



# 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



# Foreword

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

**Artificial intelligence (AI) is revolutionising 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 organisation.**

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.<sup>1</sup>

Unsurprisingly, such benefits make executives eager to integrate AI into their businesses and accelerate the value it delivers. **But organisations 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 organisation'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 organisations 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 organisations 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 organisation.

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

—**Kelly Henney**, Partner,  
**Privacy and Data Protection Lead Partner,**  
**Trusted AI Lead Partner, KPMG Australia.**

<sup>1</sup> 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 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 organisation, 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 organisations

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 organisation and AI system?

Not every organisation 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.



# 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.

Organised 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 organisations can leverage as they build out their control catalogues.






# Accountability

## 10 pillars of the Trusted AI framework

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

 Click each pillar below to explore

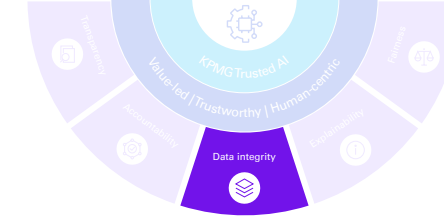




# Data integrity

## 10 pillars of the Trusted AI framework

Click each pillar below to explore



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  |
|---|--|--|--|
| <b>Lack of data integrity in AI systems</b> | 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.  | <p>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 organisation values. Training and awareness campaigns are performed for relevant stakeholders to enforce compliance. The policies and procedures are reviewed and updated, as needed, periodically.</p> <p>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.</p> |
|   | 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. | <p>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.</p>  |

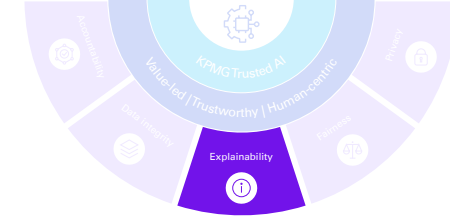


# Explainability

## 10 pillars of the Trusted AI framework

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.

 Click each pillar below to explore

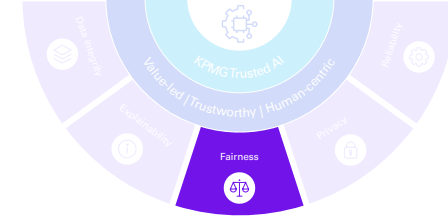




# Fairness

## 10 pillars of the Trusted AI framework

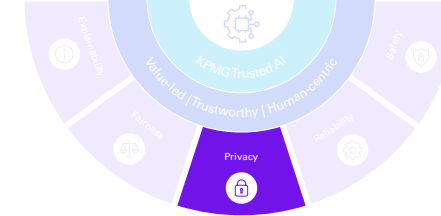
✦ Click each pillar below to explore



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



# 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

| Risk Category                               | Risk Consideration                              | Risk Description  | Illustrative Control Considerations  |
|---|---|---|--|
| <b>Privacy violations from AI solutions</b> | 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 organisation 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 utilised by AI solutions to ensure that the use of data remains in compliance with the organisation'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 unauthorised access or disclosure of personal, official use, confidential and strictly confidential data, which could compromise user or organisation 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

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





# Safety

## 10 pillars of the Trusted AI framework

Click each pillar below to explore

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



| Risk Category   | Risk Consideration   | Risk Description  | Illustrative Control Considerations   |
|---|--|---|---|
| <b>Inadequate response to AI-generated safety threats</b> | AI system errors are improperly resolved                           | Errors in the AI system remain undetected, detected late or not acted upon timely, resulting in unauthorised changes, system unavailability, security breaches, data loss or other incidents. | <p>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 authorised persons within the organisation and review documents are retained.</p> <p>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.</p> |
|   | Generation of harmful or unreliable content (e.g., hallucinations) | Generative AI outputs may be harmful, offensive, biased or misleading and could negatively impact the organisation, 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.  |
| <b>Threat to humans</b>                                   | 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

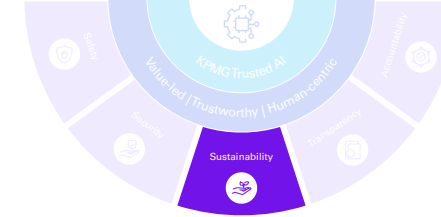
Robust and resilient practices should be implemented to safeguard AI solutions against bad actors, misinformation or adverse events.



 Click each pillar below to explore



# 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

| Risk Category   | Risk Consideration   | Risk Description   | Illustrative Control Considerations  |
|---|--|--|--|
| <b>Overarching risk associated with AI sustainability</b> | Failure to prioritise 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 organisation's standards, and develop a strategy for demonstrating how the AI system will meet the goals throughout its lifecycle.  |
|   | Failure to prioritise 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 organisational 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

Click each pillar below to explore



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 Category                                    | Risk Consideration                          | Risk Description  | Illustrative Control Considerations  |
|--|---|---|--|
| <b>Distinguishing human vs. AI content</b>       | 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 generalisability 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.  |
| <b>Lack of transparency in AI and data usage</b> | 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.                      |
|  |   |   | AI-generated or manipulated content is labeled or watermarked (e.g., CP2A) to ensure transparency and lineage of AI created content.   |
|  | 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. | 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 personalisation 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 categorisation 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   |
|-----------------------|---|--|--|
| <b>Accountability</b> | 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. |
| <b>Fairness</b>       | 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.   |
| <b>Data integrity</b> | 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.   |
| <b>Transparency</b>   | 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 personalisation 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 AITrust services and solutions helps companies put the framework into action.

Our services include:

- 01 Trusted AI strategy:** Assist organisations 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 organisations 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 organisations 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.com/TrustedAIservices](https://www.kpmg.com/TrustedAIservices)

## Need a customised AI Risk and Controls Guide?

KPMG can help customise and tailor the AI Risk and Controls Guide to meet the specific needs and challenges of your organisation, 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 operationalise 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



**Kelly Henney**  
Partner  
Privacy and Data Protection Lead Partner  
Trusted AI Lead Partner  
KPMG Australia  
E: [khenney@kpmg.com.au](mailto:khenney@kpmg.com.au)



**Dermot Mccutcheon**  
Director  
Trusted AI Solutions  
KPMG Australia  
E: [dmccutcheon@kpmg.com.au](mailto:dmccutcheon@kpmg.com.au)



**Bryan McGowan**  
Global Trusted AI Leader  
KPMG International  
E: [bmcgowan@kpmg.com](mailto:bmcgowan@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.au](https://www.kpmg.com.au)

©2025 KPMG, an Australian partnership and a member firm of the KPMG global organisation of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organisation.

The information contained in this document is of a general nature and is not intended to address the objectives, financial situation or needs of any particular individual or entity. It is provided for information purposes only and does not constitute, nor should it be regarded in any manner whatsoever, as advice and is not intended to influence a person in making a decision, including, if applicable, in relation to any financial product or an interest in a financial product. Although we endeavour to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act on such information without appropriate professional advice after a thorough examination of the particular situation.

To the extent permissible by law, KPMG and its associated entities shall not be liable for any errors, omissions, defects or misrepresentations in the information or for any loss or damage suffered by persons who use or rely on such information (including for reasons of negligence, negligent misstatement or otherwise).

Liability limited by a scheme approved under Professional Standards Legislation.