About Adversa.AI

Comments are provided by Eugene Neelou and Alex Polyakov, experts in cybersecurity, artificial intelligence, and risk management. Both are founders of Adversa.AI, a world-class Israeli startup on a mission to increase trust in AI systems.

Recently, Adversa.AI has released the unique report called "The Road to Secure and Trusted AI: The Decade of AI Security Challenges" that demonstrates the progress toward AI trustworthiness over the past 10 years across academia, government, and industry.

Besides many insights and infographics, the Adversa Report has introduced 2 essential resources for the development of the AI Risk Management Framework:

1. The Map of Trustworthy AI covers all principles and requirements for trusted AI.
2. The Lifecycle for Secure AI describes activities for all security stages for AI systems.

Please find our comments first and the mentioned resources in the end.

Adversa's Comments for the NIST AI RMF

Challenges in Managing AI Risk (RFI Topic 1)

There are various challenges in how AI-related risks are managed, including technical and organizational problems. However, reports show that AI risks are often mismanaged because companies are not simply aware of them.

To learn about such risks so they could be addressed in the first place, companies should have AI risk awareness training programs not just for AI engineering but for everyone involved in AI projects. This is similar to how cybersecurity is everyone's concern and security awareness programs are essential for cyber resilience.

We propose that NIST could list activities such as risk awareness explicitly as a prerequisite for successful AI risk management. Thus, organizations will know what they need to protect from.

Principles of AI Trustworthiness (RFI Topic 3)

There are many dimensions of AI trustworthiness. Unfortunately, terms for them often have unclear and confusing definitions. Also, the same terms may be used interchangeably or describe both characteristics and principles of AI trustworthiness. The absence of system thinking confuses AI stakeholders and prevents them from addressing AI risks efficiently.

Moreover, the dimensions of AI trustworthiness are managed by very different teams with different skillsets. To be actionable, the terms must be grouped by types of AI risks and aligned with organizational structure so it's easier to discuss with the right AI stakeholders and identify the right teams to manage those AI risks.

We suggest the following groups of principles for AI trustworthiness:

1. "Reliable AI" principles describe how AI systems operate in normal conditions.
   Reliable AI principles include characteristics of Robustness, Accountability, Transparency.

2. "Resilient AI" principles describe how AI systems operate in adversarial conditions.
   Resilient AI principles include characteristics of Security, Privacy, Safety.

3. "Responsible AI" principles describe how AI systems comply with social norms.
   Responsible AI principles include characteristics of Fairness, Ethics, Sustainability.

We propose that NIST could provide a clear hierarchy and definitions while ensuring that principles cover all characteristics of AI trustworthiness. The terms should be defined the way it's clear how different dimensions of AI trustworthiness relate, overlap, or complement each other. Thus, it would be clear which teams are more suitable to manage certain types of AI risks.

AI Governance Issues (RFI Topic 12)

AI governance is an essential part of the AI risk management framework. It's not possible to address AI risk efficiently without handling organizational, technical, and operational issues.

We propose that NIST could cover all stages of AI development from the first ideas and prototypes to the maintenance of production AI systems. It's important to emphasize that AI risk management is not a one-time project but rather a regular process that requires team training, operational guidelines, incident response plans, and appropriate tooling. Thus, organizations will be empowered to overcome challenges toward efficient AI risk management.

Adversa's Resources for AI Risk Management

Adversa.AI is a world-class team that develops AI risk management methods and technologies.

Recently, Adversa.AI has released the unique report called "The Road to Secure and Trusted AI: The Decade of AI Security Challenges" that demonstrates the progress toward AI trustworthiness over the past 10 years across academia, government, and industry.

