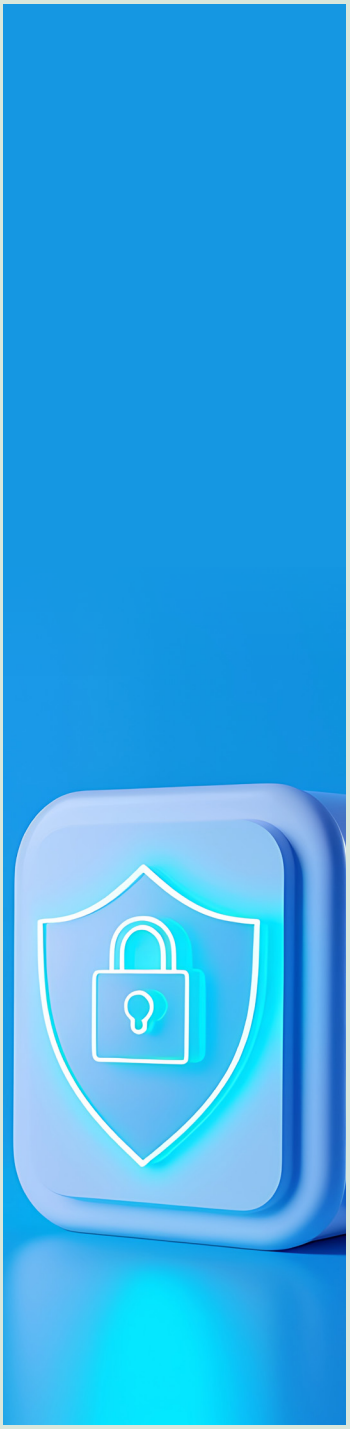
WHITE PAPER



BUILDING TRUSTWORTHY AI: A CYBERSECURITY-FIRST APPROACH TO RESPONSIBLE INNOVATION

Abstract

Artificial Intelligence (AI) has rapidly evolved from a niche technology to a core enabler of business transformation across industries. Organizations are leveraging AI to drive strategic decision-making, optimize operations, and enhance customer experience. However, as AI systems increasingly influence critical business and societal outcomes, the stakes for ensuring their reliability, fairness, and security have never been higher.

Trustworthy AI is no longer just a technical aspiration- it is a business and regulatory imperative. Insecure and poorly governed AI models can have severe consequences across business, societal, and political domains. This includes amplifying existing biases, widespread security breaches, data manipulation, financial losses, and degradation of public trust and non-compliance with emerging regulations such as the EU AI Act, NIST AI Risk Management Framework, and ISO/IEC 42001. A lack of security governance compounds all other risks, allowing small failures to become systemic crises.

This whitepaper provides guidance on building trustworthy AI and managing associated Cybersecurity risks through a comprehensive approach that spans the entire AI lifecycle—from data collection and model development to deployment, monitoring, and retirement. It emphasizes embedding principles of preemptive Cybersecurity, privacy, transparency, accountability, and ethical governance at every stage. Organizations that prioritize these principles will not only mitigate cybersecurity risks but also unlock sustainable value from their AI investments.

Infosys®
Navigate your next

Overview & Industry Problem

As the adoption of Artificial Intelligence (AI) accelerates across diverse industries, it brings transformative potential—but also introduces a new set of unforeseen challenges. AI systems, characterized by their complexity, opacity, and autonomous decision-making capabilities, pose unique risks that traditional cybersecurity frameworks are unable to address. These frameworks often fall short in identifying and managing the dynamic risk and evolving nature of AI, leaving organizations vulnerable to operational disruptions, financial losses, reputational damage, legal liabilities, and regulatory non-compliance.

The industry stands at a critical juncture: on one hand, AI offers unprecedented efficiency, innovation, and competitive advantage; on the other, its rapid deployment without appropriate guardrails can amplify systemic risks and erode stakeholder trust. If AI systems are developed and deployed without robust governance, security, and ethical oversight, it can have catastrophic impact on organizations and humanity. The absence of standardized controls and oversight mechanisms has led to a fragmented risk landscape, where vulnerabilities can propagate rapidly across interconnected systems, increasing exposure to operational, financial, legal, and reputational threats.

