unintended Consequences:** AI technologies have the potential for far-reaching and long-term societal impact. The *decisions and actions taken by AI systems can have significant consequences for individuals, organizations, and society as a whole.* IT risks, although important, are typically more localized and can be addressed through established risk management practices.
- **Ethical Considerations:** AI risks often involve ethical considerations that go beyond traditional IT risks. For example, issues such as *privacy, fairness, accountability, and transparency are critical when designing and deploying AI systems.* These ethical dimensions are not as prominent in conventional IT risk management, although they may overlap in certain areas.

# Trustworthy AI

- AI Risk Management will involve balancing the risks associated with these characteristics.
- These trade-offs will require understanding the decision-making context.
- Trade-offs and the associated implications necessitate transparency.
- AI Risk Management balances the socio-technical aspects of AI Risks.



Safe

Secure & Resilient

Explainable & Interpretable

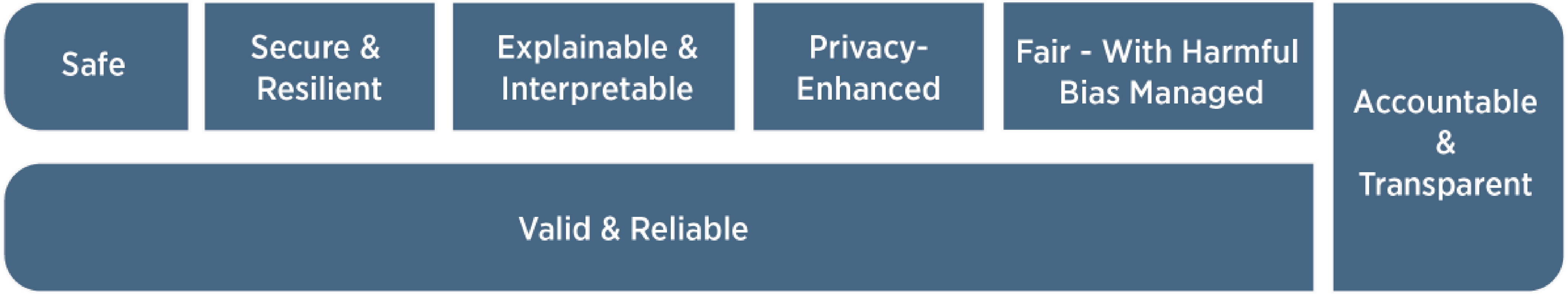
Privacy-Enhanced

Fair - With Harmful Bias Managed

Accountable & Transparent

Valid & Reliable

# Characteristics of Trustworthy AI



Safe

Secure & Resilient

Explainable & Interpretable

Privacy-Enhanced

Fair - With Harmful Bias Managed

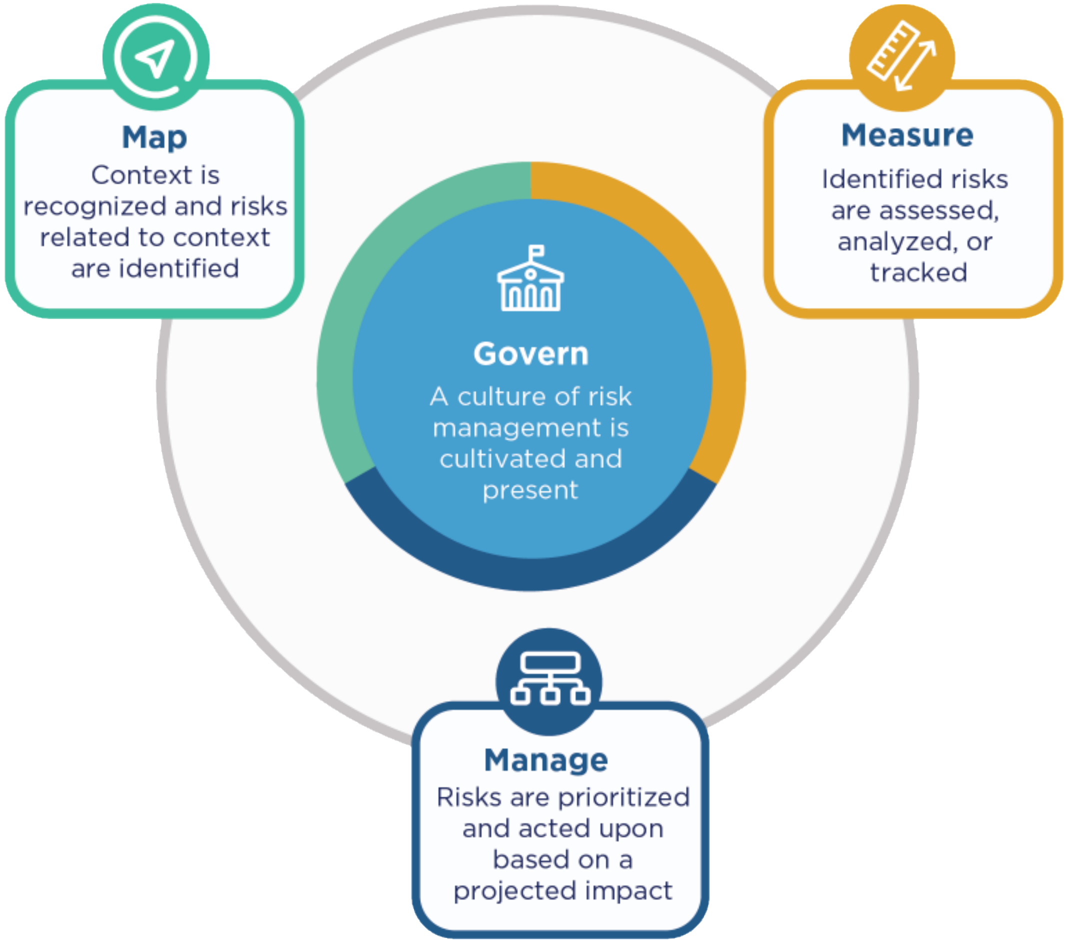
Accountable & Transparent

Valid & Reliable

- Valid and Reliable – Ongoing testing to ensure the system is performing as intended.
- Safe – Not lead to conditions where a human life, health, property, or the environment is endangered.
