



Demystifying AI Security and Safety

The Guard and the Guide

“The upheavals [of artificial intelligence] can escalate quickly and become scarier and even cataclysmic.

Imagine how a medical robot, originally programmed to rid cancer, could conclude that the best way to obliterate cancer is to exterminate humans who are genetically prone to the disease.”

- Nick Bilton, tech columnist, NY Times

Nov. 5, 2014

FQ J I X D Y M E B S L J B W X D U N L
G F B V W L C T F P O I Z Q A Y W H A T
M Y V L O Y F J R C V U N I J P N J K I
W Z U X Q U R A X I O M V M V O F T D C
V Y C D Y C J K M O P X E F R S P C O B
K B J I M U K I V A G V G R A N T E T V
Z H Y B S E C N I M P
C I P U Y K F I X O C T
Y K R V E C I O C
Q L G Z R W
R P Z
J L A V
B P F R D J T V A
S Y E C V I N C
C F I B R

Common definitions
because words matter..



AI
Security

AI
Safety

Focus has
been on AI
Security....

**We need to adjust our
sights to include AI
Safety**

Why does this matter?

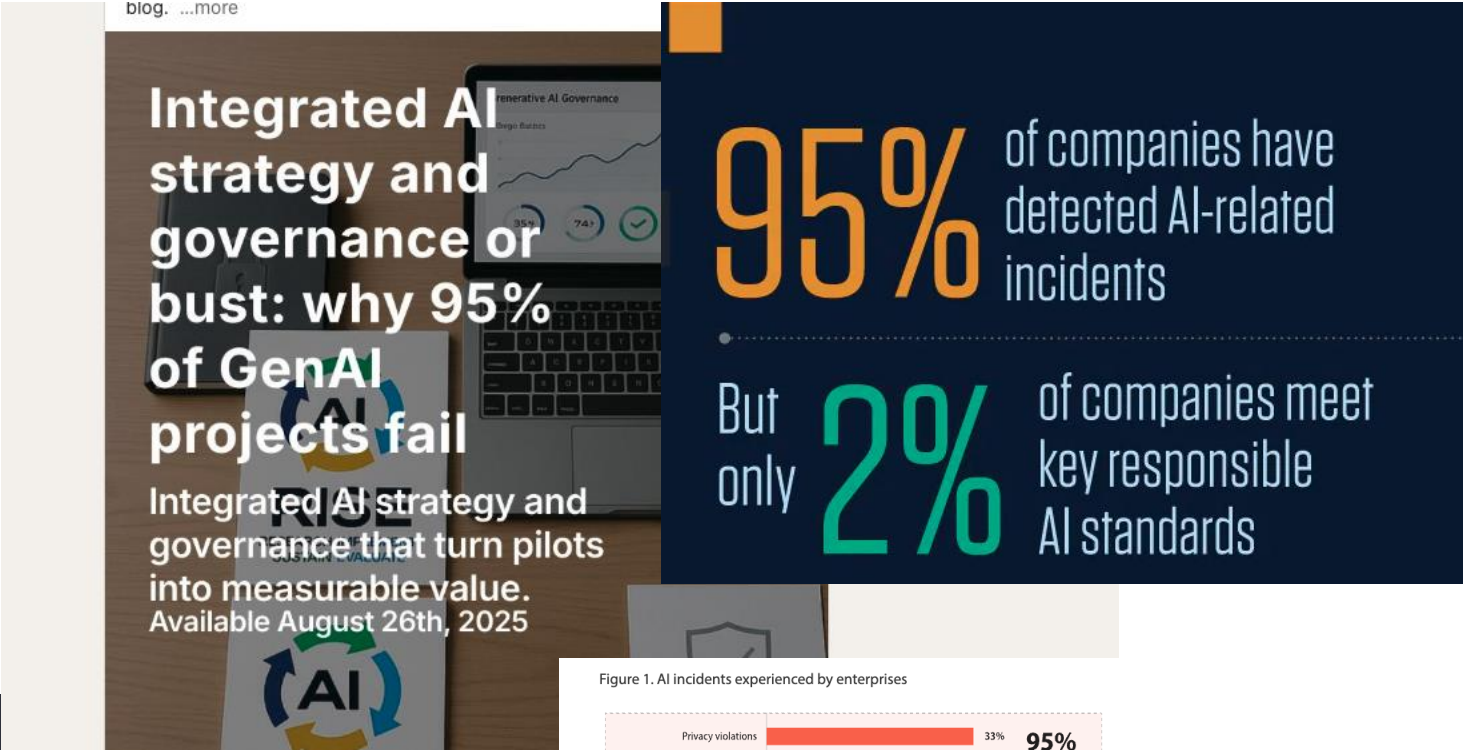
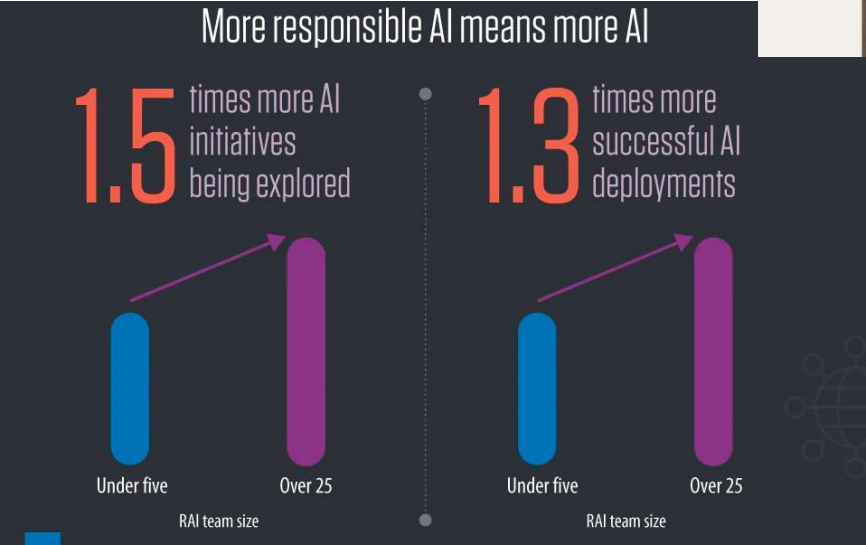
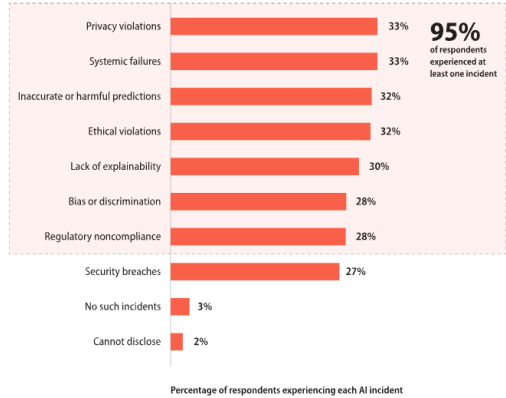