Key AI-specific risks include:

- **Phishing and other attacks:** Generative and agentic AI can enable hyper-personalized phishing attacks, deepfakes, and automated exploit development, making reconnaissance and attack execution faster and more effective.
- **AI-driven malware:** They can evolve in real time to evade detection, lowering the barrier for cyber criminals and increasing both attack volume and sophistication.
- **Model hallucinations:** Generation of false or misleading outputs that can compromise decision-making.
- **Opaque decision-making:** Lack of transparency and explainability, raising concerns around accountability and ethical oversight.
- **Adversarial attacks:** Threats such as model evasion, model extraction, data poisoning, and data leakage that exploit AI vulnerabilities.
- **Privacy violations:** Improper handling of personal and sensitive data, leading to breaches and non-compliance.
- **Regulatory gaps:** Emerging frameworks like the EU AI Act, GDPR, ISO/IEC 42001, and CCPA highlight the need for AI-specific governance.
- **Inadequate monitoring and reporting:** Limited visibility into AI system behavior post-deployment.
- **Insufficient governance and cybersecurity controls:** The rise of shadow AI and unregulated model usage expands the threat landscape.
- **Intellectual property concerns:** Use of scraped public or protected data for model training without proper authorization or attribution.

These challenges underscore the urgent need for a new paradigm in AI risk management—one that integrates cybersecurity, governance, and ethical principles throughout the AI lifecycle.

Refer a statistical view of AI specific Risks, source

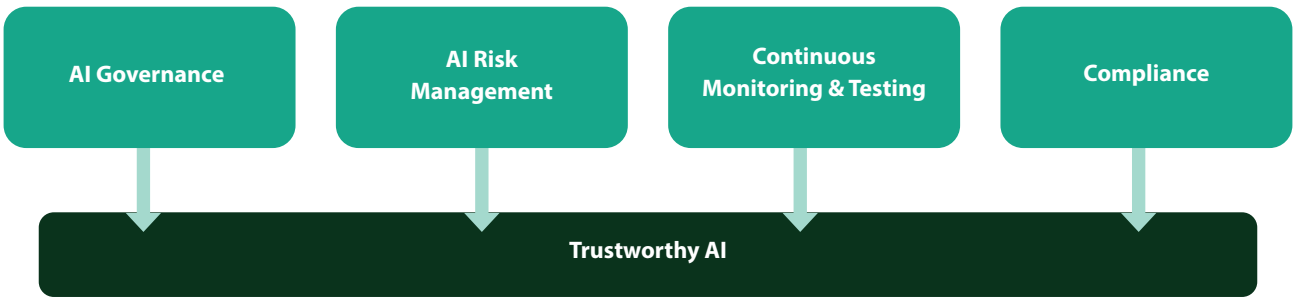
2024 KPMG AI Security
Benchmark Survey Results

Recommended Solution

Building blocks of Trustworthy AI

Building trustworthy AI requires a comprehensive framework anchored on four pivotal pillars: **AI Governance, AI Risk Management, Continuous Monitoring & Testing, and Compliance**. AI Governance ensures ethical alignment and accountability, while AI risk management identifies and mitigates vulnerabilities across the AI lifecycle. Continuous monitoring safeguards against model drift, bias, and security threats, ensuring fairness and reliability post-deployment. Compliance reinforces adherence to global regulations, protecting fundamental rights and fostering stakeholder confidence. Together, these components create a resilient foundation that enables organizations to innovate responsibly while minimizing risks and maintaining public trust. The proposed approach promotes ethical use, identifying and mitigate risks, and aligns regulatory requirements, fostering trust and innovation.

