Deloitte.

12th eGovernment Forum

Designing Trustworthy AI: AI Security by Design & Attack Vectors

October 21, 2025

AI use cases in Government and Industry

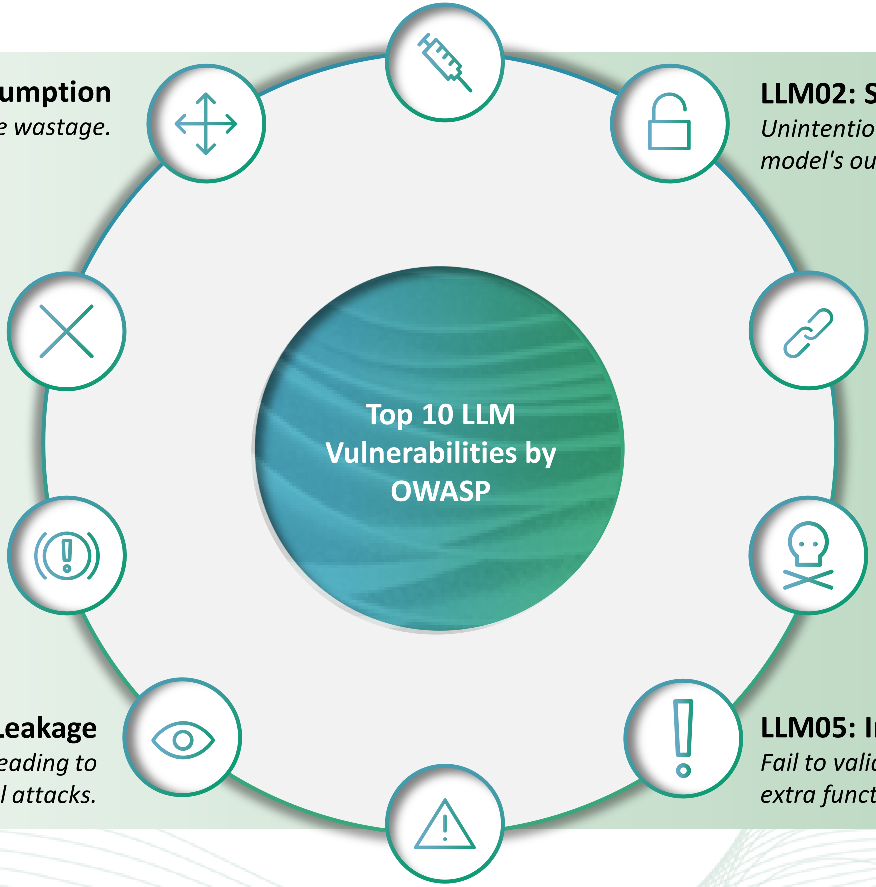
The impact of the vulnerabilities listed below is interconnected between the components of the system. The effects of one weakness will cascade throughout, highlighting the critical need to safeguard every component and ensure the integrity and resilience of the overall system.

Representative Use Cases by Value…	Energy, Resources & Industrials	Financial Services & Insurance	Government & Public Sector	Technology, Media, & Telecom	Life Sciences & Health Care	Education Sector
Cost Reduction	Automation & Robotics	Insurance Claims Processing and Validation	Fraud and Abuse Detection	Autonomous Coding	Autonomous Triage Services	Adaptive Learning Platforms for Student Support
Speed to Execution	High-Performance Asset Design	Fraud Detection in Customer Transactions	Regulation / Policy Document Generation	Ad, Media, Gaming Content Generation	Personalized Healthcare Treatment Plans	Automated Grading and Assessment Tools
Reduced Complexity	Predictive Insights for Utility Outages	Know Your Customer Analysis	Autonomous Governmental Operations	Autonomous Infrastructure Operations	Medical Imaging Analysis Reports	Curriculum Design through Learning Analytics
Transformed Engagement	Field Workforce Safety Virtual Assistant	Digital Banking Assistants	Benefits Administration Service Recommendations	Autonomous Network Operations	Virtual Personal Health Assistants	AI Tutoring and Virtual Class Assistants
Fueled Innovation	Rapid Nanomaterial R&D	Intelligent Marketing and Distribution Analytics	Climate Pattern and Impact Monitoring	Real-time Language Translation for Virtual Meetings	Predictive Insights on Personal Healthcare Data	Immersive Learning through AI-Generated Simulations
New Markets	Asset Performance Simulation	Financial Forecasting & Portfolio Generation	On-Site Autonomous Service Desks	Synthetic Data Generation for AV Simulation	Biomedical Data Science for Disease Discovery / Prevention	Global Remote Education through AI Platforms

© 2025 Deloitte Consulting WLL Co. All rights reserved.