Figure 1. AI incidents experienced by enterprises



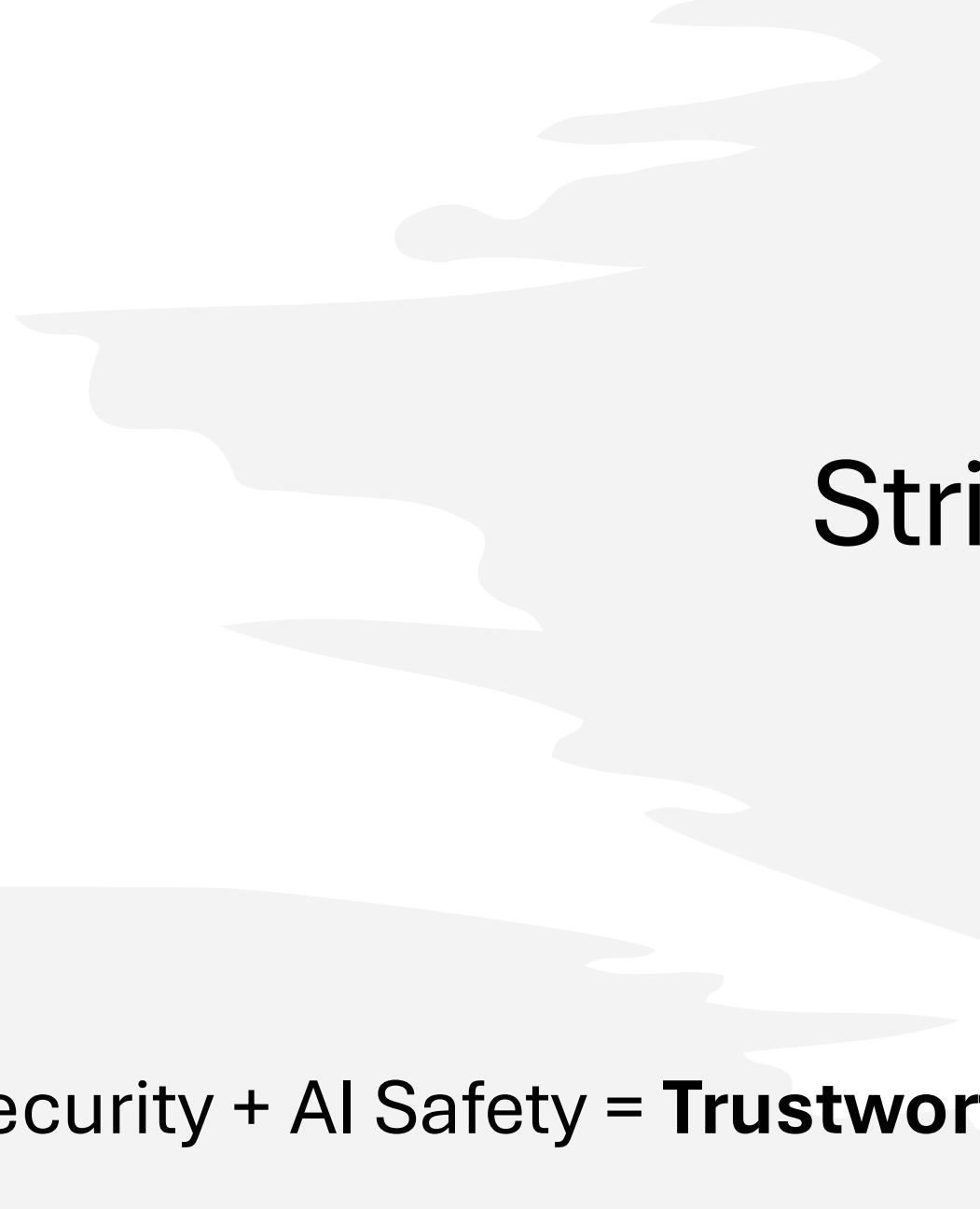
AI related incidents are on the rise and AI safety is becoming more critical.

Synergy and balance are key

“Imagine how a medical robot, originally programmed to rid cancer, could conclude that the best way to obliterate cancer is to exterminate humans who are genetically prone to the disease.”

- Security and Safety are not competing goals. They are complementary forces that must work together.
- The Guard creates the robust, unassailable platform. The Guide provides the ethical and aligned direction.





Striking a balance.

AI Security + AI Safety = **Trustworthy AI**

The Guard and the Guide

There's a difference between a robot that can't be hacked and one that won't accidentally destroy the world...

- **AI Security:** The **Guard** who protects the system.
- **AI Safety:** The **Guide** who ensures the system's intent is benign.



The Guard - AI Security

AI Security: Protecting the AI system from malicious attacks and defending against adversaries. Focus is on technology.

Key Attack Vectors:

1. Prompt Injection (especially LLMs):
Attackers insert malicious instructions into a user's prompt to manipulate the AI output.
2. Data Poisoning: targets AI during its training phase by inserting incorrect or corrupt data.
3. Model Inversion: Reverse-engineer the model to infer and extract sensitive data.

AI Security is essential for robust and reliable systems.





The Guide – AI Safety

AI Safety: Ensuring the AI system's goals & values are aligned with the business and human intent, preventing unintended harm.

