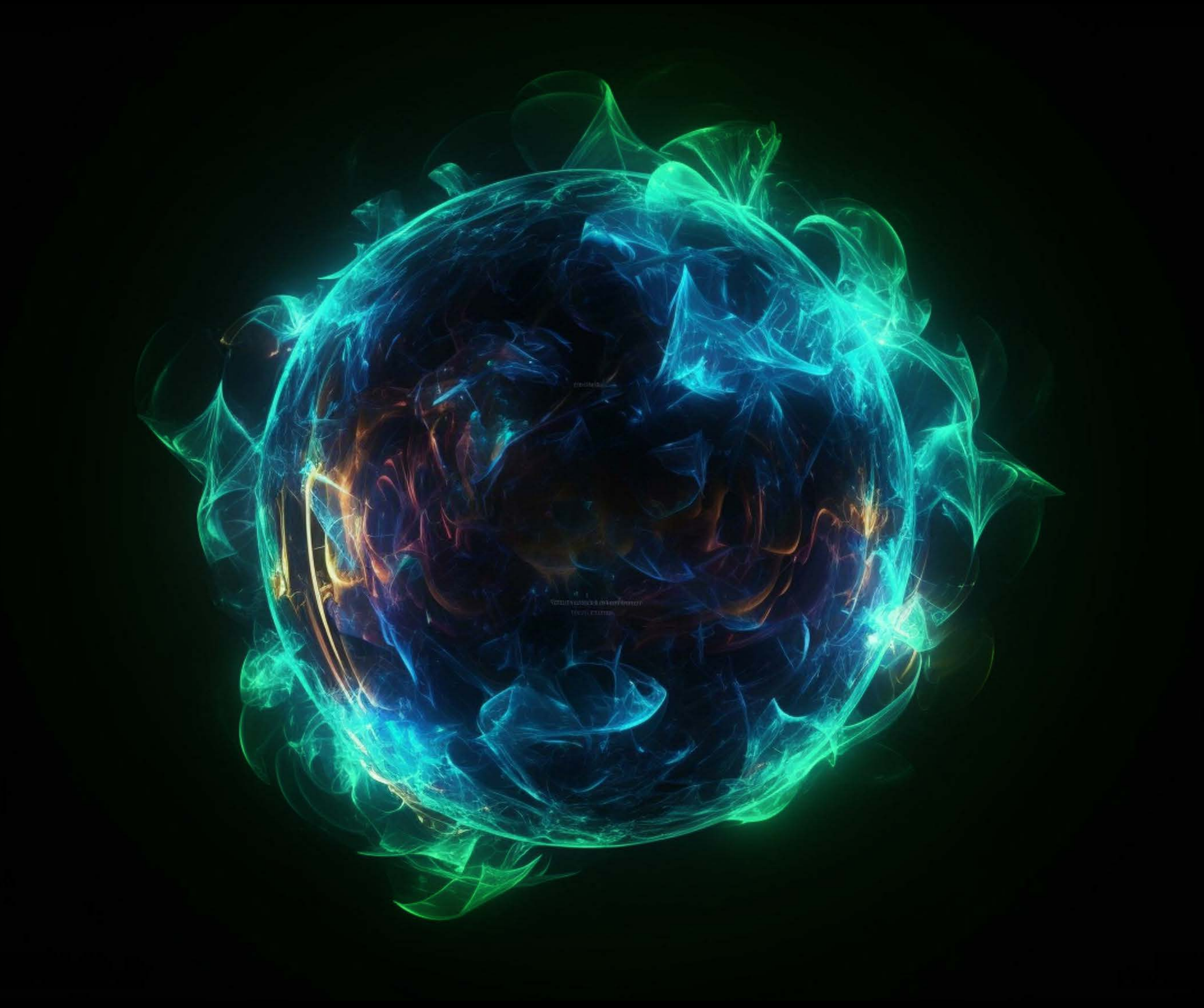
Deloitte.



AI Risk Management

Risk mitigation "now" and strategic insights "next"

March 2024

Table of contents

Introduction 04

Enhancing trustworthiness at every stage of the AI lifecycle 04

A layered approach to building a trustworthy AI 05

AI risk universe—Illustrative 07

Deloitte’s Trustworthy AI framework 11

Need for governance structure across the AI lifecycle 16

Way forward 17

Introduction

In today's growing market, Artificial Intelligence (AI) is an imperative for various industries. Organisations are exploring the use of AI for several solutions, including automation, to deliver value and bring efficiency to operations. If companies are relying heavily on AI, they need to ensure ethical assurance and trustworthiness to make their AI systems dependable.

A solid framework can help organisations navigate this journey and gain confidence against various regulatory requirements as the AI landscape evolves.

Enhancing trustworthiness at every stage of the AI lifecycle

Ideation and design

Accounting for applicable regulations for the business/industry and local or target geography to ensure compliance by design from the initial stages of the AI project.