Architecture Diagram



1. AI Governance:

Why it matters: Governance is the foundation for responsible AI adoption. It ensures fairness, transparency, accountability, and alignment with organizational values, reducing reputational and regulatory risks. AI Governance provides the foundation for ethical, transparent, and accountable AI practices, ensuring compliance and risk management throughout the AI lifecycle.

- **Legal & Regulatory Compliance as the Foundation**
  - Understand, document, and manage all applicable AI-related laws and regulations (e.g., EU AI Act, GDPR, ISO/IEC 42001, NIST AI RMF).
  - Integrate these requirements into organizational policies and governance frameworks.
- **Embedding Trustworthy AI Principles into Organizational culture**
  - Incorporate fairness, transparency, accountability, and safety into policies, standards, and practices.
- **Structured Risk Management & Executive Oversight**
  - Establish transparent processes for mapping, measuring, and mitigating AI risks aligned with organization risk appetite
  - Establish AI Governance oversight and decision-making body with clear roles and responsibilities, with executive leadership accountable for risk-related decisions.
  - Plan ongoing monitoring and periodic reviews of governance and risk processes
- **Lifecycle Management & System Inventory**
  - Maintain a comprehensive inventory of AI systems prioritized by risk level.
  - Implement AI onboarding, safe decommissioning and phasing-out procedures to prevent residual risks.
- **Inclusive & Diverse Decision-Making for Risk Mitigation**
  - Ensure that multidisciplinary and diverse teams are involved in decision-making to promote inclusivity and minimize bias
  - Define human-AI oversight roles for business-critical activities and accountability in system operations.
  - AI literacy training to be planned for all stakeholders

2. AI Risk Management Frameworks:

Why it matters: AI risk management is inherently complex due to the dynamic nature of AI systems and the diverse actors involved across the AI lifecycle. These actors—ranging from developers and data scientists to third-party vendors—introduce unique risks at different stages, often operating in varied contexts. This complexity makes it challenging to detect, measure, and manage risks effectively.

Third-party risks, in particular, require optimal assessment strategies, as external components (e.g., pre-trained models, APIs, datasets) can introduce vulnerabilities that are difficult to trace or control. Without a structured approach, organizations may struggle to maintain oversight and accountability, especially as AI systems scale and evolve.

- **The Role of AI Risk Management Frameworks**

  AI risk management frameworks are essential tools for proactively identifying, assessing, and mitigating risks such as:

  - Bias and fairness issues
  - Privacy violations
  - Safety and security vulnerabilities
  - Regulatory non-compliance

  These frameworks provide a **structured methodology** that supports the development of **trustworthy and responsible AI**, helping organizations:

  - Build stakeholder confidence
  - Aligning with evolving global regulations
  - Ensure ethical and secure AI deployment

- **Selecting the Right Framework**

  Choosing an appropriate AI risk management framework depends on several factors:

  - **Regulatory alignment:** Compatibility with local and international laws (e.g., EU AI Act, NIST AI RMF)
  - **Lifecycle coverage:** Ability to address risks across data collection, model development, deployment, and monitoring
  - **Ecosystem integration:** Fit with existing cybersecurity, governance, and compliance structures
  - **Scalability:** Applicability across diverse use cases and organizational sizes
  - **Industry relevance:** Tailored to sector-specific risks (e.g., healthcare, finance, energy)
  - **Quantifiability:** Support for measurable risk metrics and KPIs
  - **Stakeholder engagement:** Facilitation of cross-functional collaboration and decision-making

  By prioritizing risks and balancing innovation with mitigation strategies, these frameworks empower organizations to make informed, responsible decisions.