How AI Can Be Compromised

Highlighting the top 10 vulnerabilities targeting Large Language Models as defined by OWASP.



Top 10 LLM Vulnerabilities by OWASP

LLM01: Prompt Injection
Manipulate models behavior by embedding hidden inputs that influence the generated output.

LLM02: Sensitive Information Disclosure
Unintentional sensitive data input can lead to its exposure in the model's output.

LLM03: Supply Chain
Supply chain vulnerabilities impact data, models, platforms causing bias, security breaches, or system breakdowns.

LLM04: Data and Model Poisoning
Manipulate training data or models to introduce biases, or vulnerabilities that affect the performance and security.

LLM05: Improper Output Handling
Fail to validate model outputs, allowing users unintended access to extra functionality.

LLM06: Excessive Agency
Pose risks by allowing models to independently decide and execute actions.

LLM07: System Prompt Leakage
Expose hidden sensitive information in model instructions, leading to potential attacks.

LLM08: Vector and Embedding Weaknesses
Exploit weaknesses in vector and embeddings to inject harmful content , manipulate outputs, or access sensitive data

LLM09:Misinformation
Provide fake information through hallucinations, and overlaying causing risks, reputational damage, and legal problems.

LLM10: Unbounded Consumption
Generate outputs without limits, leading to resource wastage.

© 2025 Deloitte Consulting WLL Co. All rights reserved.

Secure GenAI Adoption and Risk Mitigation

Adopting Generative AI requires careful consideration to address potential risks. Embracing innovation while ensuring security will unlock the full potential of generative AI effectively and responsibly.

Doing Nothing

The most common and risky approach, could expose you to data leakage risks, potential legal consequences, and many other risks.

Single-Purpose Solutions

Relying solely on data protection with no security integration and observability leads to gaps in threat detection.

Monolithic Security Solutions

Relying on a secure LLM provider leaves other LLMs in your ecosystem unprotected, and presents risks of vendor lock-ins.

Enterprises spent on GenAI solutions an estimated of

$19.4 Billion

In 2023

$151 Billion

By 2027

Reference - International Data Corp.

With no robust internal Governance policies set to secure GenAI solutions, enterprises will not trust the output generated by proprietary LLM models. Below we are highlighting key controls to securely adopt GenAI:

1 ***Privilege Control***

Ensure controlled access to the LLM's backend systems, ensuring permissions are limited.

2 ***Human Intervention***

To reduce risks of executing unauthorized actions, add an extra layer of security by enforcing human approval.

3 ***Proper Segregation***

Ensure external data sources are properly segregated, ensuring minimal external influence on the user's prompts.

4 ***Periodic Monitoring***

Regularly perform audits to ensure any detected system vulnerabilities are properly addressed.

5 ***Zero-Trust Approach***

Treat the model as an untrusted entity by implementing the zero-trust framework, to ensure inputs are properly validated.

6 ***Supply Chain Data***

Ensure the training data of external sources are legitimate. Maintain attestations using Machine Learning Bill of Materials (ML-BOM).

7 ***Training Data***

Assess and validate the authenticity of the data sources used for pre-training, fine-tuning, and many other processes.

8 ***ML SecOps Approach***

Integrate adversarial robustness into the training lifecycle through adversarial training techniques.

9 ***Input Validation***

Implement input validation and sanitization to ensure malicious content is filtered out.

