Deploying trustworthy AI: An illustrative risk and controls guide

The guide to AI risks and underlying control considerations for risk, technology, compliance, and legal leaders

visit.kpmg.us/TrustedAI

Foreword

AI is on the rise. Controls can help manage the risks.

Artificial intelligence (AI) is revolutionizing sectors, 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 accomplishments enterprises can achieve with AI are seemingly limitless. According to the KPMG 2024 CEO Outlook, 64 percent of global CEOs say AI is a top investment priority, despite uncertain economic conditions with top expected benefits being increased efficiency and productivity, an upskilled workforce, and increased enterprise innovation.[1]

Unsurprisingly, such benefits make executives eager to integrate AI into their businesses and accelerate the value it delivers. **But organizations can only harness AI's full potential once they ground such initiatives in trust, managing its complexities and risks in a responsible, ethical, and transparent manner.** As the scale and complexity of AI adoption advances across business operations, such complexities become increasingly difficult to navigate.

The stakes are also rising for those tasked with ensuring the safe deployment and use of AI applications—risk and compliance departments, cyber and information security teams, data and privacy offices, legal teams, and internal audit. AI systems that are not properly governed and controlled can hinder returns on AI investments, lead to regulatory compliance violations, result in data and IP loss, or damage the organization's reputation.

Ultimately, it will be key to ground AI systems in pragmatic and scalable risk management practices to **deploy AI boldly, quickly, and responsibly—unlocking its transformative benefits.** Establishing a robust risk and controls guide for managing AI risks is a critical step in developing an AI risk management program.

KPMG has published a first-of-its kind illustrative AI risk and controls consideration guide. The guide—aligned to the KPMG Trusted AI framework—provides a structured approach for organizations to begin identifying AI risks and designing proportionate control considerations to mitigate those risks. While existing AI frameworks and standards identify risks at different stages of the AI lifecycle, this guide delves into the underlying control activities, outlining suggestive control considerations businesses should contemplate for managing AI risks.

Please note: This guide is meant to be an informative aid for helping organizations like yours appropriately manage AI-specific risks. It provides illustrative examples of potential control considerations to address a large, though not complete, set of AI-specific risks. Intentionally focused solely on AI risks, it is designed to complement existing risk management frameworks that address general technology risks across domains such as security, data privacy, and third-party risk management. As such, you should first identify control considerations from this guide that are relevant to your business, and then carefully integrate them with your existing risk and control frameworks to help ensure a thorough view of risks across your organization.

We hope that this guide helps your organization begin to navigate the complex landscape of AI risks and drive innovation in a trusted manner.

—Bryan McGowan
Global Trusted AI Leader, KPMG International

[1] KPMG 2024 US CEO Outlook

How to put this guide into practice

Who is this guide for?

This guide can serve as a resource for any anyone leading or involved in AI risk management and governance, including risk and compliance departments, cyber and information security teams, data and privacy offices, legal teams, and internal audit.

Start with these questions.

How does the risk and related set of control considerations align to existing risk taxonomies in my business?

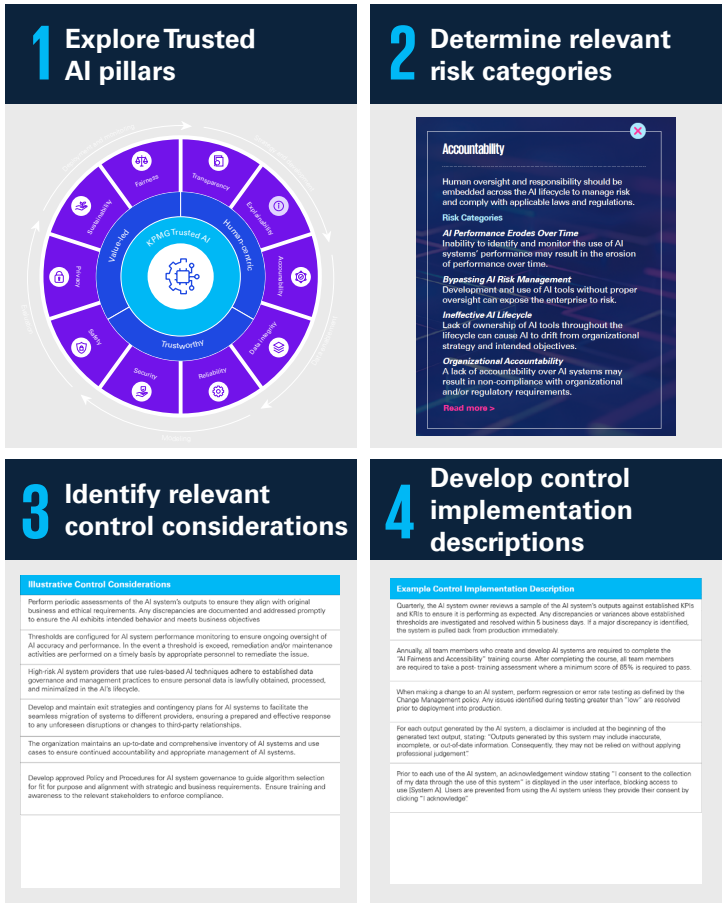
This guide is aligned to the 10 pillars of the KPMG Trusted AI framework, and was developed around leading AI frameworks and regulations, such as ISO 42001, the National Institute of Standards and Technology (NIST) AI risk management framework, and the EU AI Act. This is meant to be complementary to existing risk taxonomies within your organization, such as IT general controls and data governance controls.

