Deploying Trustworthy AI: An Illustrative Risk and Controls Guide

A guide outlining AI risks and corresponding control considerations for risk, technology, compliance, and legal leaders.

Foreword

The increasing prevalence of AI demands effective risk management through appropriate controls. Artificial intelligence is revolutionizing industries, transforming business structures, and even altering our way of life and work. It also holds the potential to significantly reshape the future of your organization.

The achievements enterprises can realize with AI appear limitless. 68% of CEOs consider AI a top investment priority, despite uncertain economic conditions, with the leading expected benefits being increased efficiency and productivity, an upskilled workforce, and increased enterprise innovation.

It's no surprise that these potential benefits make executives eager to integrate AI into their businesses and quickly realize its value. However, organizations can only fully leverage AI's potential by grounding their initiatives in trust, responsibly, ethically, and transparently managing its complexities and risks. As AI adoption scales and becomes more complex across business operations, navigating these complexities becomes increasingly challenging.

The stakes are also increasing for those responsible for ensuring the safe deployment and use of AI applications —including risk and compliance departments, cybersecurity and information security teams, data and privacy offices, legal teams, and internal audit. AI systems that lack proper governance and controls can impede returns on AI investments, lead to regulatory violations, result in data and intellectual property loss, or damage an organization's reputation.

Ultimately, grounding AI systems in practical and scalable risk management practices will be crucial for deploying AI boldly, quickly, and responsibly, thereby unlocking its transformative benefits. Developing a robust risk and controls guide for managing AI risks is a critical step in building an effective AI risk management program.

This guide offers a structured approach for organizations to begin identifying AI risks and designing appropriate control considerations to mitigate those risks. While other AI frameworks and standards identify risks at various stages of the AI lifecycle, this guide delves into the underlying control considerations.

This guide explores the underlying control activities, outlining suggested control considerations that businesses should consider when managing AI risks. It's important to note that this guide is intended as an informational resource to help organizations like yours appropriately manage AI-specific risks.

It provides illustrative examples of potential control considerations to address a broad, but not exhaustive, set of AI-specific risks. Deliberately focused solely on AI risks, it's designed to complement existing risk management frameworks that address general technology risks across areas like security, data privacy, and third-party risk management.

Therefore, you should first identify the control considerations from this guide that are relevant to your business and then carefully integrate them with your existing risk and control frameworks to ensure a comprehensive view of risks across your organization. We trust this guide will assist your organization in navigating the complex landscape of AI risks and driving innovation in a trusted manner.

How to put this guide into practice

Who should use this guide?

This guide is a useful resource for anyone leading or working on AI risk management and governance, including risk and compliance departments, cybersecurity and information security teams, data and privacy offices, legal teams, and internal audit.

Getting Started: Key Questions
Consider these initial questions:

- How do these risks and controls align with our existing risk categories?

This guide is based on the 10 pillars of the trusted AI framework and incorporates leading AI frameworks and regulations like ISO 42001, the NIST AI Risk Management Framework, and the EU AI Act. It's designed to work with your current risk categories, such as IT general controls and data governance controls.

- How do we apply these controls across the AI lifecycle?

Organizations should consider several factors when identifying and implementing controls throughout the AI lifecycle.

Organizations should consider several factors, including the nature and purpose of the AI system, the data flow, configuration, and logic that govern its operation, and the learning types and data sources it utilizes.

How do we adapt these control considerations to our specific organization and AI system?

Not every organization or AI system will need every control listed, and you may need additional controls based on your specific deployments. Users of this guide should consider existing and relevant risk and control frameworks, such as IT general controls, data governance controls, access and security controls, and API controls. Additionally, factors like the nature of the AI deployments (third-party, internally developed, proprietary data sources, etc.) and specific configurations or techniques (like retrieval augmented generation) can influence both risks and AI system operation. These considerations will help determine the specific risks present and the control activities required.