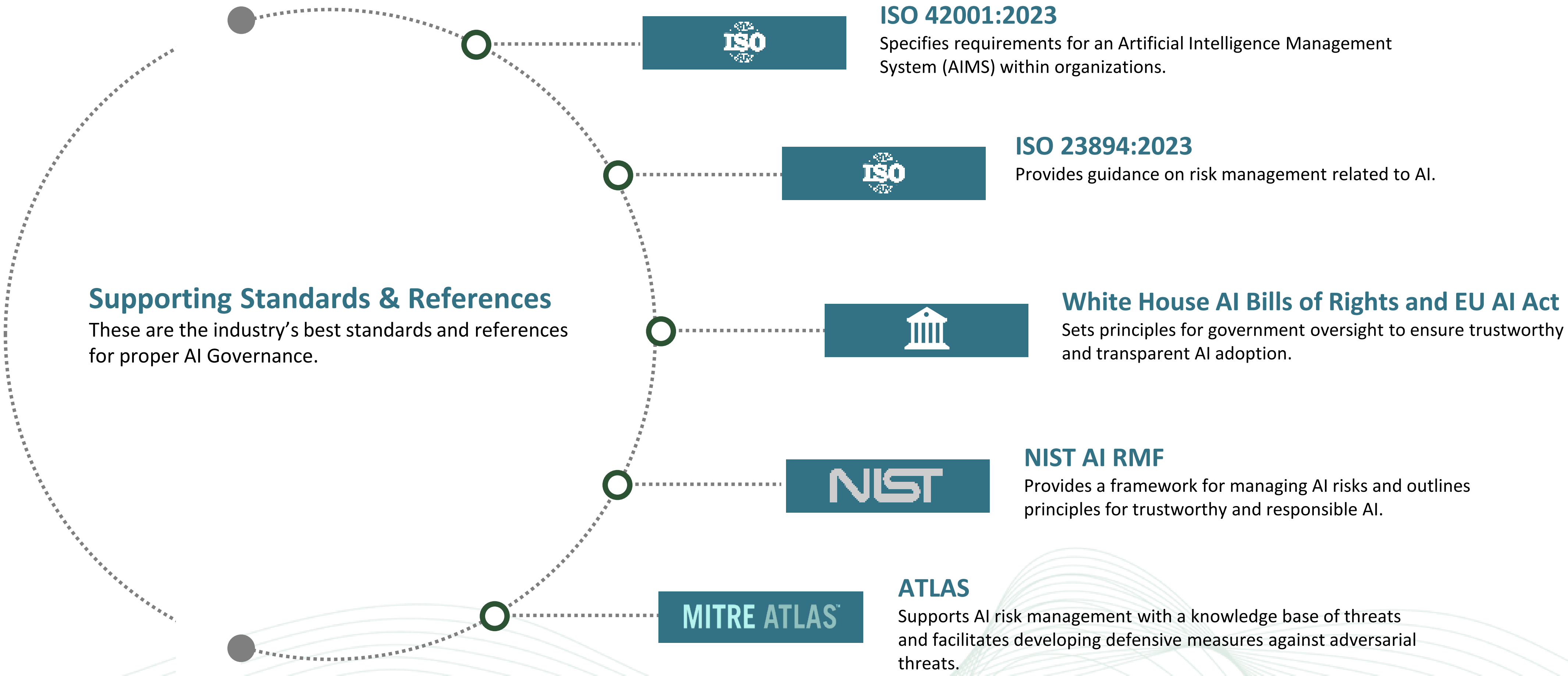
10 ***Awareness Promotion***

Promote awareness among personnel and provide guidelines for secure LLM usage.

© 2025 Deloitte Consulting WLL Co. All rights reserved.

AI Governance Industry Standards

Below we highlight key industry standards and governmental acts that support AI governance within an organization which enables effective risk management and ensures compliance.



Supporting Standards & References

These are the industry's best standards and references for proper AI Governance.

ISO 42001:2023

Specifies requirements for an Artificial Intelligence Management System (AIMS) within organizations.

ISO 23894:2023

Provides guidance on risk management related to AI.

White House AI Bills of Rights and EU AI Act

Sets principles for government oversight to ensure trustworthy and transparent AI adoption.

NIST AI RMF

Provides a framework for managing AI risks and outlines principles for trustworthy and responsible AI.