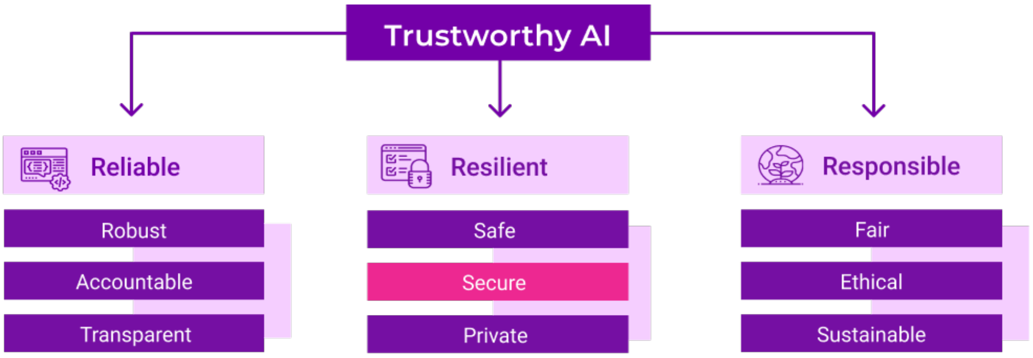
Besides many insights and infographics, the Adversa Report has introduced 2 essential resources for the development of the AI Risk Management Framework:

1. The Map of Trustworthy AI covers all principles and requirements for trusted AI.
2. The Lifecycle for Secure AI describes activities for all security stages for AI systems.

Please find these resources below.

Adversa’s Map of Trustworthy AI

The Map of Trustworthy AI covers principles and requirements for AI systems.



“Reliable AI” principles describe how AI systems operate in normal conditions.

- Robustness means the abilities of AI systems to function as intended, behave predictably especially in situations of uncertainty, and fail gracefully in case of fatal errors.
- Accountability explains how to operationalize AI systems in a responsible way by providing human oversight and control over AI behaviors and outcomes.
- Transparency introduces mechanisms for traceable and explainable AI decisions as well as policies that respect human rights to know when they interact with AI systems.

“Resilient AI” principles describe how AI systems operate in adversarial conditions.

- Safety is required for mission-critical situations and human-computer interactions, where an AI system should prevent any harm to living beings or the environment.
- Security makes AI systems resilient to malicious activities and adversarial attacks such as general software security attacks or specific attacks against AI algorithms.
- Privacy demands AI systems to implement adequate data governance mechanisms, data protection, and access controls for collected and inferred data.

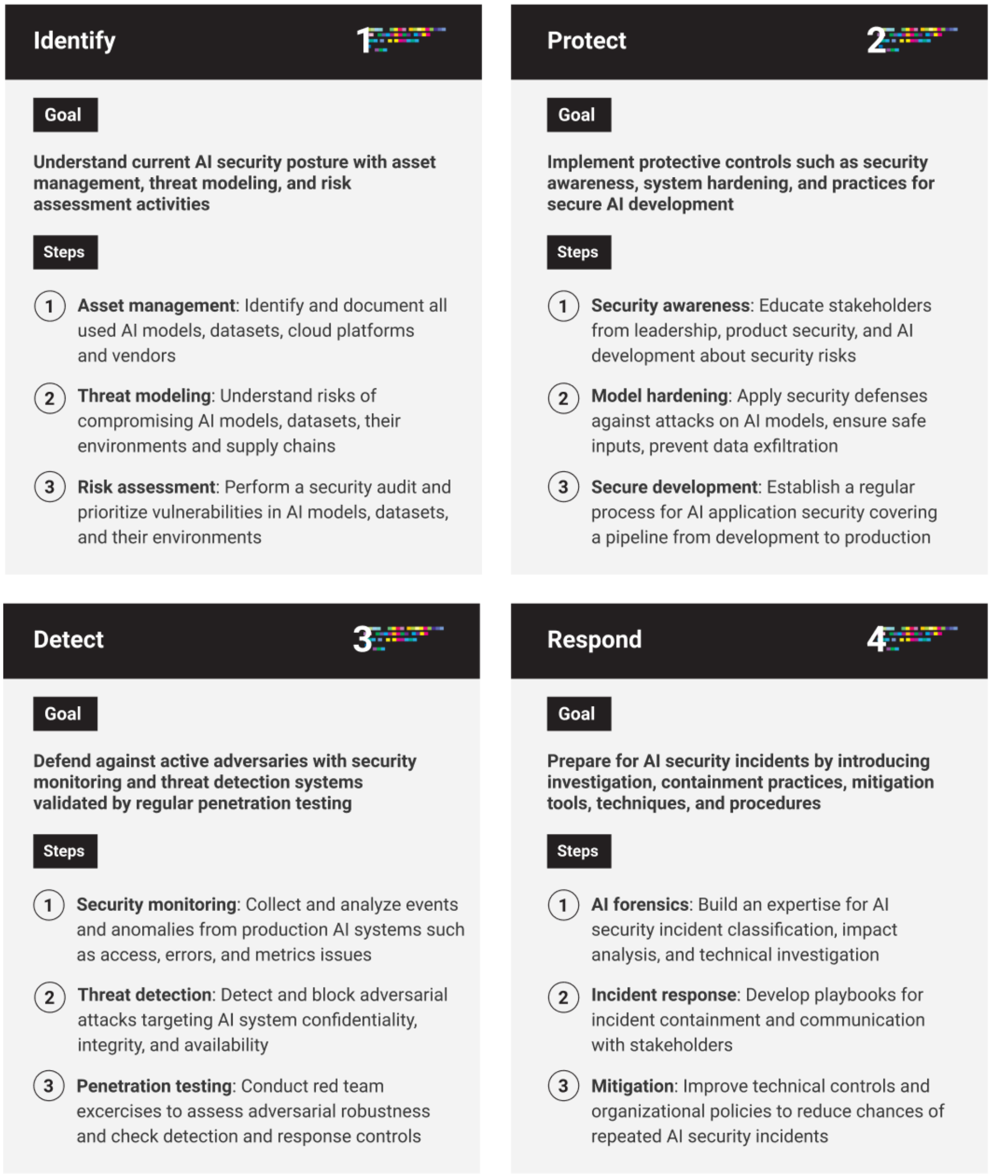
“Responsible AI” principles describe how AI systems comply with social norms.