Model requirement

Establishing clear and comprehensive requirements during the model requirement stage of the AI lifecycle to ensure successful development and deployment; data scientists and engineers could be involved proactively during this stage to minimise the risk of project failures, costly rework, and potential ethical or legal issues.

Model development

Forming standards and best practice guidelines for developers, ensuring their technologies adhere to compliance requirements at every stage of the AI lifecycle.

Model deployment

Demanding concrete and trustworthy demonstration from developers and/or vendors, ensuring their AI systems adhere to relevant ethical, legal, and technical standards.

Data requirement

Certifying prerequisites in available data for AI models, such as adequacy, representativeness, and high quality to prevent bias, discrimination, and unreliable results.

Data cleansing

Conducting data cleansing (error detection, standardisation, and normalisation) to eliminate errors, ensure consistency, and optimise model performance in AI projects.

Data labelling

Ensuring accurate and detailed labels with bias mitigation to avoid errors and manage data effectively.

Model training and testing

Conduct thorough validation and testing of the model using diverse datasets, including both training and validation data. Perform sensitivity analysis and stress testing to assess the robustness and reliability of the model under different scenarios. Use adversarial testing to identify vulnerabilities and potential security risks, such as adversarial attacks.

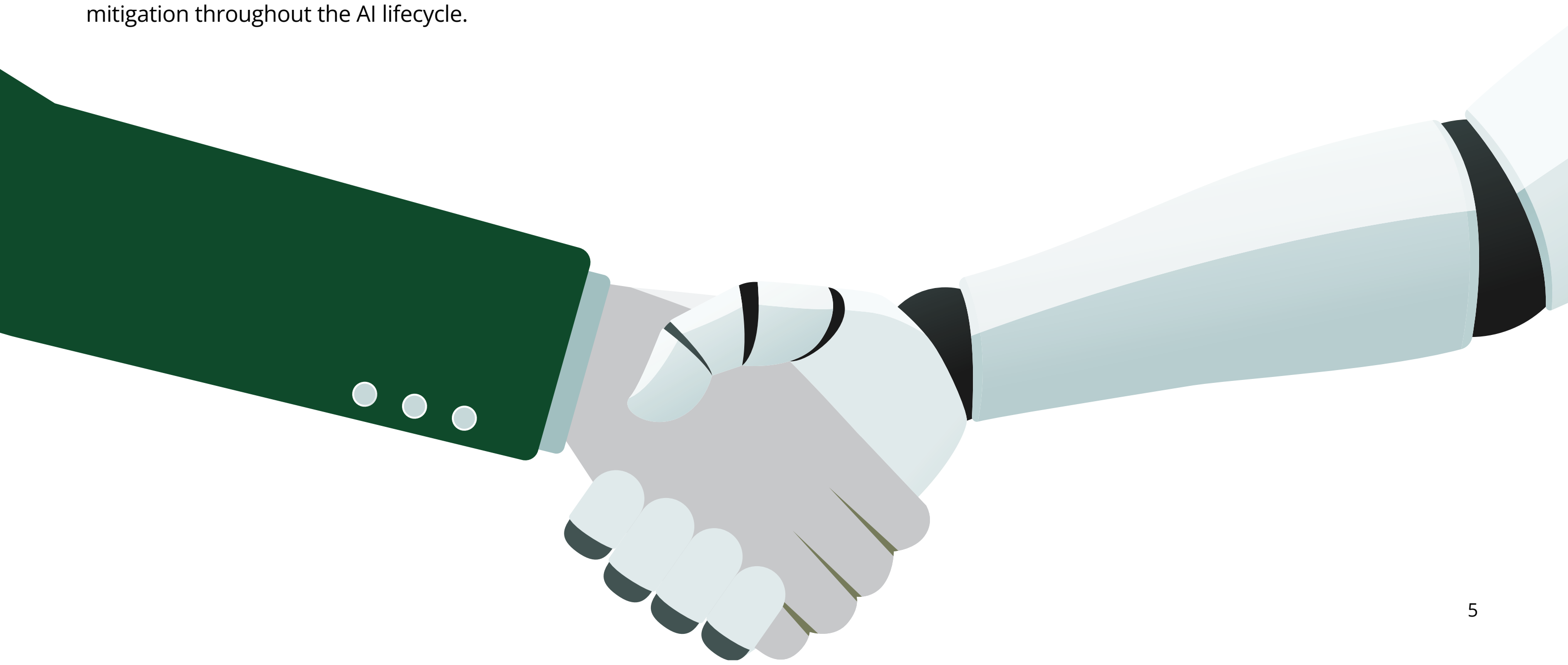
Model monitoring

Conducting continuous performance tracking, data drift detection, model retraining, maintaining transparency, and confirming compliance with regulations and ethical standards to maintain model reliability and accountability in decision-making.