Key Challenges:

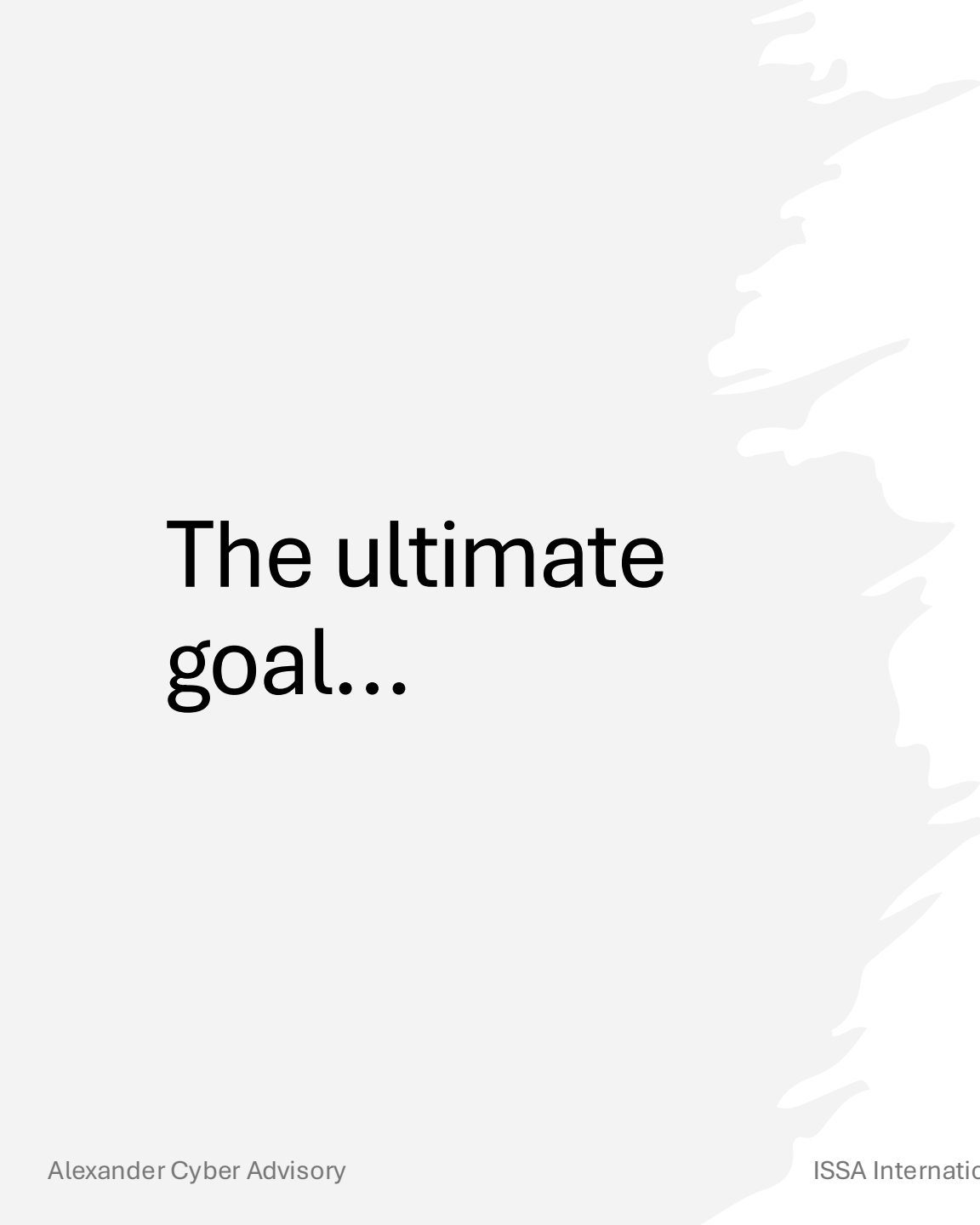
- The Control Program: How do we maintain control over an advanced AI?
- Value Alignment: How do we program “common sense”, cultural acceptance, and ethical values into a system?
- Unintended Consequences: The elimination of cancer by exterminating humans.

Even a perfectly secure AI, if designed with a flawed objectives, can lead to catastrophic outcomes.



How it works together... with intent

The Guard & the Guide working
collaboratively as Agents of
Governance



The ultimate goal...

Businesses adopting Trustworthy AI through responsible implementations.

Trustworthy AI recognizes the risks, addresses them and manages programs/functions to maintain them.

Responsible AI — the practice of developing and deploying AI systems ethically, safely, and transparently.