Examples of Prominent AI Risk Management Frameworks:

Standard	Approach	Focus	Benefits	Audience
NIST AI RMF (U.S.)	Voluntary guidance for managing AI risks	Trustworthy design, development, and use	Enhances trust, reduces risk, supports compliance	Enterprises, public sectors, cybersecurity experts
ISO/IEC 42001:2023	Certifiable global AI management system standard	Governance, lifecycle management	Certification, global standardization, streamlined compliance	All AI-developing or deploying organizations
EU AI Act	Mandatory risk-based regulation for AI in the EU	Safety, transparency, non-discrimination	Legal certainty, EU citizens' rights protection	Global AI providers targeting EU market
Google SAIF	Security-first AI lifecycle framework	Model integrity, threat management	Resilience, complements NIST RMF	Cloud-based AI developers, cybersecurity teams
China	Mandatory and voluntary standards for high-risk AI services	Generative AI, algorithmic control	Innovation-state balance, national security	China-based or operating AI services in China
Singapore	Voluntary principles-based framework with sector-specific mandates	Fairness, transparency, accountability	Practical tools, innovation-friendly	Asia-Pacific organizations, especially finance
OECD	Best practices and benchmark for Ethical AI	Human-centric values, robustness	Policy-level guidance, global consensus	International organizations, Policymakers, Private sector organizations, Civil society

- **Integrating Preemptive Cybersecurity**

Preemptive cybersecurity aligns with AI risk management by proactively mitigating threats that could compromise AI integrity, data confidentiality, and system availability. It supports key risk principles like resilience, transparency, and accountability.

- **Assessing the AI Attack Surface:** Identify vulnerable AI systems, data pipelines, and endpoints across the global attack surface grid (GASG).
- **Integrating AI-Powered Threat Detection:** Deploy tools like Darktrace and CrowdStrike Falcon for anomaly detection and autonomous response.
- **Applying Moving Target Defense (MTD):** Use platforms such as SentinelOne Singularity to dynamically shift system configurations and reduce predictability.
- **Deploying Cyber Deception Platforms:** Implement solutions like Acalvio ShadowPlex to create realistic decoys and gather attacker intelligence.
- **Aligning with Zero-Trust Architecture:** Ensure all access is explicitly verified and segmented using tools like Illumio for microsegmentation.

3. Continuous Monitoring & Testing:

Why it matters: Continuous Monitoring and Testing are essential to maintain the integrity and trustworthiness of AI systems after deployment. They help detect model drift, performance degradation, and emerging biases, ensuring outputs remain accurate, fair, and secure over time. By implementing robust monitoring processes and automated testing tools, organizations can proactively identify risks, prevent harm, and uphold compliance with ethical and regulatory standards—ultimately safeguarding user trust and enabling responsible innovation.Legal & Regulatory Compliance as the Foundation

- Create a baseline reference of model during production rollout to establish monitoring and alert thresholds.
- Document and monitor the data flow through systems, tracking transformations, movements, dependencies, and the end-to-end data journey.
- Document and test use cases for model accuracy, bias testing, fairness testing.
- Conducting bias, fairness, and explainability tests
- Red teaming and adversarial simulations tests
- Perform Fundamental rights impact assessment and Data protection impact assessment as applicable to protect personal data and sensitive personal data.
- Implement activity logging, anomaly detection, monitoring and incident management.

Recommended Tools & Platforms

- Arize AI – End-to-end AI observability with bias detection and explainability metrics.
- Fiddler AI – Fairness and bias monitoring for predictive and generative models.
- Deepchecks – Open-source ML testing and validation framework.
- Azure AI Foundry – Continuous monitoring integrated with Azure Monitor.
- WhyLabs – Privacy-first real-time guardrails for AI models.
- IBM AI Fairness 360 (AIF360)- Open-source toolkit with 70+ fairness metrics and 10+ bias mitigation algorithms for ethical AI development
- SHAP (SHapley Additive exPlanations),
- LIME (Local Interpretable Model-agnostic Explanations)- Model-agnostic tool that builds interpretable models locally around predictions to explain black-box decisions.
- Amazon SageMaker Model Monitor- Fully managed service for monitoring data quality, model drift, bias, and feature attribution in production ML models.
- Google Vertex AI- Unified platform for building, deploying, and monitoring ML and generative AI models with integrated MLOps and explainability tools.
- Mend AI Discovery Platform Discovers third-party AI models and inference providers during SCA (Software Composition Analysis) scans.
- Microsoft Learn – Threat Modeling AI/ML- Offers structured guidance to identify AI/ML dependencies and threat boundaries during security design reviews.
- CycloneDX- implement lightweight, security-focused software bill of materials specification by OWASP for tracking components, vulnerabilities, and dependencies across software and AI supply chains.