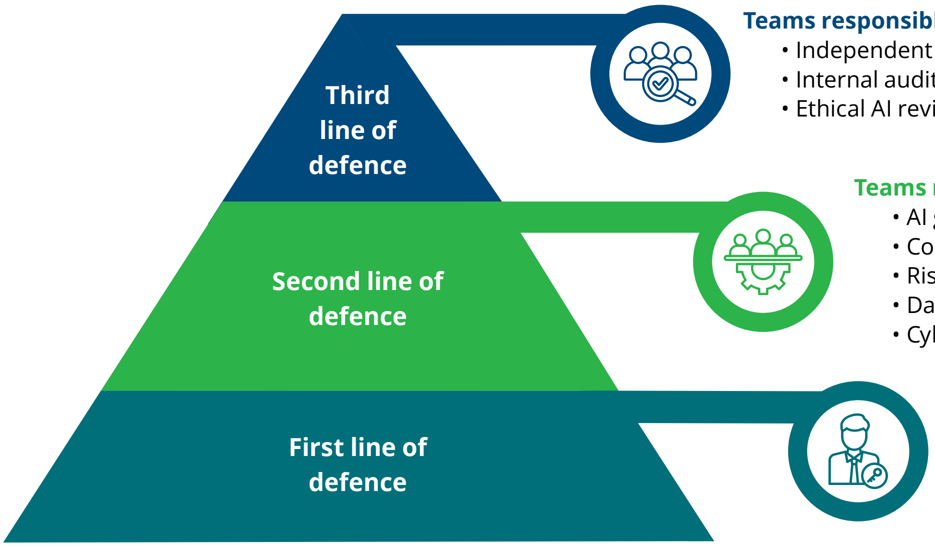
A layered approach to building a trustworthy AI

To achieve a strong AI governance and risk management, it is crucial to establish multiple security layers when deploying AI programs. The three Lines of Defence (3LoD) model is a fundamental framework that delineates three integral layers of defence, each with unique responsibilities and accountabilities. At the core of this framework lies the pivotal role of personas, seamlessly integrated across these lines of defence.

Through this process, organisations establish a resilient AI governance structure and foster transparency, accountability, and risk mitigation throughout the AI lifecycle.



Lines of defence



Teams responsible:
- Independent AI assurance and audit team
- Internal auditors
- Ethical AI review board

Teams responsible:
- AI governance team
- Compliance and ethics team
- Risk and compliance function
- Data privacy officer
- Cyber security experts

Teams responsible:
- Business unit owners
- AI developers
- AI/ML engineers

Assurance checks

First line

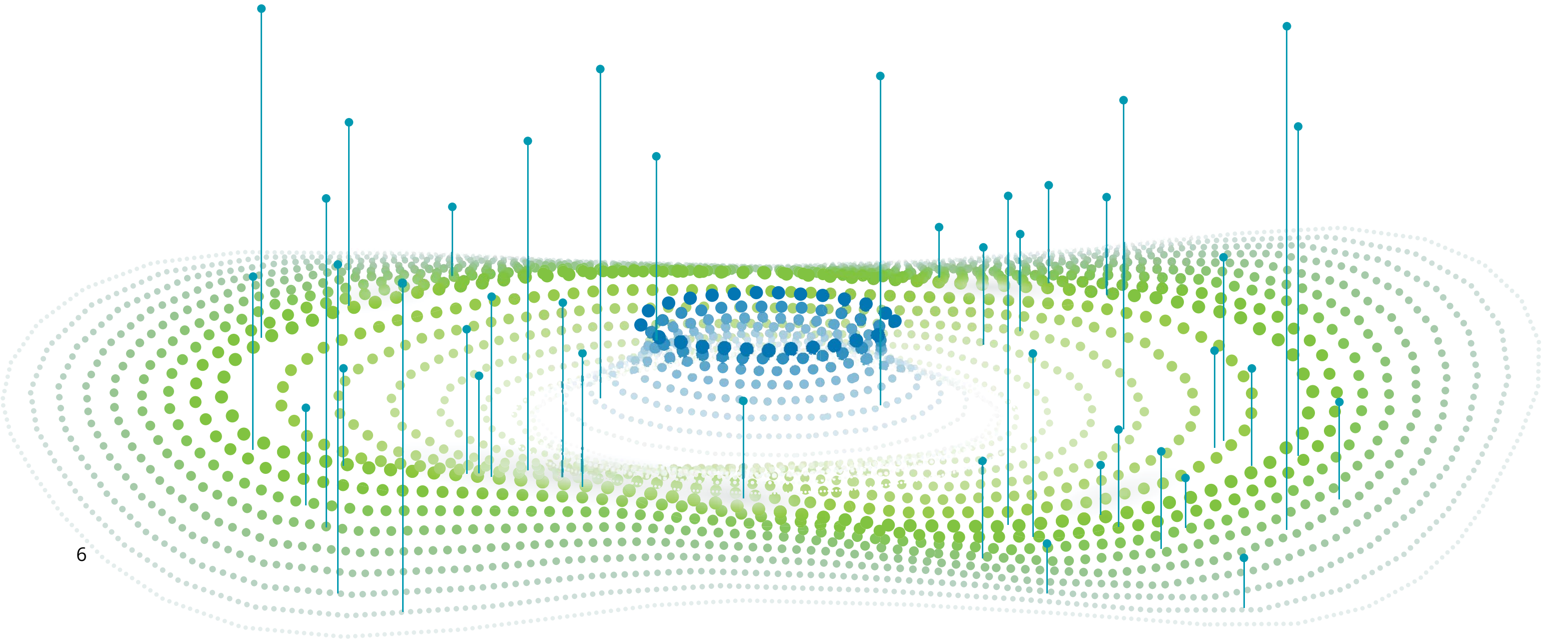
- Enable increased first line of defence testing by model owners through stress testing and continuous testing
- Automating model validation and monitoring
- Define model parameters and refine model development processes