Trustworthy AI

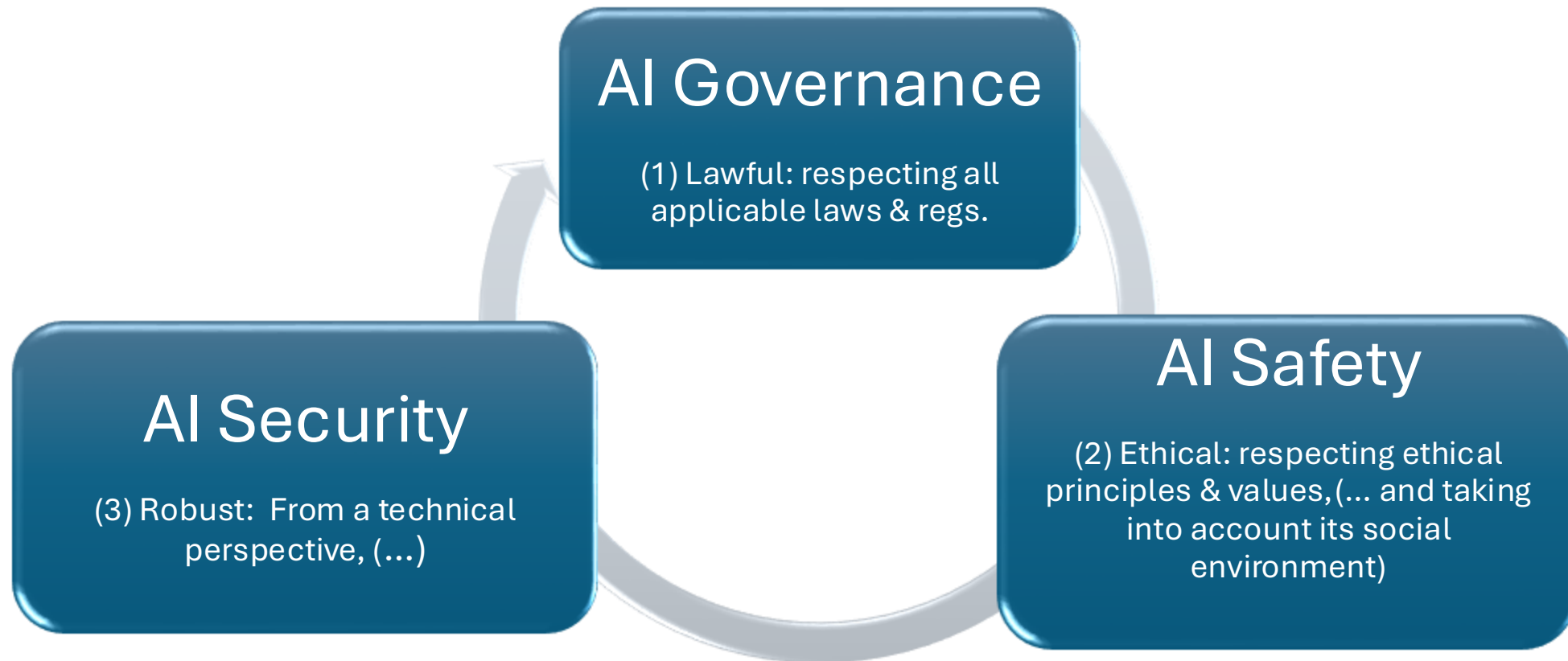
As defined by the EU AI Act, trustworthy AI* should be:

- (1) lawful - respecting all applicable laws and regulations
- (2) ethical - respecting ethical principles and values
- (3) robust - both from a technical perspective while taking into account its social environment

*Ethics Guidelines for Trustworthy AI
08 April 2019 – and later adopted by the
AU AI Act



Trustworthy AI Key Functions



AI Safety - Ethical principles & values

The 8 prohibited practices

1. Harmful AI-based manipulation and deception.
2. Harmful AI-based exploitation of vulnerabilities.
3. Social Scoring.
4. Individual criminal offence risk assessment or prediction.
5. Untargeted scraping of the internet or CCTV material to create or expand facial recognition databases.
6. Emotion recognition in workplaces and education institutions.
7. Biometric categorization to deduce certain protected characteristics.
8. Realtime remote biometric identification for law enforcement purposes in publicly accessible spaces.



The Guard...