How should the control considerations be applied across the AI lifecycle?

To identify and implement control considerations across the AI lifecycle, there are several factors organizations should consider, such as understanding the nature and use of the AI system; data flow, configuration, and logic that influences operation; and learning types and data sources used.

How can we design and implement the control considerations to fit our own organization and AI system?

Not every organization or AI system may need to implement every control or there may be additional controls based on your specific deployments. Users of this guide should consider existing risk and control taxonomies in place and relevant to AI, such as IT general controls, data governance controls, access and security controls, application programming interface (API) controls, etc. Additionally, users should consider, for example, the nature of the AI deployments, and whether AI systems are third party, internally developed, leverage proprietary data sources, or have other configuration or techniques in play (such as retrieval augmented generation) which may influence risks and AI system operation. These considerations help to inform what risks may be present and, therefore, control activities required.



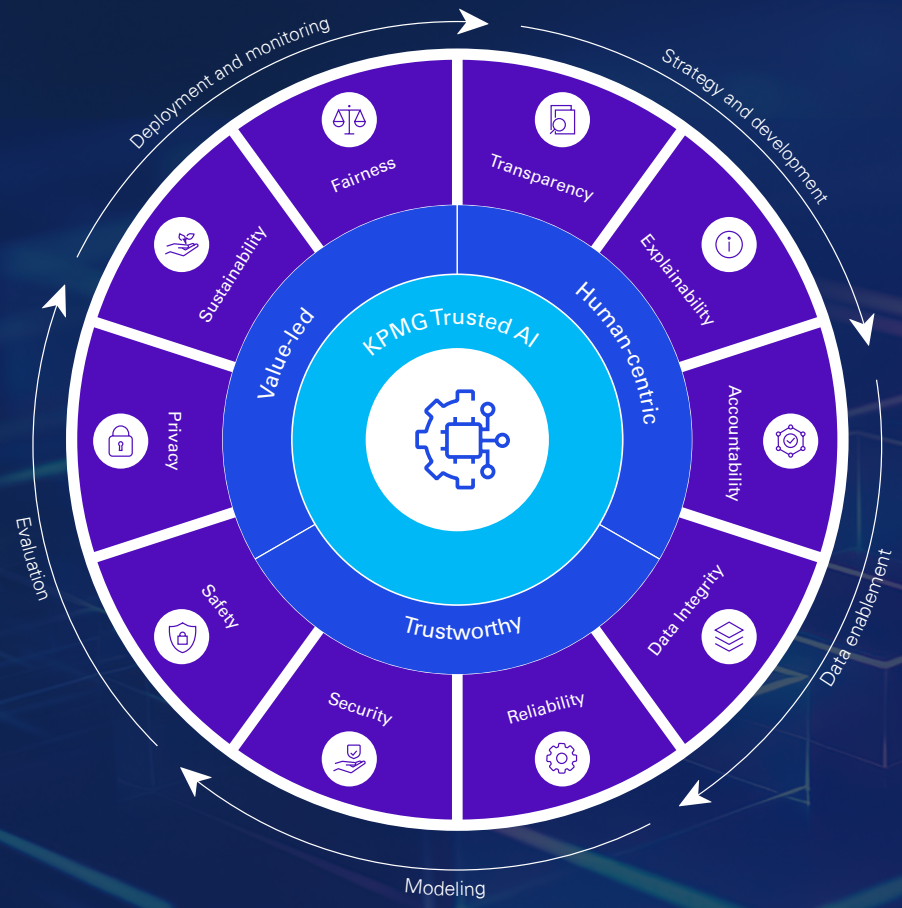
Get started by exploring the KPMG Trusted AI framework

Trusted AI pillars of risk and controls guide

About the KPMG Trusted AI framework

The AI Risk and Controls Guide is aligned to our Trusted AI framework, which is rooted in a values-driven, human-centric, and trustworthy approach to AI development and deployment. The Trusted AI framework helps our own firm, and our clients, develop and deploy AI solutions that address ethical concerns and comply with regulatory standards.

Organized under the 10 pillars of the KPMG Trusted AI framework, this guide outlines an initial inventory of AI risks, each with a set of control considerations that organizations can leverage as they build out their control catalogues.



Transparency

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

Risk Categories

Distinguishing Human vs. AI Content
Failure to distinguish between human-generated and AI-generated content can lead to misinformation, confusion, compromise the integrity of information sources, and/or lead to consumer mis-trust.

Lack of Transparency in AI and Data Usage
Lack of transparency in AI and data usage can undermine user privacy, cause unaccountability for errors or harm, and the potential to violate ethical standards, thereby eroding public trust in such technologies.