Trusted AI Pillars of Risk and Controls Guide

ng

About the Trusted AI framework

This AI Risk and Controls Guide aligns with our Trusted AI framework, which is based on a values-driven, human-centric, and trustworthy approach to AI development and deployment. The Trusted AI framework helps organizations develop and deploy AI solutions that address ethical concerns and comply with regulatory standards.

Organized around the 10 pillars of the Trusted AI framework, this guide provides an initial list of AI risks, each accompanied by a set of control considerations that organizations can use when developing their control catalogues.

Accountability

10 pillars of the Trusted AI framework

Human oversight and responsibility should be integrated throughout the AI lifecycle to manage risk and ensure compliance with applicable laws and regulations.

- **Accountability**
- Data Integrity
- Explainability
- Fairness
- Privacy
- Reliability
- Safety
- Security
- Sustainability
- Transparency

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
AI Performance Erodes Over Time	AI system errors are improperly resolved	Errors in the AI system remain undetected, detected late, or not acted upon timely, resulting in unauthorized changes, system unavailability, security breaches, data loss, or other incidents.	Perform periodic assessments of the AI system's outputs to ensure they align with original business and ethical requirements. Any discrepancies are documented and addressed promptly to ensure the AI exhibits intended behavior and meets business objectives.
			Thresholds are configured for AI system performance monitoring to ensure ongoing oversight of AI accuracy and performance. In the event a threshold is exceeded, remediation and/or maintenance activities are performed on a timely basis by appropriate personnel to remediate the issue.
Bypassing AI Risk Management	Inadequate AI governance	Lack of control over AI system/system modifications, deployment, and inappropriate access (including authentication and authorization) may lead to incidents, unauthorized usage, and data loss, resulting in operational, integrity, financial, or reputational damage.	High-risk AI system providers that use rules-based AI techniques adhere to established data governance and management practices to ensure personal data is lawfully obtained, processed, and minimalized in the AI's lifecycle.
			Develop and maintain exit strategies and contingency plans for AI systems to facilitate the seamless migration of systems to different providers, ensuring a prepared and effective response to any unforeseen disruptions or changes to third-party relationships.
			The organization maintains an up-to-date and comprehensive inventory of AI systems and use cases to ensure continued accountability and appropriate management of AI systems.
	Inappropriate modification to the AI system	Inappropriate modifications are made to the AI system which could lead to errors and vulnerabilities being introduced to the system.	Develop approved policies and procedures for AI system governance to guide algorithm selection for fit for purpose and alignment with strategic and business requirements. Ensure training and awareness to the relevant stakeholders to enforce compliance.

Data Integrity

10 pillars of the Trusted AI framework

Data used in AI solutions should be collected in compliance with relevant laws and regulations and evaluated for accuracy, completeness, appropriateness, and quality to ensure reliable decisions.

- Accountability
- Data Integrity
- Explainability
- Fairness
- Privacy
- Reliability
- Safety
- Security
- Sustainability
- Transparency

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Data Integrity Issues in AI Systems	Inadequate data governance	Insufficient data governance for learning, training, or testing data can result in biased, inaccurate, or unreliable outputs and ineffective AI systems.	Policies and procedures define data management requirements, covering the collection, analysis, labeling, storage, and filtering of data, along with criteria for using training and test datasets. These ensure compliance with regulatory requirements and organizational values. Regular training and awareness campaigns are conducted for relevant stakeholders to promote compliance. The policies and procedures are reviewed and updated periodically as needed.
			Conduct quality checks and comprehensive measures, like data gap analysis, to ensure the training, validation, and testing data are high-quality, accurate, and complete. Any discrepancies or deficiencies are promptly identified, documented, and resolved.
	Insufficient controls for data interactions	Inadequate methods for facilitating and controlling data interactions (e.g., transfers) between AI systems and data sources or other entities (e.g., applications, APIs) can lead to data corruption or loss, system misuse, or unauthorized access.	When making changes to an AI system, the training and testing data is reviewed for relevance and accuracy in relation to the changes. Additional data is incorporated as needed to train and test any new system capabilities or features.