4. Compliance:

Why it matters: Compliance is a critical pillar of trustworthy AI, ensuring adherence to global regulations and ethical standards. It helps organizations avoid legal and financial risks while reinforcing accountability and transparency. By mapping controls to frameworks such as EU AI Act, GDPR, ISO/IEC 42001, and NIST AI RMF, and conducting regular audits, organizations can maintain trust with regulators, customers, and stakeholders, creating a secure foundation for responsible AI innovation.

- Document control objectives based on regulatory requirements aligned with organizational standards and policies
- Embed compliance checks at every stage—design, development, deployment, and monitoring—to proactively address risks.
- Implement Common Control Frameworks to manage multi-jurisdictional compliance and reduce duplication of efforts.
- Automate control validation to enhance visibility, reduce human error, and ensure real-time compliance.
- Schedule periodic audits and maintain detailed records of compliance activities for transparency and regulatory reporting

Recommended Best Practices

Establish Clear AI Governance Structures

Define roles, responsibilities, and accountability mechanisms across the AI lifecycle

Embed Ethical Principles from Design to Deployment

- Align AI systems with values like fairness, transparency, and human rights
- Collect external feedback on societal impacts and integrate adjudicated inputs into system design

Ensure Transparency and Explainability

- Make AI decisions understandable to stakeholders through model documentation and interpretability tools

Implement Robust Metadata Management

Track data lineage, model training steps, and decision logic to support traceability and compliance

Conduct Regular Risk Assessments

- Use AI risk management frameworks to identify and mitigate risks like bias, privacy violations, and safety concerns
- Conduct regular adversarial testing
- Implement Human-in-the-loop for high-risk decisions
- Continuous monitoring: Use automated tools for drift detection and compliance alerts

Promote Multistakeholder Engagement

Involve diverse voices—including civil society, domain experts, and impacted communities—in AI development

Adopt International Standards and Principles

Align with global frameworks like OECD AI Principles and ISO/IEC standards to ensure interoperability

Monitor AI Systems Continuously

Establish feedback loops and performance monitoring to detect drift, anomalies, and unintended consequences

Kill switch

Incorporate a kill switch, also known as an emergency shutdown mechanism, is a critical safety feature that allows immediate termination of an AI system to prevent harm or unpredictable behavior. It can be manual, ensuring human oversight and accountability, or automated, triggered by predefined risk thresholds for rapid containment.

Ensure Regulatory Compliance

- Stay updated with evolving laws like the EU AI Act and integrate compliance checks into development workflows
- Develop AI-specific policies integrated with enterprise governance frameworks like ISO/IEC 42001 and NIST AI RMF for global compliance
- Review Model cards and datasheets documentation to understand the model's purpose, capabilities, limitations, and ethical considerations

Foster a Culture of Responsible Innovation

- Encourage ethical awareness, training, and incentives for teams working on AI technologies

Metrics & KPIs for Trustworthy AI Security

To ensure AI systems are secure and resilient, organizations must track measurable indicators that reflect the performance of preemptive cybersecurity strategies. Key metrics include:

1. **Threat Detection Accuracy**
   - Percentage of true positives vs. false positives in AI-driven threat detection systems (e.g., Darktrace, CrowdStrike).
2. **Mean Time to Detect (MTTD) & Mean Time to Respond (MTTR)**
   - Time taken to identify and mitigate threats using autonomous response tools like SentinelOne.
3. **Deception Engagement Rate**
   - Number of attacker interactions with decoy assets (e.g., Acalvio ShadowPlex), indicating successful diversion from real systems.
4. **Attack Surface Reduction Score**
   - Quantitative measure of reduced exposure through moving target defense and microsegmentation (e.g., Illumio).
5. **Compliance Coverage Index**
   - Percentage of AI systems aligned with regulatory frameworks such as the EU AI Act and NIST AI RMF.
6. **Security Incident Recurrence Rate**
   - Frequency of repeated incidents post-mitigation, indicating resilience and learning capability of AI systems.
7. **Automated Patch Deployment Rate**
   - Speed and coverage of vulnerability remediation through AI-driven exposure management.

Tracking these KPIs helps organizations validate the trustworthiness of their AI systems and continuously improve their cybersecurity posture.

Conclusion

The journey toward building trustworthy AI is not a one-time effort but an ongoing commitment to ethics, security, and transparency. Securing AI systems demands a paradigm shift—from reactive cybersecurity to proactive, lifecycle-integrated governance. As AI becomes embedded in core business functions, organizations must evolve their risk management strategies to address:

- Emerging threats such as data poisoning, model inversion, and deepfake misuse.
- Regulatory convergence, with global standards like ISO/IEC 42001 and the EU AI Act setting the tone for cross-border security compliance.
- AI transparency, explainability and accountability, driven by public demand and ethical imperatives.
- AI sustainability, ensuring models are environmentally responsible and socially equitable.
- To build trustworthy AI, organizations must:
- Integrate AI governance with enterprise risk and compliance programs.
- Invest in explainability, fairness, and adversarial robustness.
- Foster cross-functional collaboration between cybersecurity, legal, data science, and executive leadership
- Embrace continuous learning and adaptation as AI technologies evolve.

Future success depends on AI systems that are not opaque "black boxes" but clear partners in decision-making. We must invest in advanced explainability techniques, ensuring that all stakeholders—from data scientists to end-users and regulators—can understand and interpret how conclusions are reached. This transparency is the bedrock of accountability.

Trust is inherently fragile. It can only be sustained by embedding robust privacy protections and state-of-the-art security measures into the very architecture of every algorithm and system. Adopting a "privacy-by-design" and "security-by-design" philosophy mitigates risk and demonstrates a non-negotiable commitment to protecting individual and institutional data.

Trust is not optional; it is the foundation upon which sustainable, AI-driven growth will stand. The opportunity now lies with executives and innovators worldwide to shape a world where AI accelerates human progress without compromising integrity or trust.

References

1. NIST AI Risk Management Framework (AI RMF)
2. ISO/IEC 42001:2023 – AI Management System Standard
3. EU Artificial Intelligence Act (2024)
4. CSA AI Controls Matrix for Trustworthy AI
5. Chambers & Partners: China AI Law Overview
6. White & Case: AI Watch – China Regulatory Tracker
7. Google SAIF Framework
8. Singapore AI Verify Foundation
9. OECD AI Principles

About the Author

Sunil Pandya
Principal Consultant

Sunil Pandya is a seasoned Cybersecurity professional with over 28 years of experience, specializing in Security Architecture reviews, AI Risk Assessment, Risk Management, Application Security, Information Security Audits, Compliance Management, Privacy, and Data Protection. He has managed multiple US clients across multiple industries, including Oil & Gas, BFSI, Healthcare, Retail, Software and Manufacturing, thus enabling clients to build resilient security architectures aligned with business goals and regulatory requirements. His specialization in managing Industry frameworks and standards such as ISO/IEC 27001, PCI DSS, NIST Cybersecurity Framework, SOC 2, and CMMC helped clients steered complex Cyber Governance, Risk, and Compliance (GRC) landscapes while supporting mission-critical business objectives. Sunil's expertise in managing Third party risk management program has benefited clients to manage third parties risk securely, thereby enhancing risk visibility and management supporting critical business objectives.

For more information, contact askus@infosys.com

Infosys®
Navigate your next