Explainability

AI solutions should be developed and delivered in a way that answers the questions of how and why a conclusion was drawn from the solution.

Risk Categories

Explainability Not Embedded in the Design
AI systems are not designed, developed, or implemented with explainability principles in mind—when explainability is not considered at the start of the AI lifecycle, the result is solutions with profound downstream implications on system use, trust, and performance.

Lack of Meaningful Human Review or Intervention
Humans need to be aware of the use of AI, provide oversight, and be able to override decisions made by AI systems.

Accountability

Human oversight and responsibility should be embedded across the AI lifecycle to manage risk and comply with applicable laws and regulations.

Risk Categories

AI Performance Erodes Over Time
Inability to identify and monitor the use of AI systems' performance may result in the erosion of performance over time.

Bypassing AI Risk Management
Development and use of AI tools without proper oversight can expose the enterprise to risk.

Ineffective AI Lifecycle
Lack of ownership of AI tools throughout the lifecycle can cause AI to drift from organizational strategy and intended objectives.

Organizational Accountability
A lack of accountability over AI systems may result in noncompliance with organizational and/or regulatory requirements.

Data Integrity

Data used in AI solutions should be acquired in compliance with applicable laws and regulations and assessed for accuracy, completeness, appropriateness, and quality to drive trusted decisions.

Risk Category

Lack of Data Integrity in AI Systems
Compromised data integrity in AI systems may lead to inaccurate or unreliable outputs, undermining decision-making processes and potentially causing operational and reputational harm.

Reliability

AI solutions should consistently operate in accordance with their intended purpose and scope and at the desired level of precision.

Risk Categories

Insufficient Support and Maintenance
Insufficient operational support and maintenance leads to an ineffective AI solution, or to AI solutions becoming ineffective over time, and/or poor decision-making during major incidents.

Insufficient Understanding of AI Architecture
IT and data components of the overall AI environment, including changes to IT infrastructure, AI models, algorithms, and data, may not be fully understood by the operational IT support at the organization, undermining the reliability and robustness of the AI systems and potentially disrupting the continuity and smooth operation of the overall business.

Security

Robust and resilient practices should be implemented to safeguard AI solutions against bad actors, misinformation, or adverse events.

Risk Categories

AI Security
Failure to embed security principles in the AI model architecture and AI development processes can lead to significant security vulnerabilities and/or unauthorized disclosure of information (including Personal Data and Intellectual Property).

Unsafe Prompt Engineering
Prompt engineering may result in unintended consequences including, but not limited to, leaks of strictly confidential information/Personal Data, creation of malicious code, social engineering, or system outages.

Safety

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

Risk Categories

Inadequate Response to AI-Generated Safety Threats
Organizational procedures and systems are insufficiently robust to quickly and effectively respond to safety threats generated or exacerbated by AI systems, leading to potential harm or hazardous situations.

Threat to Humans
AI systems may be leveraged or misused as a threat to human life and well-being, resulting in potential harm or adverse effects on society.

Privacy

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

Risk Category

Privacy Violations from AI Solutions
Failure to comply with Organization Privacy Directives and Procedures (e.g., inappropriate collection/disclosure of personal data) may result in a loss of consumer trust, regulatory non-compliance, or cause financial harm.

Sustainability

AI solutions should be designed to be energy efficient, reduce carbon emissions, and support a cleaner environment.

Risk Category

Overarching Risk Associated with AI Sustainability
Lack of a sustainable AI strategy, efficient energy consumption, and understanding of e-waste generation may result in negative environmental, ethical, societal, and operational impacts.

Fairness

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

Risk Category

Harmful Bias in AI Systems
Harmful bias in AI systems can perpetuate societal inequalities or discriminatory outcomes, which may lead to the erosion of public trust and cause legal, reputational, or financial loss.

Accountability

10 pillars of the Trusted AI framework

Click each pillar below to explore

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

Human oversight and responsibility should be embedded across the AI lifecycle to manage risk and comply with applicable laws and regulations.

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.

Accountability

10 pillars of the Trusted AI framework

Click each pillar below to explore

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

Human oversight and responsibility should be embedded across the AI lifecycle to manage risk and comply with applicable laws and regulations.

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Ineffective AI Lifecycle	Insufficient AI development program	The AI development program lacks robust management methodologies, including comprehensive testing and predefined metrics for performance, accuracy, and fairness.	During strategy and development, a clear business case for the AI system is developed, formally approved by relevant stakeholders, enacted, and maintained to ensure alignment to the organization's strategy.
Organizational Accountability	Lack of ethics governance	Employees, customers, and communities are not aware or are not acting with integrity to support ethical and trustworthy data use and use of the AI system.	Conduct regular engagement with AI stakeholders, facilitating the integration of feedback regarding the AI system's impacts.
			Establish an AI ethics code of conduct that embeds shared values and principles relevant to internal and external stakeholders to support ethical and trustworthy data use and use of the AI system. The code of conduct is reviewed and updated at least annually.
	Noncompliance with internal or external requirements	Noncompliance with internal or external requirements over internal control and compliance may lead to ineffective systems, or regulatory or market repercussions.	AI-related documentation is retained for 10 years (or following applicable laws and regulatory guidance) after market launch or service initiation to ensure compliance with relevant regulations.
			An AI system undergoes a risk reassessment periodically or when triggered by significant events to ensure compliance with regulatory requirements and adherence to organization policies. Noncompliant high-risk AI providers are formally justified, with equivalent or superior alternative systems in place.