Explainability

10 pillars of the Trusted AI framework

AI solutions should be developed and deployed in a manner that explains how and why they arrive at a given conclusion.

- Accountability
- Data Integrity
- Explainability
- Fairness
- Privacy
- Reliability
- Safety
- Security
- Sustainability
- Transparency

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Explainability Not Embedded in the Design	Failure to understand AI logic	The logic within the AI system is not fully understood or accessible to the organization, impacting business operations and resulting in financial loss or reputational damage.	During strategy and development, maintain clear, comprehensive documentation (e.g., model cards) of AI systems, including narratives, flowcharts, and data flows, to ensure explainability and transparency. To maintain the documentation's accuracy, regularly review and update the documentation to reflect any changes to the systems or datasets.
	Lack of explainable AI solution environment	Lack of understanding of AI-related IT and data components by operational IT support can undermine the effectiveness of controls, including security, software licenses, IT operations, and business continuity.	AI system impacts on subsequent business operations are clearly communicated to and comprehended by all relevant stakeholders to ensure understanding of impacts on upstream and/or downstream processes.
			Configure AI activity monitoring jobs to trace AI activities, retaining logs for necessary periods to support comprehensive audit trails.
			Policies and procedures define guidelines for explainability, minimal data usage, simplicity in system, causation analysis, and tracking methods. Training and awareness campaigns are performed for relevant stakeholders to enforce compliance. The policies and procedures are reviewed and updated, as needed, periodically.

Fairness

10 pillars of the Trusted AI framework

AI solutions should be designed to minimize or eliminate bias against individuals, communities, and groups.

- Accountability
- Data Integrity
- Explainability
- Fairness
- Privacy
- Reliability
- Safety
- Security
- Sustainability
- Transparency

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Harmful Bias in AI Systems	AI systems are inaccessible to all groups	AI systems that are designed and developed without considering the principles of accessibility may limit the user base and exclude certain communities, leading to noncompliance with legal standards, and reducing the overall usability and inclusiveness of the technology.	Conduct extensive user testing with a diverse range of participants, including those with various disabilities, to identify and address potential barriers in using the AI system. Any barriers are addressed prior to launch to ensure the AI systems is more accessible and inclusive.
			Training for all team members who create and develop AI systems is periodically conducted to ensure team members understand the diverse needs of different user groups and practical methods for implementing accessibility in AI.
	Misalignment to the organization's cultural and ethical values	Misalignment of AI systems and decision-making processes with the organization's cultural and ethical values may lead to reputational damage, loss of trust, and increased accountability issues for the organization.	Diverse stakeholders are consulted during novel strategy and perform model testing, providing feedback. Feedback is gathered throughout the model development lifecycle to determine the need for additional testing, recalibration, or training data.

Privacy

10 pillars of the Trusted AI framework

AI solutions should be designed to comply with applicable privacy and data protection laws and regulations.

- Accountability
- Data Integrity
- Explainability
- Fairness
- Privacy
- Reliability
- Safety
- Security
- Sustainability
- Transparency

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Privacy Breaches from AI Solutions	Data Subject Access Privacy	Failure to have the necessary operational infrastructure for individuals to exercise their data subject access rights promptly can lead to decreased consumer trust, regulatory non-compliance, or financial losses.	Implement awareness programs to educate data subjects about their rights concerning AI technologies, explaining how to exercise these rights and the impact of AI decision-making on their personal data.
	Privacy Directives and Regulatory Non-Compliance	Non-compliance with organizational directives and/or regulations regarding data subject processing can result in financial penalties, market losses, and reputational damage.	Regular reviews of the input, training data, and output used by AI solutions are conducted to ensure data usage complies with the organization's data privacy policies and relevant regulatory requirements.
			Monitor and assess any changes to the purpose of AI systems, ensuring that any new use of personal data is fair, lawful, and transparent.
	Data Breach Leading to Privacy Violation	Data breaches can lead to unauthorized access or disclosure of personal, official, confidential, and highly confidential data, potentially compromising user or organizational privacy, violating data protection laws, damaging reputations, or causing financial harm.	A strong oversight system is in place, encompassing ethical reviews, regular audits of data protection measures, impact assessments, and compliance checks, especially when sensitive personal data is used for AI training or production.
			Document the rationale and obtain explicit approval when acquiring data for training. Implement special precautions for AI use cases that may directly or indirectly affect vulnerable individuals or have implications for safety or rights.
			Where appropriate for the model and use case, a controlled amount of randomness (i.e., differential privacy) is added to training and prompt data to protect data privacy.