Second line

- Setting AI risk appetite
- Identifying KRIs
- Including forward-thinking risk taxonomies
- Review model docs via governance dashboards
- Establish AI risk strategy
- Enable real-time issue alerts

Third line

- Enhance transparency and accountability with internal audits by sharing model data and enabling audit trails, etc.



AI risk universe—Illustrative

Awareness of the following risks in the AI development lifecycle is crucial for promoting responsible design, ensuring ethical implementation, and fostering sustainable technological advancement.

Strategic

Risk category description

Risk of AI strategy/leadership not aligned to organisational/business objectives/leadership

Individual risks

Model requirements:

- AI strategy not coordinated with company strategies/value systems/risk appetite leads to ineffective or even malicious/unethical models

Financial

Risk category description

Risk of inadequate and incorrect decisions/recommendations due to poor AI models, resulting in direct and indirect losses or threats to the organisation, customer, brand, and reputation

Individual risks

Model evaluation:

- Financial losses, wastage of resources, and reputational losses because of wrong AI models

Data

Risk category description

Risk of unavailability of accurate, labelled, relevant, and unbiased data to develop, train, and deploy models that meet its intended purposes

Individual risks

Data labelling:

- Inaccurate models from mismatched tests, production data, and improper data tagging

Data collection:

- Risk of biased or insufficient data for model development data cleaning
- Unauthorised access disintegrates solution alignment with business goals

Data labelling:

- Test data different from production data can result in inaccurate models, while inadequate data tagging based on sensitivity can result in inappropriate safeguards.

Technology

Risk category description

Risks associated with the technology used regarding auditability, scalability, and monitoring

Individual risks

Model monitoring:

- Tech constraints limit auditability and audit logs, hindering transparency
- Lack of monitoring and feedback loops delay corrections for model discrepancies

Model deployment:

- Single points of failure in deployment without redundancy and inflexible technology limit scalability as the organisation grows.

Algorithmic

Risk category description

Risks associated with the algorithms leading to incorrect/inconsistent/biased/unethical decisions and financial and reputational implications.

Individual risks

- **Model training:** Biased data begets biased and unreliable AI models
- **Model evaluation:** Inadequate risk-based stress testing and documentation can harm models
- **Model deployment:** Insecure coding and design flaws invite vulnerabilities
- **Model monitoring:** Absence of mechanisms for monitoring changing environments

Cyber (including Data Privacy)

Risk category description

Risk of not identifying, labelling, storing, and securing Personally Identifiable Information (PII) resulting in data privacy breaches, leading to reputational backlashes and regulatory repercussions.

Individual risks

Privacy: (Data labelling and data collection)

- Insufficiently secured data in AI models, lack of opt-in/opt-out options, and unauthorised data use infringe on privacy rights.

Cyber

Risk category description

Lack of adequate access controls in place to safeguard infrastructure, application, model, and underlying code

Individual risks

Infrastructure:

- Risks pertaining to the underlying technology and resources that support the AI system. This includes servers, networks, databases, and cloud services.

Application:

- Risks involving issues related to the AI application's functionality, usability, and integration

Model:

- Risks focussing on the AI model’s performance, interpretability, and generalisation capabilities

Underlying code:

- Risks involving challenges related to the quality, security, and documentation of the AI system's codebase

People

Risk category description

Risk of unavailability of skilled people at each stage of the AI lifecycle and lack of clear segregation of roles and responsibilities in terms of human-machine interface.

Individual risks

Talent:

- Risk on the company's talent culture (skills atrophy) due to AI implementation may lead to employee resentment.

Governance:

- Insufficient AI skills
- Unclear roles and unapproved developments
- Missing human-machine interaction guidance (Override)
- Expertise loss risk
- Diversity prevents bias

Regulator

Risk category description

Risk of not catering to geographical or sectoral regulatory and compliance requirements with respect to AI models, resulting in litigations, fines, and regulatory scrutiny.

Individual risks

Model evaluation:

- Lack of clarity on regulations and its changes around privacy and data security leads to the creation of ambiguous models, financial penalties and regulatory scrutiny.

Model monitoring:

- Risks such as social engineering and privacy invasion without AI regulation
- Neglecting compliance may result in penalties and business continuity risks

Third/Fourth-party

Risk category description
Risks arising due to the involvement of third/fourth parties in the AI deployment lifecycle may lead to technology dependency and intellectual property loss.