Accountability

10 pillars of the Trusted AI framework

Click each pillar below to explore

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

Human oversight and responsibility should be embedded across the AI lifecycle to manage risk and comply with applicable laws and regulations.

Risk category	Risk consideration	Risk description	Illustrative control considerations
Organizational Accountability	Noncompliance with internal or external requirements	Noncompliance with internal or external requirements over internal control and compliance may lead to ineffective systems, or regulatory or market repercussions.	An organizational AI strategy is enacted to establish consistency in AI development and use across the organization. The strategy is reviewed and updated periodically to ensure continued alignment with business goals and risk tolerance.
			Categorize AI systems by risk levels at intake according to a defined AI Risk Tiering methodology. The AI Governance Committee conducts regular reviews of the methodology to ensure alignment with organizational standards and regulatory requirements.
			Policies and procedures define how to design, develop, and manage the risk of AI systems to ensure compliance with standards and internal controls. Training and awareness campaigns are performed for relevant stakeholders to enforce compliance. The policies and procedures are reviewed and updated, as needed, periodically.
			During strategy and development, AI system impact assessments, privacy impact assessments, and data protection impact assessments are performed to ensure proactive identification of risks, implementation of mitigations, and ongoing compliance with applicable regulations. Identified risks are documented and mitigations are agreed upon with the development team.
			Identify and document internal risk controls for all components of the AI system, including third-party technologies, to ensure comprehensive oversight and mitigation of potential risks throughout the AI lifecycle.
			Implement an accountability matrix (e.g., RACI) to define accountability of actions across relevant business functions. The matrix is reviewed and updated regularly.
			Implement an enterprise AI governance framework across the organization to ensure consistent guidance and oversight over AI system development across all relevant functions. The framework is reviewed periodically.

Data integrity

10 pillars of the Trusted AI framework

Click each pillar below to explore

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

Data used in AI solutions should be acquired in compliance with applicable laws and regulations and assessed for accuracy, completeness, appropriateness, and quality to drive trusted decisions.

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Lack of Data Integrity in AI Systems	Insufficient data governance	Lack of adequate data governance over learning, training, or testing data may lead to biased, inaccurate, or unreliable outputs and ineffective AI systems.	Policies and procedures define data management requirements, including the collection, analysis, labelling, storage, and filtration of data as well as decision-making criteria for using training and test data sets to ensure compliance with regulatory requirements and organization values. Training and awareness campaigns are performed for relevant stakeholders to enforce compliance. The policies and procedures are reviewed and updated, as needed, periodically.
			Perform quality checks and comprehensive measures, such as data gap analysis, to ensure the quality, accuracy, and completeness of training, validation, and testing data. Any discrepancies or shortcomings are promptly identified, documented, and addressed.
	Inadequate methods to facilitate and control data interactions	Lack of appropriate methods to facilitate and control data interactions (e.g., transfers) between the AI systems and data sources or other entities (e.g., applications, APIs) may result in data corruption or loss, system misuse, or inappropriate access.	During the change management process for an AI system, the training and testing data used is evaluated for relevancy and accuracy with the change. As needed, additional data is introduced to train and test new system capabilities or features.

Explainability

10 pillars of the Trusted AI framework

Click each pillar below to explore

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

AI solutions should be developed and delivered in a way that answers the questions of how and why a conclusion was drawn from the solution.

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.

Explainability

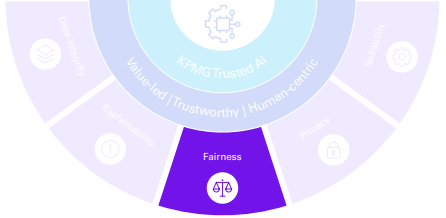
10 pillars of the Trusted AI framework

Click each pillar below to explore

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

AI solutions should be developed and delivered in a way that answers the questions of how and why a conclusion was drawn from the solution.

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Explainability Not Embedded in the Design	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.	Document the provenance of all training, validation, and testing data utilized during the AI system's lifecycle. An appointed authority carries out regular re-evaluations of data origin, ensuring documentation is current.
			Set up and upkeep a Configuration Management Database (CMDB) that catalogs all Configuration Items (CIs) to fully map out IT, data, and governance structures, ensuring clarity in data classifications, asset markings, and data flow diagrams.
Lack of Meaningful Human Review or Intervention	Insufficient review of AI outputs	Inadequate human review of AI outputs can lead to prohibited processing and unfair decisions with legal effects (where human review is legally required).	Develop and conduct role-based training for human oversight, focusing on the AI system's optimal applications, effective result interpretation, troubleshooting techniques, combating automation and other detrimental biases, and complying with automated decision-making rights and their related documentation needs.
			Document and evaluate the integration of significant human oversight in AI-driven decision processes, detailing the nature of human input, the reviewer's details, supplementary data influencing the final verdict, and specific scenarios prompting a system pause or manual override.