Reliability

10 pillars of the Trusted AI framework

AI solutions should consistently operate as intended, within their defined scope, and at the desired level of accuracy.

- Accountability
- Data Integrity
- Explainability
- Fairness
- Privacy
- Reliability
- Safety
- Security
- Sustainability
- Transparency

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Insufficient Support and Maintenance	Inadequate monitoring of AI operations	Lack of audit and effective monitoring capabilities in AI system operations may impact the ability to monitor system performance and respond to incidents timely.	As needed, develop novel risk tracking approaches for settings where AI risks are difficult to assess with current measurement techniques, ensuring comprehensive risk management even when standard metrics are unavailable.
			Automated correction, fallback, or stop/loss mechanisms are implemented in the AI system's design to ensure the AI system corrects, or when necessary, halts unintended behavior. Humans are alerted and the issue(s) are quickly remediated.
			Continuous evaluation and necessary recalibration of system performance, including training data and algorithms, features against established incident alerts to uphold the system's accuracy and reliability, adhering to predefined thresholds.
			Regularly identify and track both existing and emergent AI risks, ensuring responsive adaptation to real-world performance and contexts.
			User-friendly and accessible mechanisms are in place for employees, users, and other stakeholders to report errors, biases, or vulnerabilities in the AI system. End-user reports are collected, reviewed, tested, and remediated as needed to validate that the system is performing consistently. Residual risks and potential impacts are documented.

Safety

10 pillars of the Trusted AI framework

AI solutions should be designed and implemented to safeguard against harm to people, businesses, and property.

- Accountability
- Data Integrity
- Explainability
- Fairness
- Privacy
- Reliability
- Safety
- Security
- Sustainability
- Transparency

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Insufficient Response to AI-Generated Safety Threats	AI system errors are not resolved effectively.	AI system errors go undetected, are detected late, or are not addressed promptly, potentially leading to unauthorized changes, system unavailability, security breaches, data loss, or other incidents.	A selection of AI-driven threat response decisions is regularly reviewed to ensure they are ethical, responsible, and aligned with business objectives. This review is conducted by authorized personnel, and documentation of the reviews is retained.
			Anomaly detection systems are implemented to identify suspicious activities within a system, such as prompt injection, data poisoning, abuse, evasion, privacy attacks, increased traffic on a communication channel, and indirect prompt injection.
	Production of harmful or unreliable content (e.g., hallucinations)	Generative AI outputs can be harmful, offensive, biased, or misleading, potentially negatively impacting the organization, communities, or society.	Feedback loops are integrated into the AI system to continuously validate and verify its outputs, ensuring the AI does not generate harmful, inaccurate, or unintended content (including hallucinations) that deviates from its intended use, business objectives, or defined parameters.
Threat to Humans	Insufficient human intervention	A lack of awareness of AI use by humans, coupled with insufficient oversight, can prevent the ability to override or correct decisions made by AI systems.	Establish approved policies and procedures for disclosing AI-generated or manipulated content (e.g., deepfakes) that resembles existing persons, objects, places, or events. Provide training and awareness to relevant stakeholders to ensure compliance.
			Human moderators respond to reports of AI misuse or inaccurate outputs/decisions, ensuring AI system decisions are properly reviewed and addressed. Any necessary corrective action is taken promptly.

