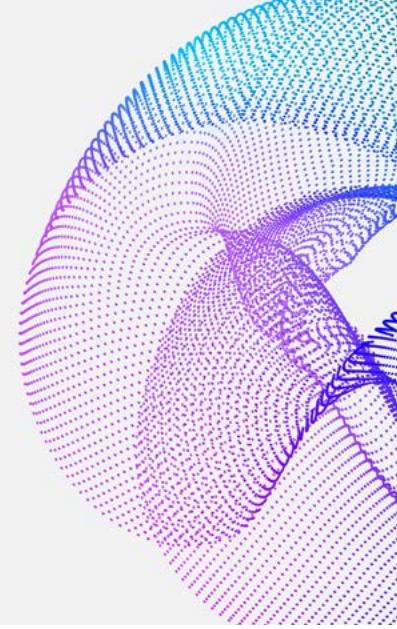# Achieving trustworthy AI

November 2020

A model to design, manage and use trustworthy AI systems to enable the long-term value of AI to be realised and captured.

An AI system refers to all components and lifecycle stages of a data driven solution including design, data, algorithms and processes. The development, procurement, deployment and governance of trustworthy AI systems requires a connected, firm wide approach that aligns, integrates and connects key functional areas of the organisation. The six segments of the model represent soft, permeable boundaries.
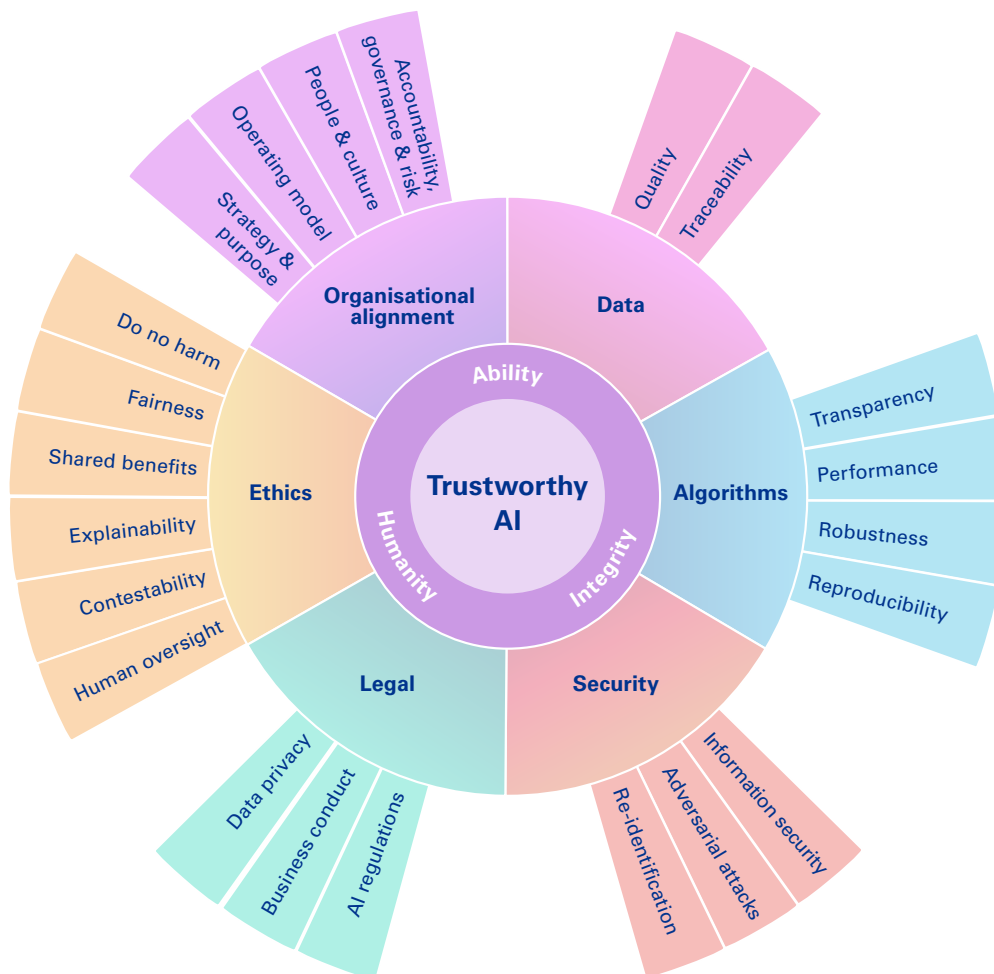
A model to design, develop, deploy, procure and operate trustworthy data driven systems and their components.

Vulnerability	What does good look like?