Fairness

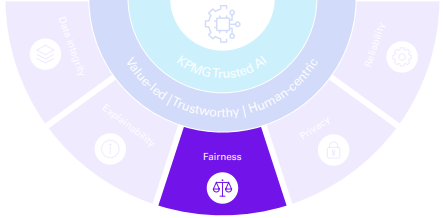
10 pillars of the Trusted AI framework

Click each pillar below to explore

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

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

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.



Fairness

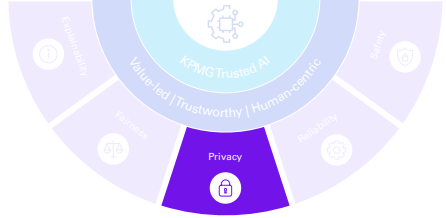
10 pillars of the Trusted AI framework

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

Click each pillar below to explore

- 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	Unfair results due to bias and inclusivity	Lack of attention to bias and inclusivity in AI systems, along with failure to identify and assess group sensitivities during system development, may result in discriminatory outcomes, reduced fairness, and exclusion of certain user groups, impacting the fairness of outcomes and consumer trust.	Algorithms are selected to align with organizational guidance with respect to fairness and risk, ensuring strategic alignment with the organization's objectives.
			Conduct periodically fairness assessments, documenting outcomes and comparing them against pre-defined risk tolerance levels to ensure ongoing adherence to fairness objectives. Remediation strategies are deployed and documented as necessary.
			Evaluate and record the AI system's capability to process diverse sub-population data accurately, both before and after deployment, using bias assessments. Mitigation strategies are implemented for any identified biases to prevent algorithmic discrimination. All findings, actions, and rationales are thoroughly documented, alongside any counterbalancing measures.
			Post-deployment, continuously monitor outputs for bias against ethical/legal standards. Any issues related to detected biases are thoroughly documented, and specific corrective actions are promptly implemented for effective remediation.
	Unrepresentative training data	Potential bias and lack of inclusivity in solution development can arise from failing to identify and assess group sensitivities, impacting the fairness of outcomes.	Evaluate all datasets for inclusivity, identifying and addressing gaps with a remediation plan, including public databases, to eliminate existing biases. All steps and findings are documented.



Privacy

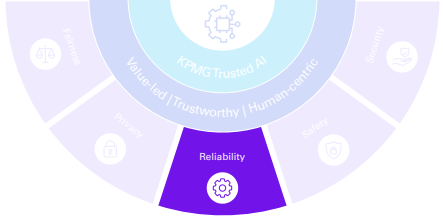
10 pillars of the Trusted AI framework

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

Click each pillar below to explore

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

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Privacy Violations from AI Solutions	Data subject access privacy	Lack of operational infrastructure to enable individuals to exercise their data subject access rights timely may result in a loss of consumer trust, regulatory noncompliance, or cause financial harm.	Launch awareness programs aimed at educating data subjects about their rights in relation to AI technologies, and explaining how to exercise these rights and the implications of AI decision-making on their personal data.
	Privacy directives and regulatory noncompliance	Lack of compliance and alignment with organization directives and/or regulations on processing data subjects may lead to financial penalties, market losses, and reputational damage.	Reviews are periodically conducted over the input, training data, and output utilized by AI solutions to ensure that the use of data remains in compliance with the organization's data privacy directives and relevant regulatory requirements.
			Monitor and assess AI system purpose changes, ensuring any new personal data use is fair, lawful, and transparent.
	Privacy violation due to data breach	Potential data breaches may result in the unauthorized access or disclosure of personal, official use, confidential, and strictly confidential data, which could compromise user or organization privacy, violate data protection laws, lead to reputational damage, or cause financial harm.	A robust oversight system is implemented, including ethical reviews, regular audits over data protection measures, impact assessments, and compliance checks, particularly when the use of sensitive personal data for AI training or production is undertaken.
			Document rationale and explicit approval when obtaining data for training. Special precautions are implemented for AI use cases that may directly or indirectly affect vulnerable individuals or have safety or rights implications.
			To a degree 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

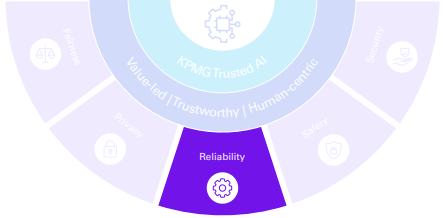
10 pillars of the Trusted AI framework

Click each pillar below to explore

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

AI solutions should consistently operate in accordance with their intended purpose and scope and at the desired level of precision.

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.