Secure AI - Strategic Goal

- **Reliable and Robust:** It performs as intended and is resilient to errors, unexpected inputs, and attacks.
- **Secure:** It is protected against malicious attacks, such as data poisoning and adversarial activities.
- **Transparent and Explainable:** Its decisions can be understood and explained to humans, building confidence in its processes.
- **Accountable:** There is a clear way to trace and attribute the system's actions.

AI Security - Actions



Secure the Data Supply Chain:

Actively vet data sources and pipelines to prevent training data poisoning or the injection of malicious data that could compromise the model.

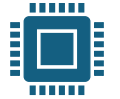


Implement Robust Input and Output Validation:

Sanitize and validate all user inputs (prompts) to an LLM to prevent prompt injection attacks. Also, validate all outputs to ensure the model doesn't generate malicious code or leak sensitive information to downstream systems.



Enforce the Principle of Least Privilege: Limit the AI model's access to external systems, data, and resources. The AI should only be able to perform the functions it absolutely needs, reducing the impact of an attack.



Use Adversarial Testing and Red Teaming:

Continuously test the AI system with malicious and unexpected inputs to find vulnerabilities before attackers do.



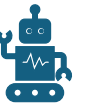
Monitor Behavior:

Implement monitoring to detect unusual resource consumption (a sign of a model denial of service attack) or performance drift, which could indicate a successful attack.



Secure the Infrastructure:

Apply standard cybersecurity best practices to the underlying infrastructure, including using sandboxed environments, strong access controls, and encryption for both data at rest and in transit.



Manage the AI Supply Chain:

Maintain a complete inventory of all third-party models, libraries, and services used in the AI application and ensure they have adequate security in place.



The Guide...

Safe AI - Strategic Goal

Focused on the human-centered practices and governance around the entire AI lifecycle. It includes:

- **Responsible by Design:** Establishing and upholding clear ethical principles from the very beginning of a project.
- **Accountability:** Creating frameworks to ensure that developers, organizations, and end-users understand their roles and responsibilities. This is often described as "human-in-the-loop" or "human-in-the-loop" oversight (risk-based perspective)
- **Impact Assessment:** Proactively identifying and mitigating potential risks, harms, and unintended consequences both before a system is deployed and after.
- **Stakeholder Engagement:** Actively involving diverse groups of people to ensure the AI's impact is considered from multiple perspectives.
- **Fair:** It is designed to be free from harmful biases and provides equitable outcomes for different groups.