ATLAS

Supports AI risk management with a knowledge base of threats and facilitates developing defensive measures against adversarial threats.

© 2025 Deloitte Consulting WLL Co. All rights reserved.

Deloitte’s Trustworthy AI Framework

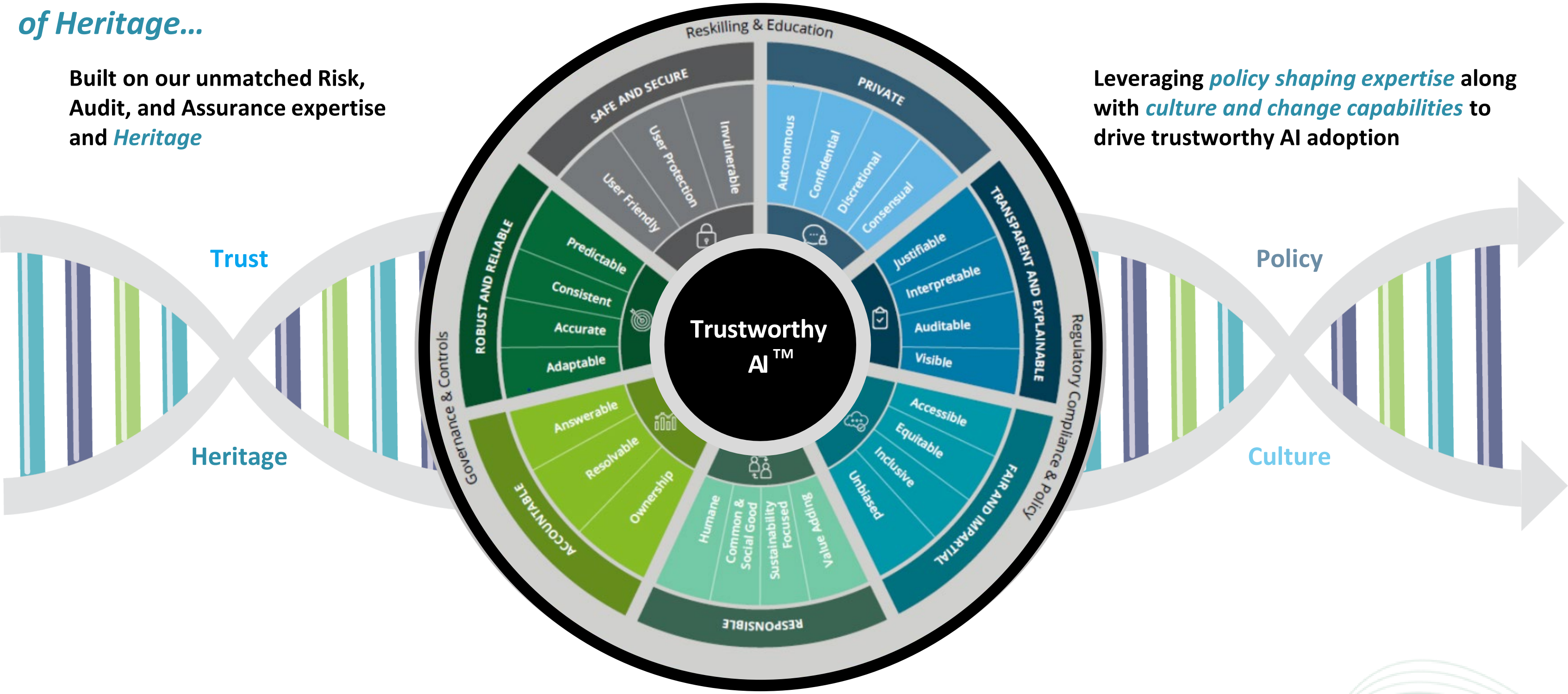
Deloitte has developed the Trustworthy AI framework which is rooted in our Risk, Audit and Assurance heritage, to responsibly harness the power of AI for the benefit of shareholders, and society at large.