Individual risks

- Unclear vendor roles hinder ownership
- Vague contract terms challenge risk management
- Inadequate security controls risk fines and reputation damage

Societal

Risk category description
Risk of incorrect, inconsistent, biased decisions and recommendations made by AI model leading to issues, such as loss of jobs and exclusion of services causing socio-economic disparity.

Individual risks

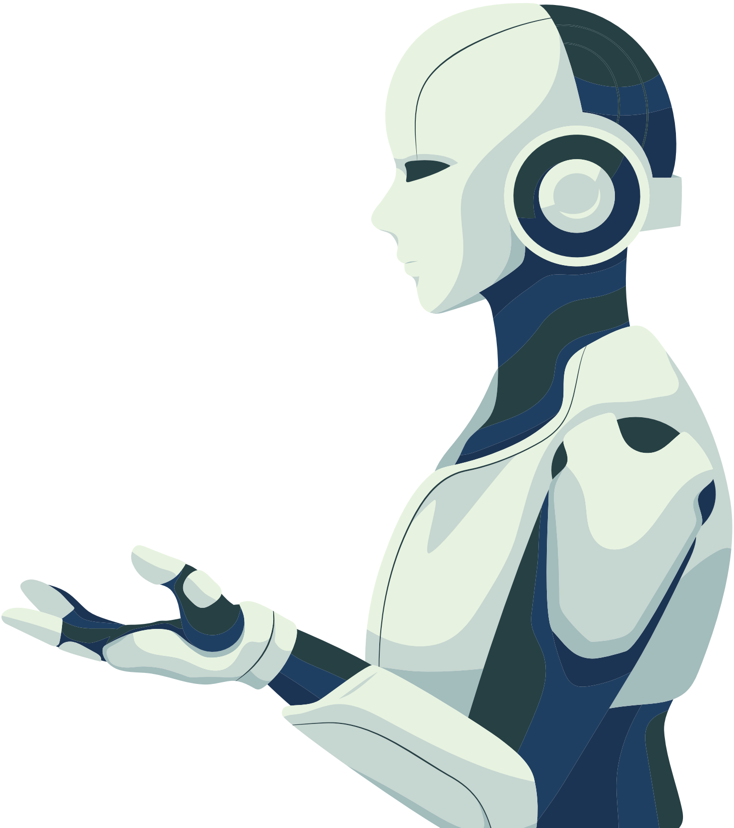
- A lack of societal expectation management erodes trust in AI adoption.
- Non-transparent AI models contribute to societal bias and exclusion.

An independent assessor should address various risks associated with AI models, as meeting regulatory requirements will bolster the entity's trust:

Independent assurance:
To establish confidence and trust in AI systems, it is necessary to demand well-defined, consensus-driven standards and credible evidence from developers, vendors, and executives. This evidence should demonstrate the validity and suitability of the assurance for a specific use case. This can be an internal and/or external assurance team (auditors, certification bodies, etc.)

Regulations and standards compliance:
Seeking assurance involves the essential reliability of AI systems falling under their regulatory purview, ensuring compliance with regulations and best practice guidelines.