Reliability

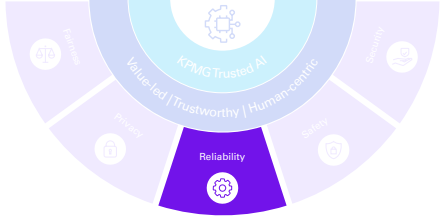
10 pillars of the Trusted AI framework

AI solutions should consistently operate in accordance with their intended purpose and scope and at the desired level of precision.

Click each pillar below to explore

- 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 resiliency and continuity of AI services	Lack of resiliency in AI systems and services, including inadequate backup and restore capabilities and insufficient availability in case of a disaster, may result in extended downtimes and failure to provide critical functions and/or services in a safe, accurate, and timely manner.	Implement failover mechanisms such as automatic backup system switching and frequent system backups, including component snapshots and rollback capabilities, as a fail-safe against unexpected failures to ensure the AI system has the ability to manage unforeseen circumstances without compromising its overall performance or reliability.
			Include advanced support and warranty arrangements in contracts with AI vendors, ensuring system availability and effectiveness via clear service levels and monitoring.
			IT architecture documentation is maintained and updated on an as-needed basis for each AI system to ensure that AI systems are "resilient-by-design" (redundancy and high availability). At a minimum, documentation describes redundancy, availability, and AI-specific risk mitigation (e.g., split brain effect).
			Manage availability and capacity for both IT infrastructure and the AI system, ensuring optimal system performance, stability, operation, and scalability for future needs within the architecture design.
	Lack of a robust quality management system	Lack of a comprehensive and systematically documented quality management system for high-risk AI systems may lead to noncompliance with regulatory requirements, resulting in the deployment of AI systems that are unsafe, ineffective, or violate ethical standards and a loss in consumer trust.	Document and maintain quality control mechanisms and validation results used throughout the development lifecycle of the AI system to ensure the integrity, safety, and efficacy of high-risk AI systems through rigorous design, development, and quality assurance practices.
			Implement automated post-deployment monitoring mechanisms to ensure the safety and reliability of high-risk AI systems.
			Throughout the AI lifecycle, maintain records of all information pertinent to the resources of high-risk AI system, including development details, modifications, compliance, data management, risk management activities, and post-deployment monitoring activities to ensure comprehensive and effective management of supporting resources.



Reliability

10 pillars of the Trusted AI framework

Click each pillar below to explore

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

AI solutions should consistently operate in accordance with their intended purpose and scope and at the desired level of precision.

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Insufficient Understanding of AI Architecture	Insufficient testing of AI system	Lack of robust testing mechanisms and development for AI systems or poor development may lead to undetected errors, resulting in inaccurate outputs, poor decision-making, and reduced reliability.	A recognized certificate authority is used for code signing, enabling operating systems and other tools to verify signature validity. Code signing processes, including certificate renewal, rotation, revocation, and protection are periodically reviewed.
			Document comprehensive test plans (e.g., UAT, SIT), including scope, objectives, and scenarios, with regression tests to safeguard against vulnerabilities. Test execution, results, and approvals are thoroughly 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.

Click each pillar below to explore

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

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Inadequate Response to AI-Generated Safety Threats	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.	A subset of AI-only threat response decisions is periodically reviewed to ensure that decisions are ethical, responsible, and aligned with business objectives. The review is performed by authorized persons within the organization and review documents are retained.
			Anomaly detection systems are implemented to detect suspicious activities (e.g., prompt injection, data poisoning, abuse, evasion, or privacy attacks; increased traffic in a communication channel; and indirect prompt injection) within a system.
	Generation of harmful or unreliable content (e.g., hallucinations)	Generative AI outputs may be harmful, offensive, biased, or misleading and could negatively impact the organization, communities, or society.	Feedback loops within the AI System are implemented to continuously validate and verify system outputs to ensure that the AI is not generating content (including hallucinations) that is harmful; inaccurate; or deviates from intended use, business objectives, or defined parameters.
Threat to Humans	Lack of human intervention	Human unawareness of AI use and lack of proper oversight may result in the inability to override and/or correct decisions made by AI systems.	Develop approved policies and procedures to disclose AI-generated or manipulated content (e.g., deep fakes) that resembles existing persons, objects, places, or events. Ensure training and awareness to the relevant stakeholders to enforce compliance.
			Human moderators reply to reports of AI misuse or inaccurate outputs/decisions, ensuring the AI system's decisions are appropriately vetted and responded to. Any needed reversal in action is taken in a timely manner.

Security

10 pillars of the Trusted AI framework

Click each pillar below to explore

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

Robust and resilient practices should be implemented to safeguard AI solutions against bad actors, misinformation, or adverse events.

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.

Security

10 pillars of the Trusted AI framework

Click each pillar below to explore

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

Robust and resilient practices should be implemented to safeguard AI solutions against bad actors, misinformation, or adverse events.