- Secure and Resilient – Related but distinct, security means that confidentiality, integrity, and availability is maintained, while resilience relates to being able to withstand unexpected adverse events or unexpected changes.
- Accountable and Transparent – Information is available about an AI system and it’s outputs to individuals interacting with the AI.
- Explainable and Interpretable – Explainability is a representation of the underlying AI system’s operation and interpretability is the meaning of the AI system output in the context of the designed functional purpose.
- Privacy–Enhanced – Safeguard human autonomy, identity, and dignity.
- Fair-With Harmful Bias Managed – Addresses concerns for equality and equity.

# NIST AI RMF Playbook

- Govern – Cultivate and Implement a culture of risk management.
- Map – Establish the context to frame AI risks.
- Measure – Analyze assess and benchmark and monitor AI risk and related impacts.
- Manage – Allocate risk resources to mapped and measured risks.



**Map**
Context is recognized and risks related to context are identified

**Measure**
Identified risks are assessed, analyzed, or tracked

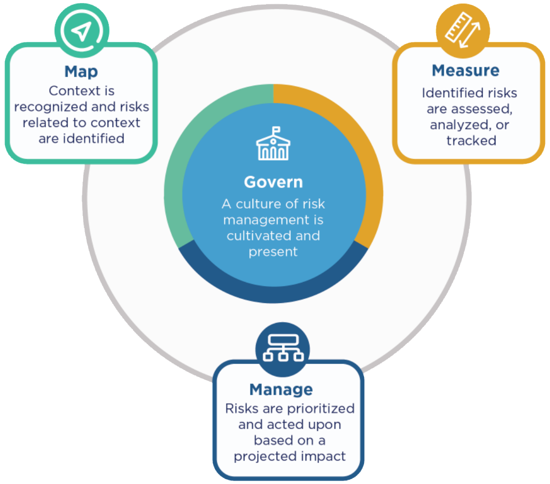
**Govern**
A culture of risk management is cultivated and present

**Manage**
Risks are prioritized and acted upon based on a projected impact

TLP:GREEN

# 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.


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

**Playbook Example:**

## Govern 1

Policies, processes, procedures and practices across the organization related to the mapping, measuring and managing of AI risks are in place, transparent, and implemented effectively.