Organisational Alignment

Strategy & Purpose	The purpose, design and use of AI systems align with the organisation's strategy, purpose and values, and are designed to engender trust.
Operating Model	Resourcing, processes, policies and operational systems are developed and updated to execute the organisation's AI strategy.
People & Culture	The right people, capabilities, knowledge and diversity, and cultural practices are in place to achieve trustworthy AI.
Accountability Governance & Risk	The chain of accountability and responsibility for the AI system (including governance of data and algorithms) across key stages of its lifecycle are clearly defined, structured and understood across the organisation, and efficiently executed.

Data

Quality	Data availability, usability, consistency and integrity are assessed to ensure data is suitable for informing the inferences produced by the algorithm, and are sufficiently comprehensive to produce accurate and reliable outcomes.
Traceability	The source and lineage of data within the system is known, documented, traceable and auditable.

Algorithms

Transparency	The technical features of the algorithm are documented and designed to enable understanding of how the model works and arrives at its solutions.
Performance	The integrity and accuracy of algorithms and processes are assessed before deployment based on valid metrics to ensure it operates as intended.
Robustness	The overall solution and processes are tested to ensure the same performance – confirmed during the development – is preserved despite possible changes in the environment during its operations. Ongoing performance monitoring and appropriate corrective action is taken across the system's lifecycle.
Reproducibility	An audit trail of documentation, evidence and logs is kept to reproduce prior results as needed.

Security

Information security	Robust and clear information security and access protocols are in place to ensure the confidentiality, integrity, access and availability of data is protected throughout the data and AI lifecycle.
Adversarial attacks	Robust cyber security measures are in place to identify and prevent adversarial machine learning attacks, hacking and other types of cyber-attacks that may compromise the performance of the AI system, breach human and legal rights, and result in unfair outcomes.
Re-identification	The risk of malicious actors re-identifying individuals by combining anonymised data with other sources is effectively identified and managed.

Legal

AI regulations	Local and global, soft and hard regulations and legislative frameworks relating to data and AI are understood and consistently adhered to across the organisation. Changes are dynamically monitored.
Data Privacy	Privacy impact assessments and procedures are in place to ensure legal compliance and stakeholders' ethical privacy expectations are met.
Business Conduct	Business conduct regulations are pro-actively identified to ensure AI systems are compliant.

Ethics

Do no harm	The risks, unintended consequences and potential for harm of an AI system are fully assessed and mitigated prior to, and during, its deployment. Particular care is given to human rights and vulnerable stakeholders.
Fairness	The outcomes of AI systems are regularly monitored to ensure they are fair, free of unfair bias and discrimination, and designed to be inclusive for diverse stakeholders.
Shared benefit	The AI system is designed to benefit a range of stakeholders, including customers, employees and end users.
Explainability	The purpose of the AI system, how it functions and arrives at its solutions, and how data is used and managed, is transparently explained and understandable to a variety of stakeholders.
Contestability	Any impacted user or stakeholder is able to challenge the outcomes of an AI system via a fair and accessible human review process, with clear mechanisms for remediation where appropriate.
Human Oversight	There is appropriate human oversight and control of AI systems and their impact on stakeholders by people with sufficient knowledge and AI literacy to ensure informed engagement, decision making and risk management.

Key contacts

University of Queensland

Nicole Gillespie
KPMG Chair in Organisational Trust
Professor of Management,
The University of Queensland
T: +61 7 3346 8076
E: n.gillespie1@uq.edu.au

KPMG

James Mabbott
National Leader, KPMG Innovate
KPMG Australia
T: +61 2 9335 8527
E: jmabbott@kpmg.com.au

Richard Boele
Global Leader, Business
& Human Rights Services
KPMG Australia
T: +61 2 9346 585
E: rboele@kpmg.com.au

Jon Stone
Partner, KPMG Digital Delta
KPMG Australia
T: +61 3 9288 5048
E: jonstone@kpmg.com.au

Zoe Willis
National Leader, Data & RegTech
KPMG Australia
T: +61 2 9335 7494
E: zoewillis@kpmg.com.au

Dr Sanjay Mazumdar
Chief Data Officer
KPMG Australia
T: +61 8 8236 7237
E: skmazumdar@kpmg.com.au

Ali Akbari
Artificial Intelligence
Capability Lead
KPMG Australia
T: +61 2 9335 7740
E: aakbari@kpmg.com.au

Rossana Bianchi
Associate Director,
Strategy, Growth & Digital
KPMG Australia
T: +61 2 9335 7036
E: rbianchi2@kpmg.com.au

Rita Fentener van Vlissingen
Associate Director, Human
Rights & Social Impact
KPMG Australia
T: +61 2 9346 6366
E: ritafentener@kpmg.com.au

KPMG.com.au