- Fairness enables inclusion, diversity, and accessibility by introducing practices for examining bias and counting the interests of everyone affected by AI systems.
- Ethics touches on moral principles of interaction between humans and AI systems, which should respect human rights and empower people, not the opposite.
- Sustainability lets organizations use AI algorithms to make people's lives better by meeting today’s needs without compromising the interests of future generations.

Adversa's Lifecycle for Secure AI

The Lifecycle for Secure AI describes the required steps for all security stages for AI systems.

The four areas represent the AI system's stages from inception to maturity and looped back for continuous improvement. The steps are ordered from basic to complex with later steps relying on preceding outcomes.



Identify 1

Goal

Understand current AI security posture with asset management, threat modeling, and risk assessment activities

Steps

1. **Asset management**: Identify and document all used AI models, datasets, cloud platforms and vendors
2. **Threat modeling**: Understand risks of compromising AI models, datasets, their environments and supply chains
3. **Risk assessment**: Perform a security audit and prioritize vulnerabilities in AI models, datasets, and their environments

Protect 2

Goal

Implement protective controls such as security awareness, system hardening, and practices for secure AI development

Steps

1. **Security awareness**: Educate stakeholders from leadership, product security, and AI development about security risks
2. **Model hardening**: Apply security defenses against attacks on AI models, ensure safe inputs, prevent data exfiltration
3. **Secure development**: Establish a regular process for AI application security covering a pipeline from development to production

Detect 3

Goal

Defend against active adversaries with security monitoring and threat detection systems validated by regular penetration testing

Steps

1. **Security monitoring**: Collect and analyze events and anomalies from production AI systems such as access, errors, and metrics issues
2. **Threat detection**: Detect and block adversarial attacks targeting AI system confidentiality, integrity, and availability
3. **Penetration testing**: Conduct red team excercises to assess adversarial robustness and check detection and response controls

Respond 4

Goal

Prepare for AI security incidents by introducing investigation, containment practices, mitigation tools, techniques, and procedures

Steps

1. **AI forensics**: Build an expertise for AI security incident classification, impact analysis, and technical investigation
2. **Incident response**: Develop playbooks for incident containment and communication with stakeholders
3. **Mitigation**: Improve technical controls and organizational policies to reduce chances of repeated AI security incidents

This lifecycle corresponds with NIST Cybersecurity Framework and Gartner's Adaptive Security Architecture, popular reference frameworks for cybersecurity lifecycle management.