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
AI Security	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 quickly respond to incidents.	Alert mechanisms are implemented to continuously identify, track, and alert any security breach and/or malfunction that may impact the operation, performance, and safety of the AI system. The AI system is superseded, disengaged, deactivated, or decommissioned, as needed. When required by international regulatory bodies, alerts are reported to the appropriate governing body.
	Lack of AI architecture segregation	Lack of architectural segregation, especially in a cloud/multi-tenant system, may lead to increased vulnerability to security breaches, unauthorized access, and data corruption in the AI landscape and cause financial loss or reputational damage.	AI system's IT architecture (components and data) is segregated from other IT infrastructure/ cloud components to ensure logical segmentation of AI systems within a multi-tenant cloud system, protecting data confidentiality, integrity, and availability.
			Security-by-design principles are embedded in the AI architecture, approach, and development methodology to ensure appropriate and sustainable level of security.
	Poor response to corrective actions prescribed by authorities	The cybersecurity risk posed by AI deployments to the organization's operations, assets, and individuals is not understood and captured by the organization through security policies and procedures, which may result in exposure to malicious attacks or data breaches.	Track and manage AI pipelines and cybersecurity risks with end-to-end visibility as part of the standard risk management process.

Security

10 pillars of the Trusted AI framework

Click each pillar below to explore

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

Robust and resilient practices should be implemented to safeguard AI solutions against bad actors, misinformation, or adverse events.

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
AI Security	Security principles for AI systems	Lack of adherence to security principles in AI design, development, and deployment, in line with the organization's existing policies and procedures, may result in security vulnerabilities, malicious attacks, data breaches, and development of unsecure or unreliable AI system.	A comprehensive inventory of APIs is maintained, tracking their access to internal systems to ensure secure and controlled API integration.
			AI system's training data should be configured securely against human or machine tampering. Checks should be automatically performed on the completeness and accuracy of the training data against tampering.
			Conduct periodic resiliency and security assessments of the AI system, adhering to organizational best practices and encompassing a range of tests to ensure comprehensive security and sustainability.
Unsafe Prompt Engineering	Prompt injection	Direct or indirect prompt injection can lead to inaccurate outcomes through malicious code injection that may result in unauthorized disclosure of personal, official use, confidential, and strictly confidential information.	Deploy a secure parsing system using custom markup languages like enhanced ChatML for OpenAI API calls, incorporating content security policies and sandboxing to securely encapsulate and execute external content, minimizing security risks.
			Develop a zero-trust architecture with dynamic trust enforcement, using ACLs, RBAC, and a secure API gateway to verify and control interactions between the LLM, external sources, and plugins, ensuring all operations are authorized and validated.

Sustainability

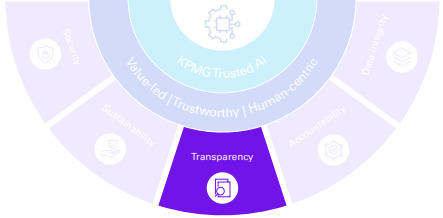
10 pillars of the Trusted AI framework

AI solutions should be designed to be energy efficient, reduce carbon emissions, and support a cleaner environment.

Click each pillar below to explore

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

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Overarching Risk Associated with AI Sustainability	Failure to prioritize the sustainable development of AI systems	Environmental impact is not considered in AI system strategy and design, which may result in energy inefficient systems.	During AI Strategy and Development, establish clear sustainability goals for the AI system, aligned to the organization's standards, and develop a strategy for demonstrating how the AI system will meet the goals throughout its lifecycle.
	Failure to prioritize the sustainable implementation and use of AI systems	Lack of sustainable implementation, use, and monitoring practices may result in system sustainability degradation and misalignment with organizational ESG commitments.	Incorporate environmental impact indicators and real-time monitoring mechanisms across the AI system lifecycle to ensure energy consumption, system efficiency, and emissions adhere to applicable environmental standards and company strategies. Gaps or improvement areas identified are quickly remediated.



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.

Click each pillar below to explore

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

Risk Category	Risk Consideration	Risk Description	Illustrative Control Considerations
Distinguishing Human vs. AI Content	Opacity of AI systems	Lack of AI system transparency can reduce 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 about its applicability and effectiveness.
			Identify and document potential negative residual risks to both downstream acquirers and end users, to provide a comprehensive overview of unmitigated risks associated with the AI system.
Lack of Transparency in AI and Data Usage	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.	Document test sets, metrics, and the tools used during the Test, Evaluation, Validation, and Verification (TEVV) processes to establish a transparent and reproducible framework for assessing the AI system's performance and reliability.
	User transparency	Insufficient transparency in the development and use of AI systems may result in a lack of accountability, making it difficult to understand the rationale behind the system's behavior, raise ethical concerns, and erode consumer trust.	AI-generated or manipulated content is labeled or watermarked (e.g., CP2A) to ensure transparency and lineage of AI created content.
			For each output generated by the AI system, users are explicitly informed of potential inaccuracies in the results, with a strong recommendation to critically review the AI system's outputs.
			Prior to each use, users of the AI system are notified of data collection and/or processing for personalization and recommendation purposes. When notified, users are presented the option to opt out of such services to ensure transparency and user choice.
			Users or those impacted by emotion recognition or biometric categorization AI systems are notified of the system's operation prior to their use.