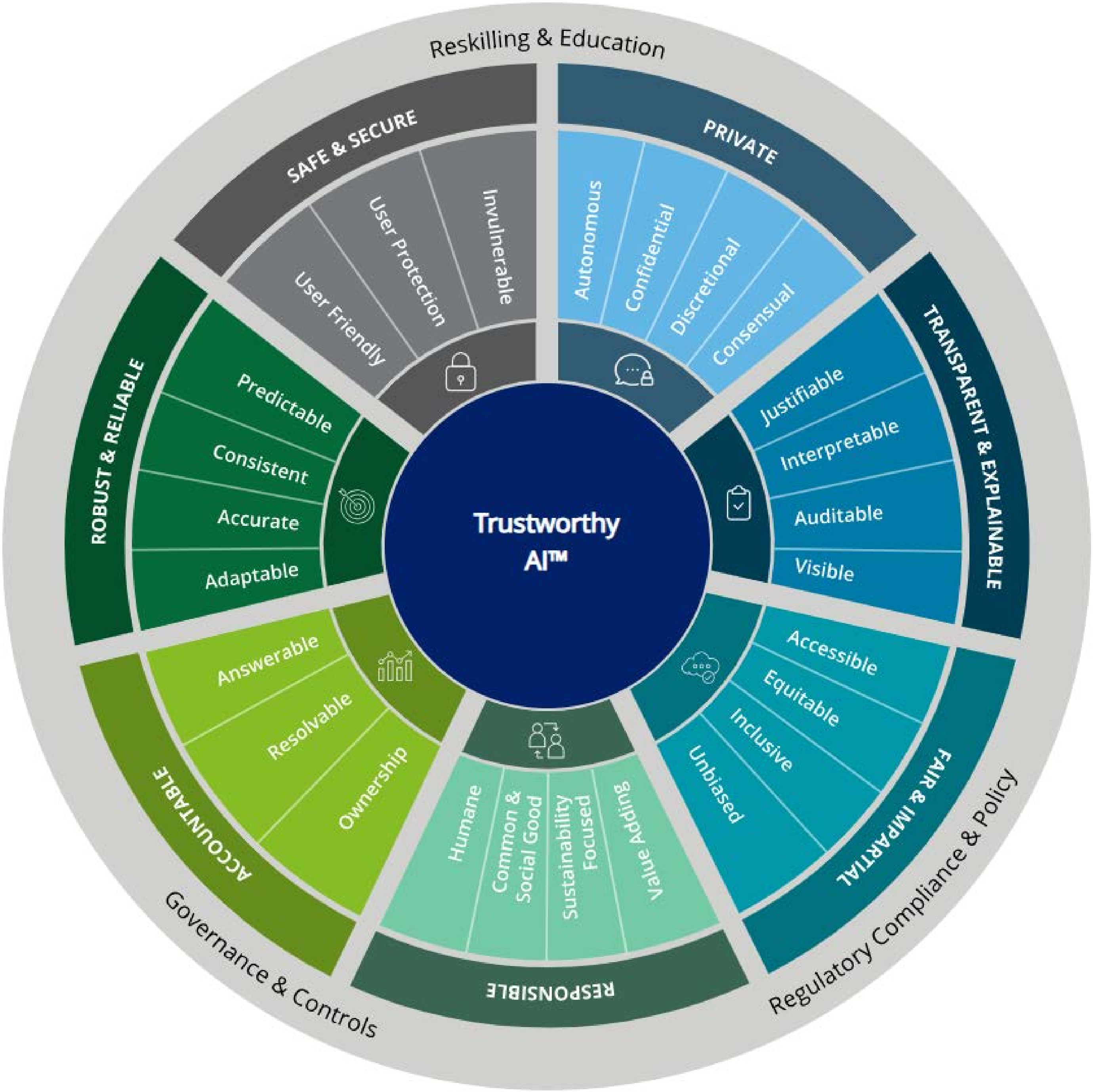
The control frameworks developed by the organisation can use the existing frameworks, such as ISO 27001, ISO 42001, COBIT, GDPR, Fairness Accountability and Transparency in Machine Learning (FAT ML), and implementation guidelines, along with best practices, such as NIST SP 800, NIST AI Risk Management Framework, CIS Controls, and OWASP.



Deloitte's Trustworthy AI™ framework

Governments, industries, and various other groups have struggled to set up an AI framework due to the challenging AI evolution across industries. To bridge the gap, we have developed a Trustworthy AI framework, putting trust at the centre of everything we do.

This helps organisations set up governance structures for AI programmes and meet regulatory compliance throughout the AI lifecycle from ideation to design, development, deployment, and Machine Learning Operations (MLOps) to empower employees, businesses, customers, and industries.



This trustworthy framework is based on the following seven dimensions

Transparent and explainable

AI models enable users to make decisions that are easy to understand, auditable, and open to inspection. This involves assessing system complexity, training methods, and efforts to enhance comprehension. It also examines how the system communicates results, reasoning, involvement in outcomes, and avenues for recourse to users and data subjects.

Fair and impartial

AI models prioritise inclusive design, promoting equitable application, access, and outcomes. An impartiality assessment examines system design to ensure fairness, by considering bias and cultural context. An integral part of this is to provide comprehensive support for displaced workers. Ongoing user bias training and diverse fairness testing are conducted to address potential biases using various definitions.

Robust and reliable

AI models produce consistent and accurate outputs, withstand errors, and recover quickly from unforeseen disruptions and misuse. AI models must maintain robustness and reliability throughout their entire lifecycle. They should operate suitably in various conditions, including normal, foreseeable, and adverse scenarios.

Private

AI models help respect user privacy by limiting data use to its intended purpose and duration. They provide opt-in/out options for data sharing and evaluate transparency in user communication regarding data policies, system risks, testing outcomes, and appropriate use. They also scrutinise privacy by detailing sensitive data types used and strategies for data protection during training and deployment.

Safe and secure

AI models are protected from risks that may cause individual and/ or collective physical, emotional, environmental, and/or digital harm.

Responsible

AI models are created and operated in a socially responsible manner. They put an organisational structure in place that can help determine who is responsible for the output of AI system decisions.

Accountable

Policies dictate responsibility for AI-related decisions. Accountability is gauged by transparent supervision of AI model creation and deployment. This ensures clarity and prevents manipulation, with effective communication of system functions and limitations. It includes validating documented design decisions, system failure reviews, and scenario planning by the AI team.

Enhancing reliability throughout the AI lifecycle

We explore stage-specific techniques to bolster reliability, linking each stage to its Trustworthy AI element, key stakeholders, guiding principles, and crucial audit points to consider.