More than 175 years of Heritage…

Built on our unmatched Risk, Audit, and Assurance expertise and *Heritage*

Uniquely positioned…

Leveraging *policy shaping expertise* along with *culture and change capabilities* to drive trustworthy AI adoption



AI Ethics and Strategy | AI Governance | AI Risk Management | AI Testing | Algorithm Assurance

Culture, Change, and Organizational Design

Infusing an Ethical AI Mindset Across our Talent

Tech & AI Ethics Program

Centre for AI

Delivering Trustworthy AI across Industries

- Financial Services
- Consumer & Industrial Products
- Healthcare & Life Sciences
- Technology

Shaping Global Policy

WORLD ECONOMIC FORUM

Notre Dame Deloitte Center for Ethical Leadership

ILLINOIS

UNIVERSITY OF OXFORD

Creating Compelling Thought Leadership…

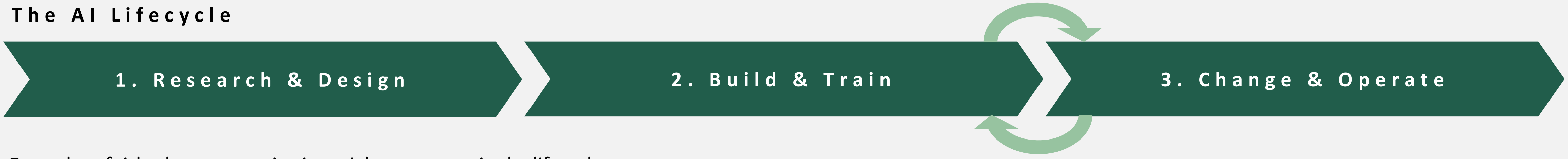
…including publications in the Wall Street Journal

© 2025 Deloitte Consulting WLL Co. All rights reserved.

How Organizations can Address AI Risks?

Identifying pertinent risks and corresponding de-risking toll-gates at each stage of the AI Lifecyle will help to align stakeholders around a consistent approach to manage risks.

The AI Lifecycle



1. Research & Design

2. Build & Train

3. Change & Operate

Examples of risks that an organization might encounter in the lifecycle

1. Research & Design

- *The solution's intended purpose is not aligned to the organization's mission and values.*
- *The solution's inherent risks (e.g. computer vision application that captures and potentially misuses customer / employee images or other PII) are overlooked during the design phase*

2. Build & Train

- *The data used to train an AI model has quality issues*
- *The organization lacks appropriate ways to secure consent from individuals whose data is used to train the AI model*

3. Change & Operate

- *The organization is unable to quickly assess which new data sources an AI-enabled solution has recently accessed on its own*
- *The organization lacks the ability to test and monitor AI solutions*

Risks (Based on the Trustworthy AI Framework)

1. Research & Design

- **Fairness:** Potential inherent bias in datasets identified and selected for use
- **Accountability:** Potential lack of clear accountability and ownership of the AI idea/use case
- **Transparency/Trust:** Potential misalignment with organization's core values and regulatory guidance/standards

2. Build & Train

- **Robustness:** AI model is not developed and properly tested prior to production
- **Fairness:** Datasets and/or model outputs have discriminatory bias on protected identifier (s)
- **Transparency:** Misalignment with regulatory guidelines and lack of transparency of the AI model's learning rulesets
- **Security:** Exposure of the AI model to internal and external threats

3. Change & Operate

- **Robustness:** AI system and data may produce unexpected outputs or be unable to adapt to changed environment or input factors
- **Fairness:** AI system develops bias or produces unexpected outputs
- **Security:** Data, physical, cyber threats, and unauthorized changes associated with AI
- **Transparency/Trust:** Misuse of AI system and customer data
- **Accountability:** Lack of accountability and ownership of unintended outcomes

© 2025 Deloitte Consulting WLL Co. All rights reserved.

Conclusion

While Generative AI offers significant benefits in organizations, the absence of proper controls may expose them to risks, emphasizing the importance of robust and flexible security controls against any possible emerging threats.

***Generative AI** adds significant benefits to our daily lives by significantly enhancing efficiency and driving innovation.*

*It's imperative that we implement robust security measures to protect the systems from **emerging threats**.*

© 2025 Deloitte Consulting WLL Co. All rights reserved.

شكراً لكم

Thank You

© 2025 Deloitte Consulting WLL Co. All rights reserved.