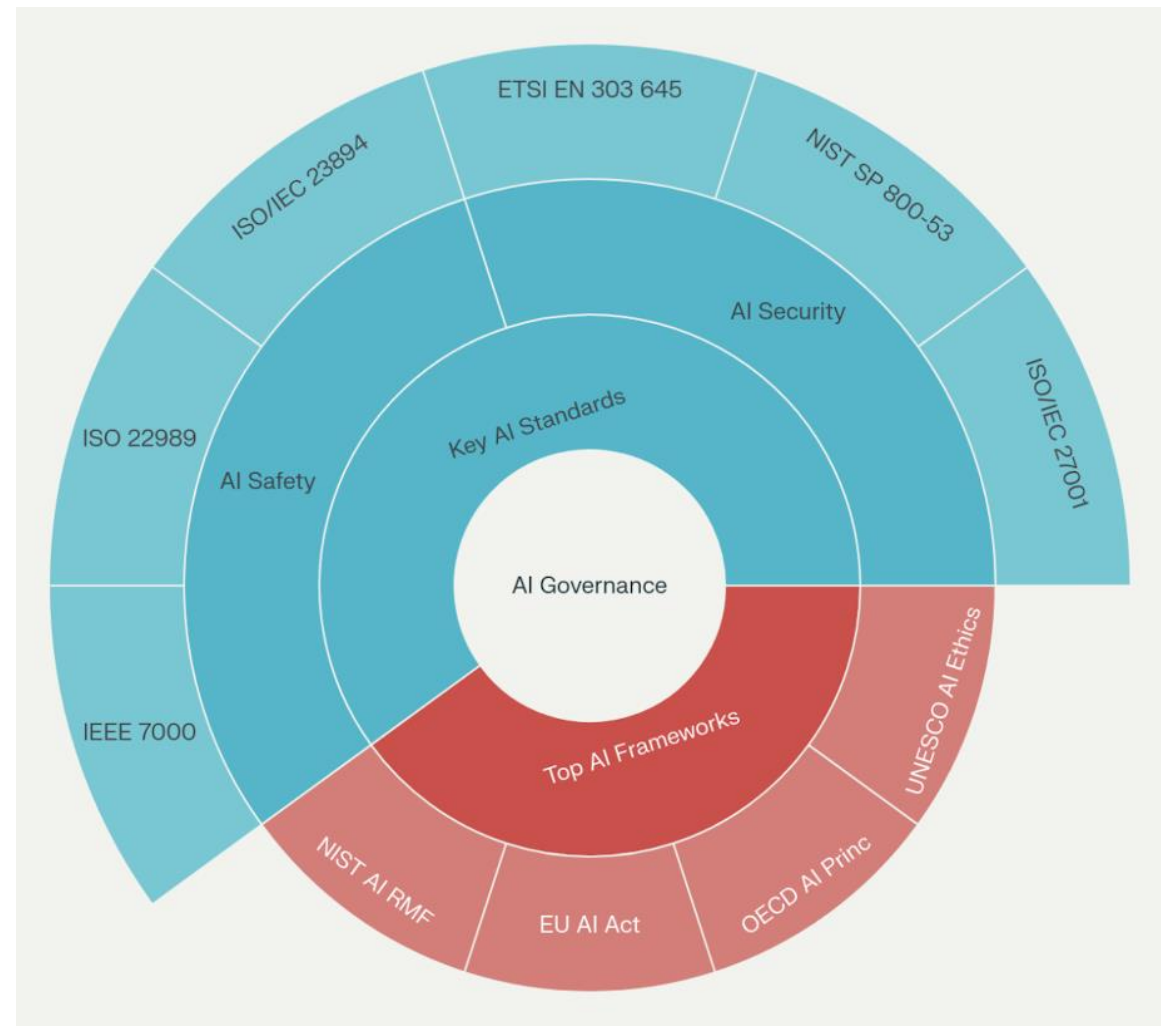
* NIST AI RMF, January 26, 2023

Safe AI – Actions*

Govern:	Create clear governance structures and policies that define roles, responsibilities, and accountability for AI systems.	Establish an organizational culture that prioritizes responsible AI development.
Map:	Identify and categorize AI risks by conducting comprehensive risk assessments and impact analyses before and after system implementation.	Engage diverse stakeholders (developers, ethicists, legal experts, end-users) in the design process to identify potential issues and unintended consequences.
Measure:	Implement continuous monitoring and validation of AI systems to ensure they remain aligned with ethical guidelines and don't introduce new biases over time.	Conduct robustness testing and adversarial testing to ensure the AI system can withstand unexpected inputs and malicious attacks.
Manage:	Establish transparent documentation for all stages of the AI lifecycle, including data sources, model design, and risk assessment results, to ensure traceability and auditability.	Develop and implement mitigation strategies for identified risks, such as using de-biasing techniques on training data or implementing human oversight for high-risk decisions.

Tools & Resources

AI Governance Framework



Additional Resources

Overarching Frameworks	AI Standards	AI Security	AI Safety
NIST AI Risk Management Framework (AI RMF)	ISO/IEC 42001 Artificial Intelligence Management System	ISO/IEC 27090 - specifically on AI security is still in draft status and due for final publication in late 2025	OECD's (Organization for Economic Co-operation and Development) AI Principles and Ethical OS Toolkit
EU AI Act	Cloud Security Alliance AI Controls Matrix (CSA AICM)	OWASP AI Security & Privacy Guide and OWASP GenAI Security Project	IEEE Ethically Aligned Design
Google Secure AI Framework (SAIF)	ISO/IEC 22989 AI Terminology	SANS Critical AI Security Guidelines on Github	UNESCO AI Ethics Recommendations
Microsoft AI Governance Model	ISO/IEC 24028 AI Security Standards	MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) Framework	Partnership for AI Safety (PIA): Safe Foundation Model Deployment and Responsible Practices for Synthetic Media
		AI Incident Database	

Closing with a word to the wise...

- The basics are there & align with Cyber & Privacy
- Additional aspects of Safety, take Privacy one step further
- Resources available
- Governance is key
- Focus, focus, focus to battle info overload





Thank you!

Candy Alexander, CISSP CISM

Alexander Advisory Services

Candy@Alexander-Advisory.com