Security

10 pillars of the Trusted AI framework

Strong and resilient practices should be implemented to protect AI solutions from malicious actors, misinformation, or adverse events.

- Accountability
- Data Integrity
- Explainability
- Fairness
- Privacy
- Reliability
- Safety
- Security
- Sustainability
- Transparency

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
AI Security	Adversarial attacks	Adversarial attacks exploiting models, data sets, or algorithms may result in unauthorized access to confidential data, model tampering, data corruption or loss, misuse, inappropriate access, or noncompliance with underlying regulations.	Design and develop AI systems with robust mechanisms in place to effectively limit outputs to essential information only. Utilize techniques such as data anonymization or differential privacy to safeguard sensitive training data, as well as to protect system and algorithm details from potential attackers and data leakage.
			Implement training data set expansion techniques as part of data cleaning process to ensure the performance and robustness of algorithms/systems and their resilience to adversarial and poisoning attacks.
			Pre-process input data to obfuscate AI system functionality, safeguarding against manipulation and protecting against potential attacks.
			Perform penetration tests and/or "Red Team" exercises for the AI system and its environment to identify potential vulnerabilities. Any identified exposures are promptly reviewed and addressed to ensure the system operates as expected.
	Copyright infringement	Intellectual property (IP) code is not accessible to the organization or is not adequately protected from IP loss/theft, resulting in an inability to maintain effective AI systems in an efficient manner.	An in-house repository containing relevant IP such as data, code, models, and 'learning data' is established and accessible, with regular backups and robust security measures including encryption to ensure IP is accessible and protected.
			IP audits are periodically conducted to ensure that all AI-related code and documentation are accounted for, properly documented, and compliant with licensing agreements.

Sustainability

10 pillars of the Trusted AI framework

AI solutions should be designed for energy efficiency, minimizing carbon emissions and promoting a cleaner environment.

- Accountability
- Data Integrity
- Explainability
- Fairness
- Privacy
- Reliability
- Safety
- Security
- Sustainability
- Transparency

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Overall Risk Related to AI Sustainability	Lack of focus on sustainable AI development	Failure to consider environmental impact in AI system strategy and design can lead to energy-inefficient systems.	During AI strategy and development, establish clear sustainability goals for the AI system, aligned with organizational standards, and develop a strategy demonstrating how the system will achieve these goals throughout its lifecycle.
	Lack of focus on sustainable AI implementation and usage	Insufficient sustainable implementation, usage, and monitoring practices can lead to decreased system sustainability and misalignment with organizational ESG commitments.	Integrate environmental impact indicators and real-time monitoring throughout the AI system lifecycle to ensure energy consumption, system efficiency, and emissions comply with applicable environmental standards and company strategies. Identified gaps or areas for improvement are promptly addressed.

Transparency

10 pillars of the Trusted AI framework

AI solutions should include responsible disclosure to provide stakeholders with a clear understanding of what is happening in each solution across the AI lifecycle.

Explore examples›

- Accountability
- Data Integrity
- Explainability
- Fairness
- Privacy
- Reliability
- Safety
- Security
- Sustainability
- Transparency

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Differentiating Human-Created from AI-Generated Content	Opacity of AI systems	Insufficient AI system transparency can diminish accountability, raise ethical concerns, and erode consumer trust.	Demonstrate the AI system's validity and reliability, and document the limitations of its generalizability beyond the tested conditions to ensure transparency regarding its applicability and effectiveness.
			Identify and document any remaining potential negative risks for both downstream recipients and end-users, providing a comprehensive overview of unmitigated risks associated with the AI system.
Insufficient Transparency in AI and Data Usage	Insufficient explainability of AI solutions	Insufficient understanding of AI-related IT and data components by operational IT support can weaken the effectiveness of controls related to security, software licensing, IT operations, and business continuity.	Document the test sets, metrics, and tools used during Test, Evaluation, Validation, and Verification (TEVV) processes to establish a transparent and reproducible framework for evaluating the AI system's performance and reliability.
	User transparency	A lack of transparency in AI system development and use can lead to diminished accountability, making it challenging to understand the reasoning behind system behavior, raising ethical concerns, and eroding consumer trust.	AI-generated or manipulated content is labeled or watermarked (e.g., using CP2A) to ensure transparency and establish the origin of AI-created content.
			For each output generated by the AI system, users are explicitly informed of the potential for inaccuracies and strongly advised to critically review the AI system's results.
			Before each use, AI system users are notified of data collection and/or processing for personalization and recommendation purposes. Users are given the option to opt out of these services to ensure transparency and user choice.
			Users or individuals affected by emotion recognition or biometric categorization AI systems are informed of the system's operation before it is used.