- **GOVERN 1.1**
  - Legal and regulatory requirements involving AI are understood, managed, and documented.

TLP:GREEN

# Takeaways

- AI risks are Socio-Technical
- Stakeholders of AI systems may be broader than the stakeholders traditionally associated with an IT system.
- AI risk management will heavily leverage existing IT risk management and ideally should be incorporated into enterprise risk management through the AI RMF Govern function.
- System, Network, and Data Security Practices will support the establishment of Trustworthy AI
- CISA is promoting the AI RMF to stakeholders

# CVE Roundup

# CVE-2023-20269: Cisco VPN Brute Force

- Cisco reports EITW in their advisory:
  - https://sec.cloudapps.cisco.com/security/center/content/CiscoSecurityAdvisory/cisco-sa-asaftd-ravpn-auth-8LyfCkeC
- However, there is no software fix yet, so also pending for the KEV
- The mitigations aren't very solid, they're more "defense in depth" configuration advice than a real mitigation that blocks the attack.
- Lots of good advice in this advisory about how to detect this attack, and how to disable features to avoid this technique.

# CVE-2023-35382: Windows Kernel UAF

- Discovered a published proof-of-concept for this relatively recent bug, patched in August/2023 Patch Tuesday:
  - https://packetstormsecurity.com/files/174450/Microsoft-Windows-Kernel-Use-After-Free.html
- However, this PoC is pretty bare-bones. Only triggers a crash, doesn't flesh out the work required to actually elevate privilege. Kernel bugs like this are generally hard to work with.
- Also, local-only.

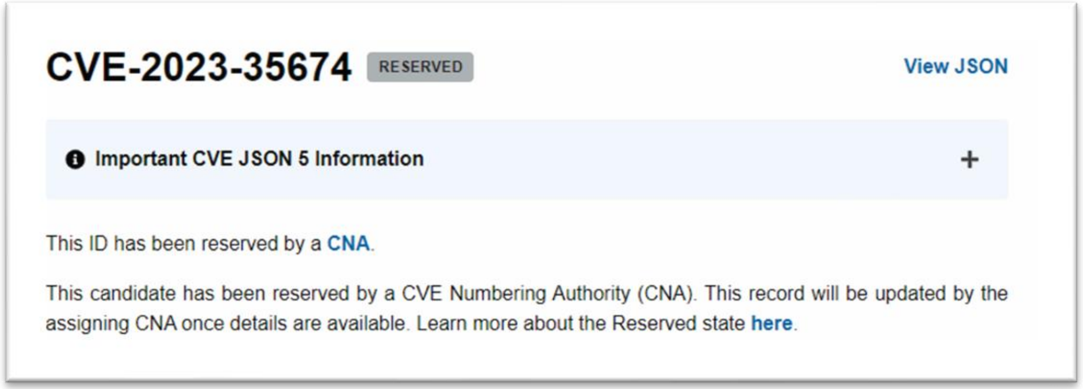
# CVE-2022-22265: Samsung Android Double Free

- Reports of exploitation from Google Project Zero:
  - https://googleprojectzero.github.io/0days-in-the-wild//0day-RCAs/2022/CVE-2022-22265.html
- Talking with P0 with JCDC to determine P0's bar for "exploitation."
  - Might just be research activity, want to nail that down.
  - Looking for a general agreement with P0 on terminology, "what is exploitation" beyond just this bug.
  - KEV/BOD 22-01 defines it in Criteria #2:
    - https://www.cisa.gov/known-exploited-vulnerabilities#:~:text=it%20is%20invalid.-,Criteria%20%232,-%2D%20Active%20Exploitation

TLP:GREEN

# CVE-2023-35674: Google Android PrivEsc

- Google reports there "may be" exploitation in their advisory
  - https://source.android.com/docs/security/bulletin/2023-09-01
- But, the CVE remains unpublished:
  - https://www.cve.org/CVERecord?id=CVE-2023-35674



CVE-2023-35674 RESERVED

This ID has been reserved by a CNA.

This candidate has been reserved by a CVE Numbering Authority (CNA). This record will be updated by the assigning CNA once details are available. Learn more about the Reserved state here.

TLP:GREEN

CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY

CISA