Designing controls for your AI systems

The control considerations in this guide offer a foundation for creating tailored control descriptions for your AI deployments. We've also included a few example control implementation descriptions for inspiration to get you started. If you have any questions, do not hesitate to reach out to our team.

Pillar	Risk Category	Illustrative Control Consideration	Example Control Implementation Description
Accountability	AI performance erodes over time	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.	Quarterly, the AI system owner reviews a sample of the AI system's outputs against established key performance indicators (KPIs) and key risk indicators (KRIs) to ensure it is performing as expected. Any discrepancies or variances above established thresholds are investigated and resolved within 5 business days. If a major discrepancy is identified, the system is pulled back from production immediately.
Fairness	Harmful Bias in AI Systems	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.	Annually, all team members who create and develop AI systems are required to complete the "AI Fairness and Accessibility" training course. After completing the course, all team members are required to take a post-training assessment where a minimum score of 85% is required to pass.
Data Integrity	Lack of Data Integrity in AI Systems	During the change management process for an AI system, the training and testing data used is evaluated for relevancy and accuracy with the change. As needed, additional data is introduced to train and test new system capabilities or features.	When making a change to an AI system, perform regression or error rate testing as defined by the Change Management policy. Any issues identified during testing greater than "low" are resolved prior to deployment into production.
Transparency	Lack of Transparency in AI and Data Usage	For each output generated by the AI system, users are explicitly informed of potential inaccuracies in the results, with a strong recommendation to critically review the AI system's outputs.	For each output generated by the AI system, a disclaimer is included at the beginning of the generated text output, stating: "Outputs generated by this system may include inaccurate, incomplete, or out-of-date information. Consequently, they may not be relied on without applying professional judgement."
		Prior to each use, users of the AI system are notified of data collection and/or processing for personalization and recommendation purposes. When notified, users are presented the option to opt out of such services to ensure transparency and user choice.	Prior to each use of the AI system, an acknowledgement window stating, "I consent to the collection of my data through the use of this system," is displayed in the user interface, blocking access to use [System A]. Users are prevented from using the AI system unless they provide their consent by clicking "I acknowledge."

How KPMG can help

The KPMG Trusted AI framework offers a pathway to help harness AI's potential in a trusted manner, and our suite of AI Trust services and solutions helps companies put the framework into action.

Our services include:

01 **Trusted AI strategy:** Assist organizations in assessing their current AI capabilities and crafting strategic roadmaps that enhance potential.

02 **AI ethics and governance:** Assist in the development of robust AI governance frameworks, controls, and operating models to help ensure AI is trustworthy. This includes comprehensive risk, policy, and controls assessments, alongside AI regulatory compliance.

03 **AI risk assessment and regulatory compliance:** Help organizations assess where they are in their Trusted AI journey by conducting risk-based AI assessments across AI use cases. This includes AI readiness, maturity assessments, AI strategy review, and assessing consistency of AI solutions with evolving frameworks and regulations.

04 **Machine learning operations:** Develop leading constructs, processes, and technologies for model management to help build trust in AI models, supporting their governance, lifecycle management, and effective deployment and monitoring.

05 **AI security:** Provide strategies, processes, and tools to help enhance AI security and privacy, helping organizations detect, respond to, and recover from cyber threats, privacy risks, and adversarial attacks.

06 **AI assurance:** Help test, examine, and report on the management processes, controls, and claims regarding the responsible use of AI technologies:

- AI assurance scoping
- AI diagnostics reviews
- AI model control testing

For more information: **visit.kpmg.us/TrustedAIservices**

Need a customized AI Risk and Controls Guide?

KPMG can help customize and tailor the AI Risk and Controls Guide to meet the specific needs and challenges of your organization, provide targeted training and education to help ensure a deep understanding and effective application of the matrix's principles, and deliver ongoing support and advisory services to navigate emerging AI risks and opportunities. Specific services we offer that can help your team tangibly implement the framework include:

- **AI governance design and operations support:** establishing or enhancing your AI governance program, policy, and operating model, or helping to scale and operationalize your AI governance program
- **Regulatory mapping:** mapping to existing taxonomies to help ensure a complete control portfolio
- **Lifecycle mapping:** aligning controls that best fit to different stages of the AI lifecycle
- **Control implementation support:** documentation, design, and implementation support for AI controls
- **AI assessments:** conducting AI assessments, compliance assessments, or risk-based governance assessments

Discover how we can help you along your Trusted AI journey.

Contact us

Bryan McGowan
Global Trusted AI Leader
KPMG International
E: bmcgowan@kpmg.com

Samantha Gloede
Managing Director
Global Trusted Leader
KPMG International
E: sgloede@kpmg.com

Xxxx Xxxx
XXXX
XXXXXX XXXX
KPMG International
E: xxx@kpmg.com

Xxxx Xxxx
XXXX
XXXXXX XXXX
KPMG International
E: xxx@kpmg.com

Xxxx Xxxx
XXXX
XXXXXX XXXX
KPMG International
E: xxx@kpmg.com

us-connectwithus@kpmg.com

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

Learn about us: **kpmg.com**