Designing controls for your AI systems

The control considerations in this guide provide a basis for developing custom control descriptions for your AI deployments. We've also included some example control implementations to help you get started. Please don't hesitate to contact our team if you have any questions.

Pillar	Risk Category	Illustrative Control Consideration	Example Control Implementation Description
Accountability	AI Performance Degradation Over Time	Regularly assess the AI system's outputs to ensure they align with original business and ethical requirements. Document and promptly address any discrepancies to ensure the AI performs as intended and meets business objectives.	The AI system owner conducts a quarterly review of a sample of the AI system's outputs against established key performance indicators (KPIs) and key risk indicators (KRIs) to verify expected performance. Any discrepancies or variances exceeding established thresholds are investigated and resolved within five business days. In the event of a major discrepancy, the system is immediately removed from production.
Fairness	Harmful Bias in AI Systems	Regular training is provided to all team members involved in creating and developing AI systems to ensure they understand the diverse needs of different user groups and practical methods for implementing accessibility in AI.	All team members involved in creating and developing AI systems are required to complete the "AI Fairness and Accessibility" training course annually. A post-training assessment with a minimum passing score of 85% is required.
Data Integrity	Insufficient Data Integrity in AI Systems	During the change management process for an AI system, the training and testing data are evaluated for relevance and accuracy in relation to the change. Additional data is incorporated as needed to train and test new system capabilities or features.	When modifying an AI system, perform regression or error rate testing as outlined in the Change Management policy. Any issues identified during testing that are classified as higher than "low" severity must be resolved before deployment to production.
Transparency	Insufficient Transparency in AI and Data Usage	For every output the AI system generates, users are clearly told that the results might be inaccurate and are strongly encouraged to carefully review them.	Each AI-generated output includes a disclaimer at the beginning, stating: "Outputs generated by this system may contain inaccurate, incomplete, or outdated information. Therefore, professional judgment should be exercised before relying on them."
		Before each use, users of the AI system are informed about any data collection and/or processing done for personalization and recommendation purposes. Users are given the option to opt out of these services to ensure transparency and user control.	Before each use of the AI system, a pop-up window appears stating, "I consent to the collection of my data through the use of this system," and blocks access to the system. Users must click "I acknowledge" to consent and gain access.

VECTOR RESEARCH PARTNERS

Working With Vector

There's a wealth of potential for PE firms to unlock via digital marketing. By pulling the right levers across your portfolio companies, more prospects and customers move through your funnel, revenue increases, and EBITDA grows. Ultimately, your investment in a business realizes its peak ROI.

Vector specializes in identifying and seizing opportunities in the digital marketing space. **We work with investors at every stage, including pre-LOI, due diligence, and post-traction, to accelerate growth plans, measurement, and value creation.**

Get the AI in Marketing report: Unlocking Scalable Strategies to Maximize EBITDA Growth

Secure Your Competitive Edge—Before Your Competitors Do This research is **not publicly available** and is reserved exclusively for a select group of firms and operators looking to leverage AI for real, measurable impact. **Access is limited**, ensuring that those who act now gain a strategic advantage in the market.

CONTACT US

or email our founder
(robert@vectorresearchpartners.com)