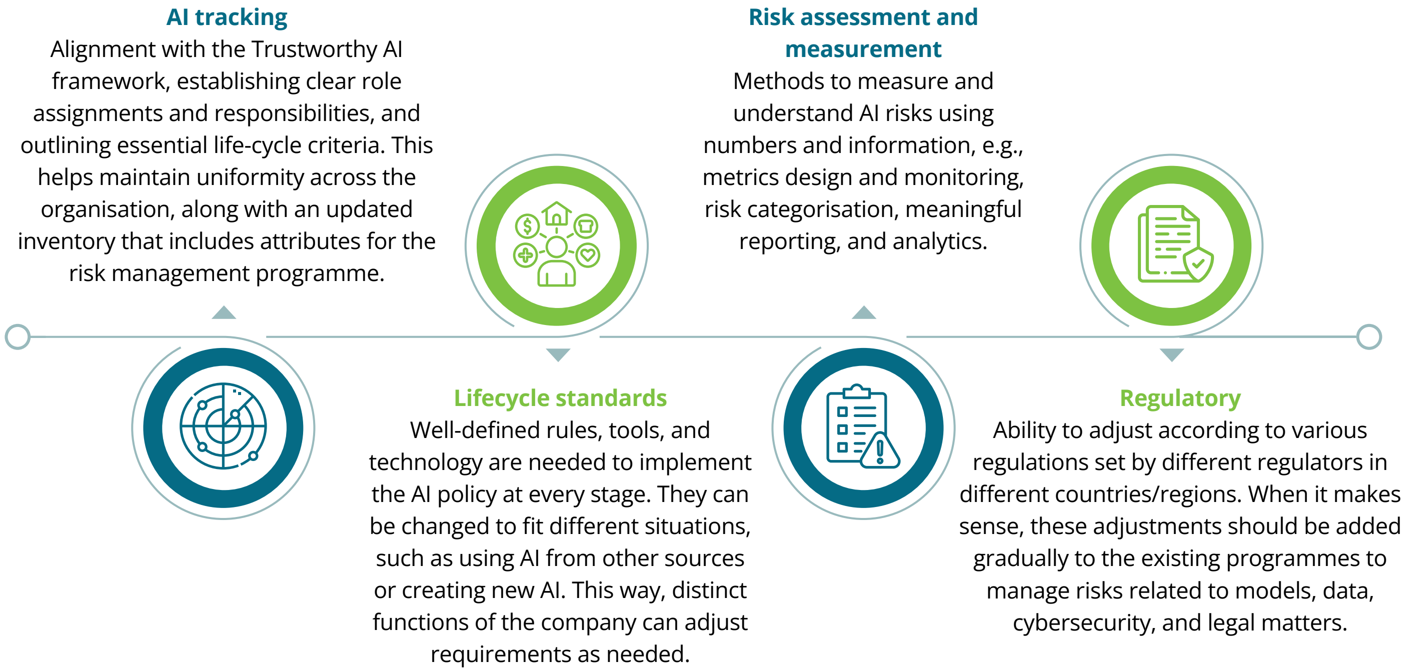
AI lifecycle stage	Trustworthy AI element	Associated persona	Principles	Audit focus
Ideation and design	• Transparent and explainable • Safe and secure	• AI architect • AI developers	• Traceability and explainability of significant decisions taken by the system • Usage of the simplest algorithm that meets performance goals • Ability to override the AI system's decision by designated people • Security of users' data • Following secure coding and security-by-design practices • Ensuring that third/fourth-party stakeholders implement all the necessary security controls	• Assess traceability and explainability implementation • Review algorithm simplicity and decision override mechanisms • Verify security measures and third/fourth-party controls
Model requirement	• Robust and reliable • Accountable	• Business unit owners • AI/ML engineers	• Alignment to the principles of both organisation and responsible AI • Reproducibility and consistency of outcomes • Implementation of appropriate grievance redressal and compensation mechanisms • Quality assurance—Human supervisory control wherever possible	• Scrutinise alignment with responsible AI principles • Validate reproducibility and grievance handling • Review human supervisory control implementation

AI lifecycle stage	Trustworthy AI element	Associated persona	Principles	Audit focus
Data cleansing	• Fair and impartial • Private	• AI governance team • Data privacy officers	• Ensuring system fairness • Minimisation of the use of sensitive data • Usage of representative datasets • Ensuring the quality and correctness of data annotations	• Assess fairness and data quality maintenance • Review procedures for sensitive data handling
Data labelling	• Fair and impartial • Private	• AI governance team • Data privacy officer	• Setting clear goals for diversity and inclusion • Countering various sources of bias • Testing the AI system with diverse user groups	• Evaluate diversity and bias mitigation • Review testing procedures with diverse user groups
Model training	• Robust and reliable	• AI/ML engineers • Risk and compliance functions	• Quality Assurance • Monitor the feedback to the system • Implementation of failover mechanisms • Optimisation of the model's inference speed • Proper integration with data sources and other AI systems • Implementation of ML Ops • Usage of risk-based stress testing techniques	• Validate quality assurance and feedback monitoring • Review failover mechanisms and stress testing implementation • Review the documentation of the training process for transparency and reproducibility • Verify the adherence to legal and compliance requirements during the model training

AI lifecycle stage	Trustworthy AI element	Associated persona	Principles	Audit focus
Model deployment	• Safe and secure	• AI developers • Cybersecurity experts	• Security of users' data • Adequate controls to prevent the possibility of a malicious attack • Ensuring the safety and security of all the stakeholders • Usage of on-device processing whenever possible	• Review security protocols and data safety measures • Validate measures for preventing attacks • Assess on-device processing implementation
Model monitoring	• Robust and reliable	• AI/ML engineers • Independent AI assurance and audit team	• Live monitoring in production to ensure that the AI system is operational • Ability to trace, diagnose and rollback, if necessary, in case of a failure • Disaster recovery and business continuity plans • Resiliency of AI systems	• Assess the efficacy of live monitoring and diagnostic capabilities • Verify the existence and effectiveness of disaster recovery and business continuity plans

Need for governance structure across the AI lifecycle

To ensure AI development and deployments, it is essential to follow the ethical principles defined by the enterprise AI policy. A governance structure at various levels ensures that AI systems are developed, deployed, and maintained responsibly, ethically, and transparently. Following is a basic outline of an AI governance structure:



AI tracking

Alignment with the Trustworthy AI framework, establishing clear role assignments and responsibilities, and outlining essential life-cycle criteria. This helps maintain uniformity across the organisation, along with an updated inventory that includes attributes for the risk management programme.

Lifecycle standards

Well-defined rules, tools, and technology are needed to implement the AI policy at every stage. They can be changed to fit different situations, such as using AI from other sources or creating new AI. This way, distinct functions of the company can adjust requirements as needed.

Risk assessment and measurement

Methods to measure and understand AI risks using numbers and information, e.g., metrics design and monitoring, risk categorisation, meaningful reporting, and analytics.

Regulatory

Ability to adjust according to various regulations set by different regulators in different countries/regions. When it makes sense, these adjustments should be added gradually to the existing programmes to manage risks related to models, data, cybersecurity, and legal matters.

In India, we do not have any regulations on AI for the development, classification, and use of non-personal and personal data in the public domain.

In the recent B20 summit (G20 Business Forum) in India, the B20 task force recommended setting up a regulatory framework for responsible AI, and the Indian government called for a global AI framework to promote the ethical development of AI.

Below are a few key considerations for setting up an effective governance structure for AI that could mobilise the people for AI governance.

- Define goals and articulate objectives.
- Set up an ethics statement.
- Establish guardrails to guide, monitor, and assess AI solutions. For example, embedded controls in the AI model could prevent specific actions from being completed.
- Define roles and responsibilities for the people responsible for the governance, development, deployment, management, and monitoring.
- Set up an inventory of AI models and procedures for tracking and maintaining AI implementations.
- Create role-specific upskilling of stakeholders and employees to guide on AI solutions and their responsible development and deployment.
- Define or optimise the existing data governance for the data.
- Develop KPIs to evaluate the AI models' performance.

Way forward

Maintaining trust in AI necessitates continuous monitoring of AI models to ensure they function as intended and align with trust criteria. This is particularly challenging with opaque AI models.

Adequate awareness of AI Risk Management across the entire AI lifecycle and relevant stakeholders along with leveraging AI Risk Management solutions to assess and validate model performance can restore balance in transparency and accuracy.
Beyond model evaluation, AI data management, privacy, cybersecurity, and post-deployment monitoring also benefit from such solutions. These tech-enabled assessments enhance AI evaluations, fostering better governance and understanding of model performance for comprehensive AI management.

Connect with us

Anthony Crasto
President, Risk Advisory
Deloitte India
acrasto@deloitte.com

Peeyush Vaish
Partner, Risk Advisory
Deloitte India
peeyushvaish@deloitte.com

Nitin Naredi
Partner, Risk Advisory
Deloitte India
nitinnaredi@deloitte.com

Samanth Aswani
Partner, Risk Advisory
Deloitte India
saswani@deloitte.com

Key contributors

Manish Dayma

Bharath Yellapu

Adarsh Mishra

Sachin Arora

Acknowledgment

Akshay Dalvi

Neha Kumari

Deloitte.

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms, and their related entities (collectively, the "Deloitte organization"). DTTL (also referred to as "Deloitte Global") and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see www.deloitte.com/about to learn more.

Deloitte Asia Pacific Limited is a company limited by guarantee and a member firm of DTTL. Members of Deloitte Asia Pacific Limited and their related entities, each of which is a separate and independent legal entity, provide services from more than 100 cities across the region, including Auckland, Bangkok, Beijing, Bengaluru, Hanoi, Hong Kong, Jakarta, Kuala Lumpur, Manila, Melbourne, Mumbai, New Delhi, Osaka, Seoul, Shanghai, Singapore, Sydney, Taipei and Tokyo.

This communication contains general information only, and none of DTTL, its global network of member firms or their related entities is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

No representations, warranties or undertakings (express or implied) are given as to the accuracy or completeness of the information in this communication, and none of DTTL, its member firms, related entities, employees or agents shall be liable or responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication.