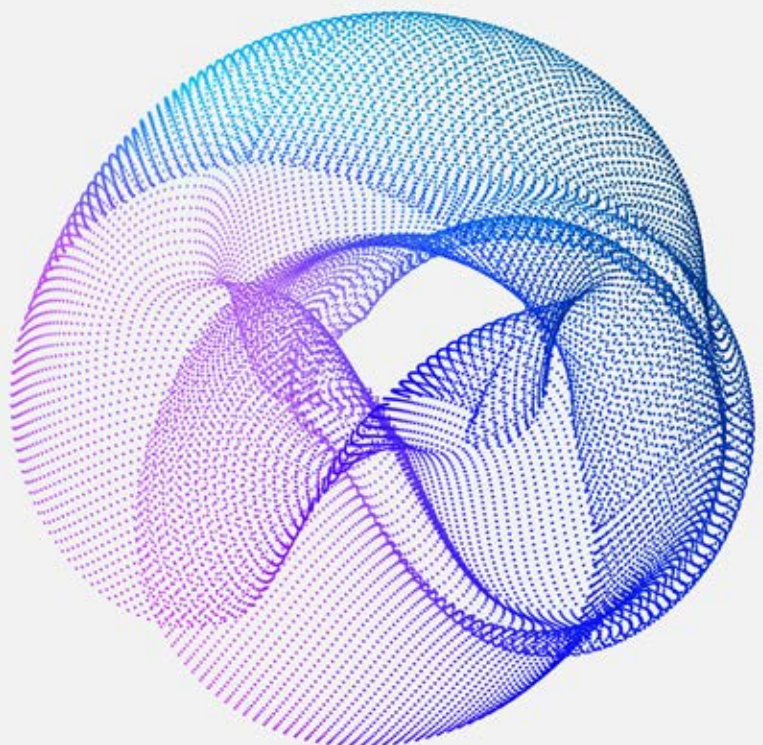




# Achieving Trustworthy AI

**A Model for Trustworthy Artificial Intelligence**



November 2020

---

[KPMG.com.au](https://www.kpmg.com.au)

# Contents

Foreword.....	1
How does an organisation go about achieving trustworthy AI? .....	2
A model for achieving trustworthy AI.....	6
Key principle Organisational Alignment.....	8
Key principle Data.....	14
Key principle Algorithms.....	18
Key principle Security.....	22
Key principle Legal.....	26
Key principle Ethics .....	30
Conclusion.....	36
Endnotes.....	37

## Acknowledgements

This report was informed by the insights and expertise shared through interviews with James Mabbott, Richard Boele, Alison Kitchen, Jon Stone, Michael Hill, Sally Calder, Mike Kaiser, Andrew Yates, Sanjay Mazumdar, Jane Gunn, Kate Marshall, Robert Warren, Zoe Willis, Vanessa Wolfe-Coote, Cath Ingram, Dr. Michelle Perugini, Sarah Haynes, and Associate Professor Mark Burdon

## University of Queensland Authors

Nicole Gillespie and Caitlin Curtis

## KPMG Authors

Rossana Bianchi, Ali Akbari and Rita Fentener van Vlissingen

## Key Contacts

James Mabbott, Rossana Bianchi, Ali Akbari, Nicole Gillespie, Rita Fentener van Vlissingen, Sanjay Mazumdar, Zoe Willis, Jon Stone and Richard Boele

## Citation

Gillespie, N., Curtis, C., Bianchi, R., Akbari, A., and Fentener van Vlissingen, R. (2020). Achieving Trustworthy AI: A Model for Trustworthy Artificial Intelligence. KPMG and The University of Queensland Report. [doi.org/10.14264/ca0819d](https://doi.org/10.14264/ca0819d)

# Foreword

Australian business and their leaders are continuing to invest in the steps needed to rebuild and retain public trust. COVID-19, despite its many diabolical effects, has provided an opportunity for us to demonstrate how we can navigate significant change and disruption with ability, integrity and humanity.

Our collective reliance on digitisation and new technologies has grown significantly through the pandemic as we embraced new ways of connecting and doing business. Organisations, leaders and their teams across the nation have swiftly transformed the way they work. Corporate cultures have become more flexible, more agile and often, more caring. I believe this has helped strengthen trust in Australian business.

Leaders should remember the fluidity they are capable of in crisis and retain that ethos as we tackle the next trust challenges: those emerging as we embrace the technology underpinning the Fourth Industrial Revolution. Investment in, and adoption of, Artificial Intelligence (AI) has continued to increase at an exponential rate as organisations and governments around the world realise the value and competitive advantage of AI. We know that trust is a key enabler and accelerator of innovation and it is critical to the ongoing acceptance and adoption of AI. But are our customers, employees and other stakeholders ready and willing to trust AI systems and the organisations that deploy them?

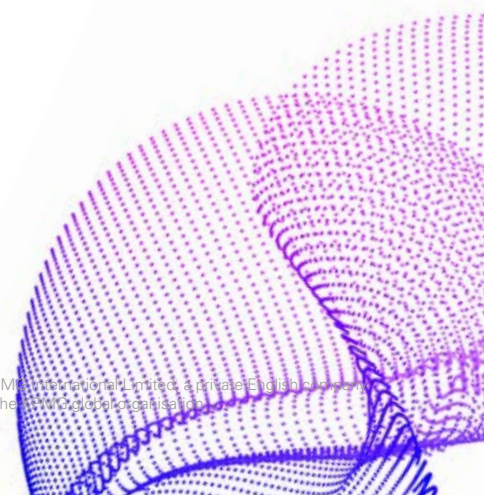
Concerns surrounding privacy violations, unintended bias and discrimination, as well as harmful or inaccurate outcomes, are fuelling a lack of trust in AI. Our recent UQ-KPMG national survey showed that trust in AI systems is low in Australia, with only one in three Australians willing to trust AI systems. Almost half of our community (45 per cent) reported being unwilling to share their information or data with an AI system, and two in five were unwilling to trust the recommendations and output of AI.

So how does an organisation go about achieving trustworthy AI systems? That is the question we address in this report, led by KPMG and Professor Nicole Gillespie from the University of Queensland Business School.

If we are to realise the promising societal benefits and economic opportunities that could be derived from AI, we need to better understand how to address the risks to people associated with AI systems. The organisations that adopt an integrated, cross-disciplinary approach to achieving trustworthy AI, as demonstrated in this report, will be the organisations able to manage reputational risk and lead the responsible stewardship of this technology.



**Alison Kitchen**  
Australian Chairman  
KPMG Australia





# How does an organisation go about achieving trustworthy AI?

In this report we set out an integrative model of the key principles and practices required, including practical guidance and examples of implementation.



## What is AI?

Artificial Intelligence (AI) refers to computer systems that can perform tasks or make predictions, recommendations or decisions that usually require human intelligence. AI systems can perform these tasks and make these decisions based on objectives set by humans but without explicit human instructions<sup>1</sup>.

## The promise of AI

AI is reshaping the competitive landscape across all sectors of the economy.

It's helping organisations make better predictions and more informed decisions, while lowering operating costs, facilitating productivity gains and driving new business models.

AI is helping us address some of humanity's most complex problems, for example:

- In financial services, AI is used to improve fraud detection and anti-money laundering processes.
- In agriculture, AI helps monitor crop and soil health and predicts the impact of environmental factors on crop yields.
- In retail, AI is enhancing the customer experience through the rapid visualisation of product layouts across stores.
- In transnational supply chains, AI tools are using satellite imaging data to map forced labour patterns and predict modern slavery hotspots.
- In healthcare, AI technology is enhancing the accuracy of medical diagnosis and improving the effectiveness and speed to market of life-saving treatments such as precision medicine.
- AI is helping the fight against COVID-19 by simulating and predicting spread patterns to inform government responses, enhancing diagnosis and helping detect mutations in the virus.

In 2019, global spending on AI systems was \$37.5 billion, and is predicted to reach US\$97.9 billion in 2023<sup>2</sup>. AI is expected to generate nearly US\$4 trillion in added value by 2022<sup>3</sup>. AI-centred start-ups attracted 12 per cent of worldwide private equity investments in the first half of 2018, up from just three per cent in 2011<sup>4</sup>.

## The risks of AI

With this rapid growth comes greater awareness of the risks associated with AI systems, including privacy violations, unintended bias and inaccurate outcomes.

High profile scandals involving AI have reduced public trust in the new technology.

Some AI technologies have been accused of reinforcing and codifying unfair biases. For example, an AI-based recidivism prediction tool, COMPAS, raised concerns around accuracy and racial bias in its decision-making recommendations<sup>5</sup>.

AI has helped spread fake or manipulative online content, including tools used to micro-target political advertising in the 2016 US Presidential election<sup>6</sup>.

And some applications of AI and automated decision making have produced inaccurate, unfair or harmful outcomes. In Australia, Centrelink used automated decision making – dubbed robodebt – to calculate and recover welfare overpayments. Errors resulted in harm to citizens and financial and reputational damage for the government, including a class action<sup>7</sup>.

AI can undermine human rights, such as privacy and autonomy, by facilitating mass surveillance programs, including facial recognition. AI could also precipitate technological unemployment.

Our recent national survey, Trust in Artificial Intelligence: Australian insights<sup>8</sup>, showed that trust in AI systems is low in Australia. Almost half of Australians are unwilling to share their information with an AI system, and two in five are unwilling to trust the recommendations or output of AI. This general suspicion will slow the potential advance of AI.

## How AI challenges trust



### Explainability

The complexity of machine learning from large datasets make it difficult if not impossible for humans to understand how the AI arrived at its outcome.



### Data privacy and security

AI learns from massive datasets, raising concerns and risks about privacy, data security, appropriate use of data and consent.



### Bias

Poor quality training data or incomplete data can cause AI output to reflect historic biases.



### Human agency and control

The self-learning capability of AI, coupled with its powerful analytic capability, raises concerns around retention of human control and agency.

## The value of trustworthy AI

Trustworthy AI has three key components. When people believe an AI system adheres to these components, they are more likely to trust in the system.

Trust underpins the acceptance and use of AI. For AI systems to work, there must be a willingness to be vulnerable to the systems, through sharing data or relying on automated AI decisions. This trust is built on positive expectations of the ability, humanity and integrity of the systems, and those developing and deploying the systems.



### Ability

AI systems are fit-for-purpose and perform reliably to produce accurate output as intended.



### Integrity

AI systems adhere to commonly accepted ethical principles and values (e.g. fairness, transparency of data collected and how it is used), uphold human rights (e.g. privacy), and comply with applicable laws and regulations.



### Humanity

AI systems are designed to achieve positive outcomes for end-users and other stakeholders, and at a minimum, do not cause harm or detract from human well-being.

Most Australians (56 per cent) agree that AI systems produce reliable, accurate output (ability), but most (67 per cent) are unconvinced that AI systems operate with integrity and humanity<sup>9</sup>.



## AI is helping restore stakeholder trust through remediation in financial services

AI systems are helping the financial services industry to increase the efficiency, quality and auditability of their remediation of customers following the 2019 Royal Commission into Banking and Financial Services. The AI system categorises the huge volume of data into a digestible form that humans can review. This enables faster compensation for vulnerable people for past breaches – a vital step in restoring trust.

“AI makes the job of the humans involved much easier. Without it, we couldn’t do the work and create a proposition that would be effective,” says Andrew Yates, KPMG National Managing Partner, Audit, Assurance & Risk Consulting.

---

“If we can’t create strong, high-trust environments, AI technologies will cost more and be slower to market.”

**James Mabbott**  
National Leader, KPMG Futures

# A model for achieving trustworthy AI

There is no silver bullet for achieving trustworthy AI systems in practice.

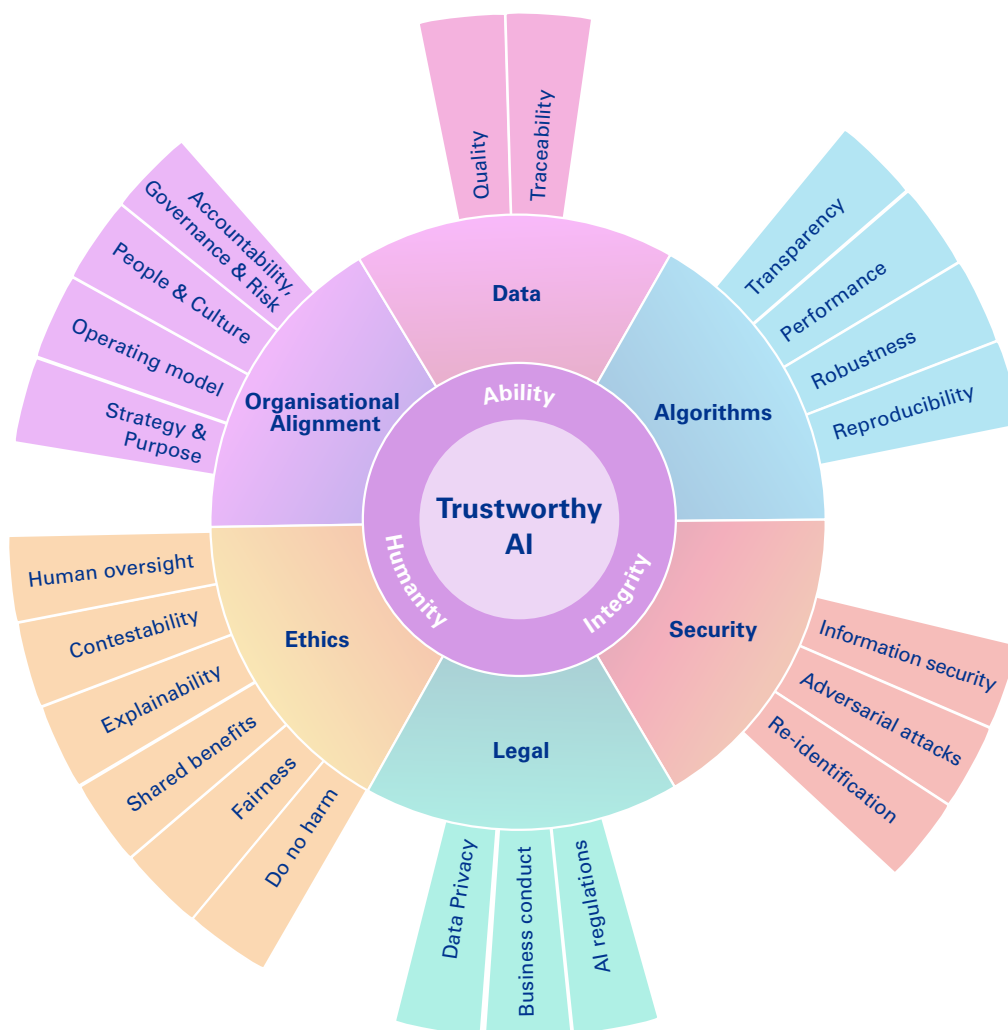
The benefits, challenges, risks and opportunities that AI offers differ from one industry and application to another. Organisations therefore need to tailor their approach and ensure it is proportional to the potential risks and impacts the AI systems pose to their stakeholders.

This requires a business-wide approach that integrates and connects key functional areas of the organisation.

Our model identifies the six dimensions that need to operate in a connected way to ensure trustworthy AI across the AI lifecycle.

It outlines key principles and practices for addressing vulnerabilities that can undermine trust in AI systems and the organisations deploying them.

It then lays out the work required to align, leverage and enhance existing organisational infrastructure and governance mechanisms.



## The Trustworthy AI Model

A model to design, develop, procure, deploy and govern trustworthy data driven systems and their components, including design, data, algorithms and processes.



## 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.

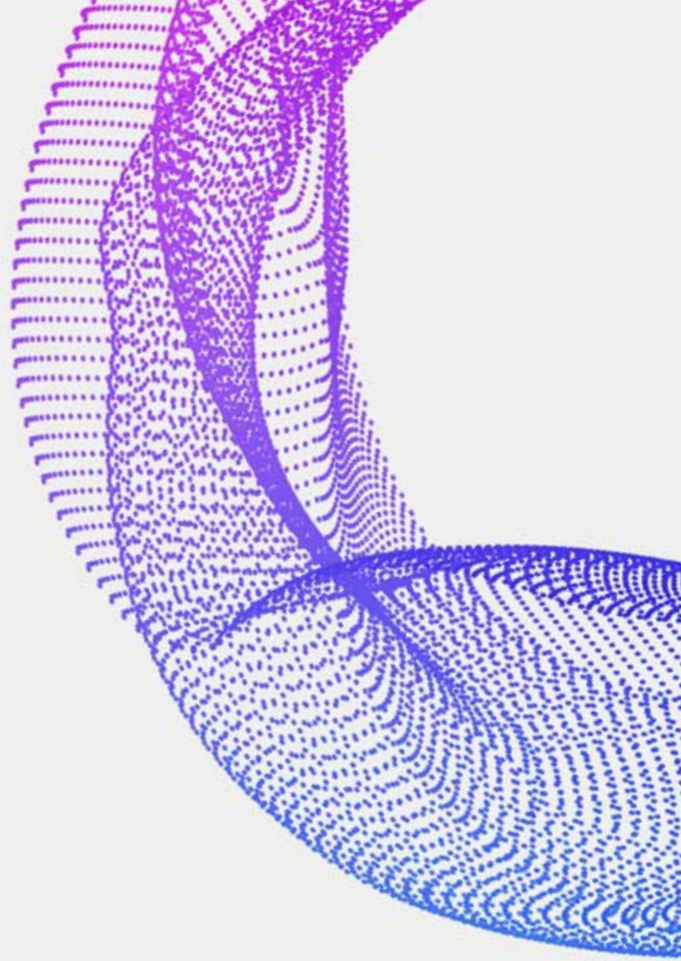
### Ethics

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 benefits	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 principle**

# Organisational Alignment



**The priority is to ensure that any AI system  
being designed, procured or implemented  
is aligned with the organisation's strategy,  
core purpose and values.**



## Align AI systems with strategy and purpose

The challenging aspects of AI, for example around bias and fairness, have the potential to significantly undermine trust if they contradict the organisation's values and the rights of stakeholders. For example, one large tech firm abandoned its AI recruitment system after it taught itself to favour males for technical roles – a clear violation of the firm's commitment to diversity and equality.

When used responsibly to support the organisation's purpose and create value for stakeholders, AI can enhance trust by demonstrating ability, humanity and integrity. For example, many financial institutions are now using AI to significantly improve the detection of credit-card fraud and money-laundering activities.

### Key considerations

- Develop a clear AI strategy and vision that articulates how the firm's use of AI will be trustworthy throughout the AI lifecycle and will support the organisation's broader purpose, strategy and values.
- Ensure the strategy for trustworthy AI is understood across the firm, including by non-technical employees.
- Support employees to raise concerns about AI-enabled products or services that may undermine the organisation's values or create trust issues.
- Involve a diverse number of customers and end-users in the design and testing of AI systems prior to release.
- Support stakeholders' understanding of how and when AI is being used in the organisation's products and services to demonstrate transparency and commitment to trustworthy AI.



Australians (57-76 per cent) believe organisations innovate with AI for financial reasons (e.g. cost saving or profit maximisation) rather than to benefit society more broadly (35-44 per cent). This imbalance is most pronounced for business, followed by government and then non-profit organisations<sup>10</sup>.

## Develop an operating model to support trustworthy AI

While AI strategy sets the direction, the operating model creates the map to arrive at the destination. A business' operating model should set out the appropriate processes and policies to ensure an efficient, connected and responsive organisation-wide approach.

Traditional IT operating models are often not equipped for the trust and ethical challenges posed by AI. They will require a new or reconfigured set of systems and processes that allow for continuous monitoring and sign-off by relevant internal stakeholders. For example, best practice processes for responsible procurement of AI systems differs markedly to those used for procuring IT systems and products.

### Key considerations

- Assess the organisation's readiness to adopt the trustworthy AI strategy, then develop and resource a change management plan.
- Establish working groups to critically review and revise policies, processes and systems.
- Establish reliable fail-safe processes and back-up solutions in case of an automated process failure, including making changes to existing IT and operational systems.

“Design is not just for technologists – you need multidisciplinary teams considering how AI should be validated and delivered.”

**Vanessa Wolfe-Coote**  
Partner, KPMG Strategy

## Ensure the right people and culture

The successful deployment of trustworthy AI relies on people and a diversity of perspectives. This is particularly important for managing the opportunities and risks of AI.

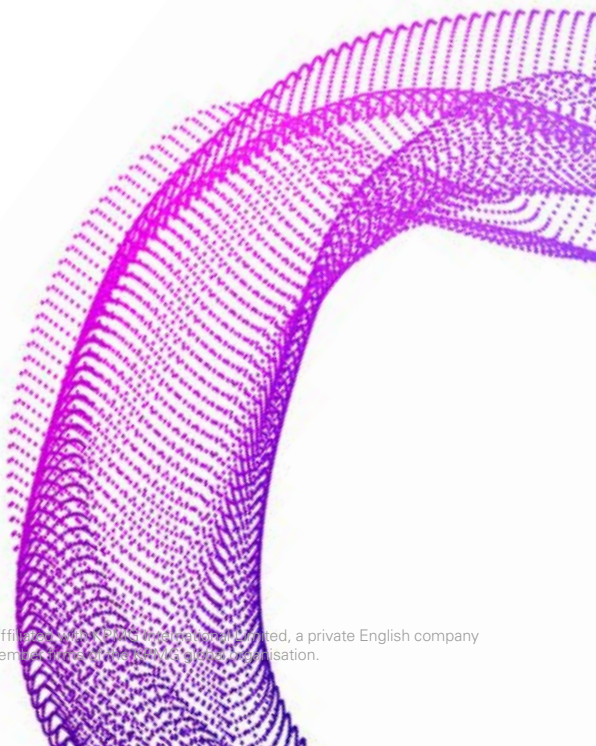
Developing and reinforcing a culture which values trustworthy AI provides the foundation for effective communication and coordination of AI implementation. This, in turn, will lead to better prevention, early detection and resolution of trust issues before they escalate and create reputational damage.

### Key considerations

- Align recruitment, learning and development to the needs of the AI strategy.
- Equip employees with the right knowledge and tools to operationalise and embed data and AI into relevant business processes and practices in a trustworthy way.
- Align key performance indicators and remuneration systems to incentivise the right behaviours.
- Create diversity and inclusion through interdisciplinary, cross functional and cross hierarchical working groups on AI.
- Role model the right cultural tone from the top by ensuring visible leadership, commitment and buy-in on trustworthy AI.

“AI cannot be imposed without a thoughtful way of engaging people, both customers and employees. They have to be brought to a clear understanding so they can trust it”

**Jane Gunn**  
Partner in Charge,  
People and Change



## Establish accountability governance and risk mechanisms

AI accountability refers to the expectation that organisations will ensure the proper functioning of AI systems in accordance with their roles and applicable regulatory frameworks<sup>11</sup>.

Organisations developing or using AI systems, whose outcomes may impact on people, need to carry out risk and impact assessments and put in place appropriate risk management processes. Where possible, organisations should leverage existing governance and risk frameworks and mechanisms, adapting these to cater for the expanded risks to the organisation and potential impacts on people from AI systems.

Establishing interdisciplinary governance boards to assess and govern AI-enabled operations, products and services is now best practice. For example, Mastercard established a governance council to review and approve the implementation of AI applications determined to be high risk. The council is chaired by senior executives, as well as data scientists and representatives from different business teams.

### Key considerations

- Adopt a code of conduct or charter that embeds shared values and principles to support ethical and trustworthy data use and AI.
- Ensure responsibility and accountability is clearly defined, allocated, understood and executed across key stages of the AI lifecycle.
- Develop internal governance, monitoring and reporting structures that provide appropriate oversight of how AI systems and technologies are brought into the organisation's operations, products and/or services<sup>12</sup>.
- Transparently document who can, and is, making key decisions throughout the AI system lifecycle.
- Carry-out an initial risk assessment and scoring to determine an AI project's level of risk to business and to stakeholders upfront and ensure the appropriate level of governance oversight and remediation is applied.
- Establish transparent and accessible processes for employees, customers and other stakeholders to report potential risks, biases or vulnerabilities in the AI system.
- Where AI systems are operating in critical functions with high risks to people, potentially impacted communities should be engaged, with a focus on the most vulnerable and marginalised stakeholder groups.
- Consider a staged release of new algorithms that have the potential to impact many, to enable robust assessment of potential impacts prior to broader release.
- Review communication channels and interactions with stakeholders of AI systems to provide disclosure



### Managing Risk: How can standards and certifications help?

Standards and certifications can facilitate the widespread adoption of trustworthy AI and help reduce risks. They can also enhance public trust by giving assurances that products that hold the certification meet technical performance or ethical standards.

Standards work in AI is being developed by international bodies such as ISO and IEEE. For example, the IEEE P7000 series of standards projects, whose stated aim is to develop standards inclusive of both technological and ethical considerations.



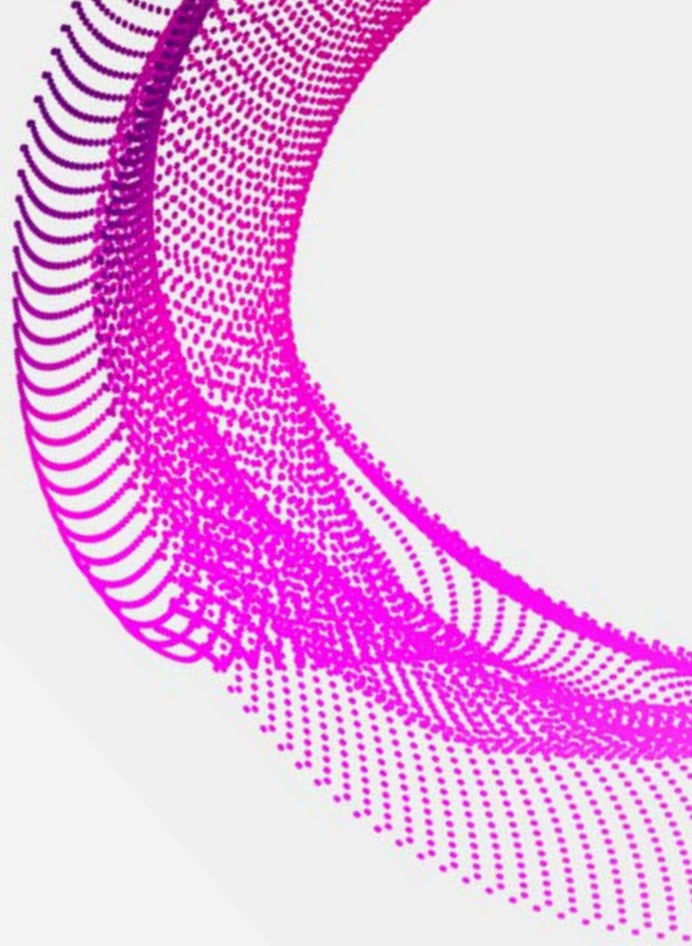




**Key principle**

Data





**AI systems learn from their input and training data. If an AI system is built on incomplete, biased or otherwise flawed data, the mistakes will likely be replicated at scale in its outputs.**



## Data

A prominent example of this pitfall, and its ensuing damaging impact on community trust, is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system<sup>13</sup>. This tool has been widely used in the United States to predict a defendant's likelihood of committing future crimes which then influences parole decisions. However, the system was found to be unreliable, due at least in part, to biases in the data used to build it.

This is not an isolated occurrence. Similar issues can arise in any AI system when the data cannot be traced confidently to its source, or when systematic quality tests are not carried out before using the data as building blocks for a new system.

In Australia, the Law Enforcement Conduct Commission (LECC) identified serious concerns about the potential for Aboriginal and Torres Strait Islander children to be

disproportionately targeted under a repeat offender monitoring scheme by New South Wales police<sup>14</sup>. If this data had been used to build an AI system before this serious issue was identified, another failure like COMPAS could have eventuated.

Such trust failures can be prevented by following best practice in assessing the quality and traceability of the data used to build AI.

## The importance of quality data



### Key considerations

- Conduct statistical tests to ensure the quantity, characteristics and representativeness of the data is suitable for the intended purpose.
- Consider, confirm and verify the ethical dimensions of your dataset, to ensure that it is diverse and does not result in unfair bias using relevant statistical tests (see section on ethics).
- Ensure all external sources of data are suitable, reliable and available in the required form and meet expected standards, based on a comprehensive knowledge of the origin and integrity of the data.
- Confirm the timeliness for the intended use and its consistency with other existing data bases.

## Ensuring data is traceable



### Key considerations

- When incorporating data into the system, keep a copy of the data in its original form along with relevant information about it – e.g. its source, time of collection, etc.
- Maintain all required data management information including the purpose, copyright, access and privacy permissions, as well as the ownership, or allowed lifetime of the data within the system.
- Keep track of how the data was collected, curated, and all process steps taken to transform and move it within the organisation, including clear documentation and version control over the processes applied to the data during its lifecycle.








**Key principle**

# Algorithms





**Machine learning algorithmic models are  
the heart of an AI system's power to make  
predictions and decisions.**

## Algorithms

Algorithms are one of the most complex components of an AI system. As such, sufficient expertise is required to minimise the potential risks and detect errors that can result from its misuse. The complexity of algorithms and approaches to building AI systems can easily cause mistakes and generate undesired results, even from good data.

For example, in the United States, an AI system was designed to automatically assign risk scores to patients to be used for referral into programmes to provide more personalised care. A complex suite of intertwined issues

was identified, involving financial, health and racial factors, which resulted in racial bias embedded in a health-care algorithm. The outcome was systemic discrimination which affected millions of African Americans<sup>15</sup>.

To prevent such bias and inaccuracies, and ensure the trustworthiness of the AI systems, best practice in assessing the performance, robustness, reproducibility and transparency of algorithms is required.



Most Australians (51-55 per cent) have high or complete confidence in Australian universities and the Australian Defence Forces to develop and use AI in the public's best interest. Only about a third of Australians (34 per cent) have high or complete confidence in technology companies, and a little over a quarter in Federal and State governments. Australians have the least confidence in commercial organisations to develop and use AI in the public's interest<sup>16</sup>.

## Open the box and make it transparent

The technical features of the algorithm should be documented and designed to enable understanding of how the end-to-end process works, and how it arrives at its outcomes.

### Key considerations

- Organisations should be meaningfully transparent about how the data is being used when an automated solution is in place.
- Clear explanation of the algorithm behaviour and the logic behind its design – without excessive technical complexities – is essential to providing a reasonable explanation when required.
- Any manual or automated steps from sourcing the data through to generation of the outcome should be visible and accessible to understand.

## Define and confirm appropriate performance levels

System developers must identify the right metrics to assess the systems performance and ensure it operates as intended prior to the deployment.

### Key considerations

- Appropriate performance targets should be set based on the sensitivity and use of the AI system.
- Prior to deployment and use, effective performance metrics need to be defined to ensure targets are achieved.
- Specialised tests should be performed to ensure outcomes are free from unfair bias.



## Establish robust monitoring systems to ensure expected performance is maintained

Outcomes and processes should be regularly tested to ensure that the same performance that was established and confirmed during the system development is upheld, despite possible changes in the environment that might occur during the system's operations. Monitoring processes should be specifically designed for each algorithm and deployed within the AI system to ensure that the desired performance levels are sustained during operations.

### Key consideration:

Correction mechanisms and/or fall-back options should be built into the system to detect and correct underperformance – or put alternative appropriate processes in place until human intervention can rectify the issue.

## Keep an audit trail to enable reproducibility of results

Just knowing the traces and origin of the data is not enough to be able to repeat and confirm the same process. It is essential to maintain and store an audit trail of documentation, evidence and logs of the development of the algorithm to reproduce results if and when required.

### Key consideration:

Maintain an appropriate version control system, documentation of development and history for all components of the system including code, training data, trained models and parameters, tools and platforms used for the development – along with their configurations.

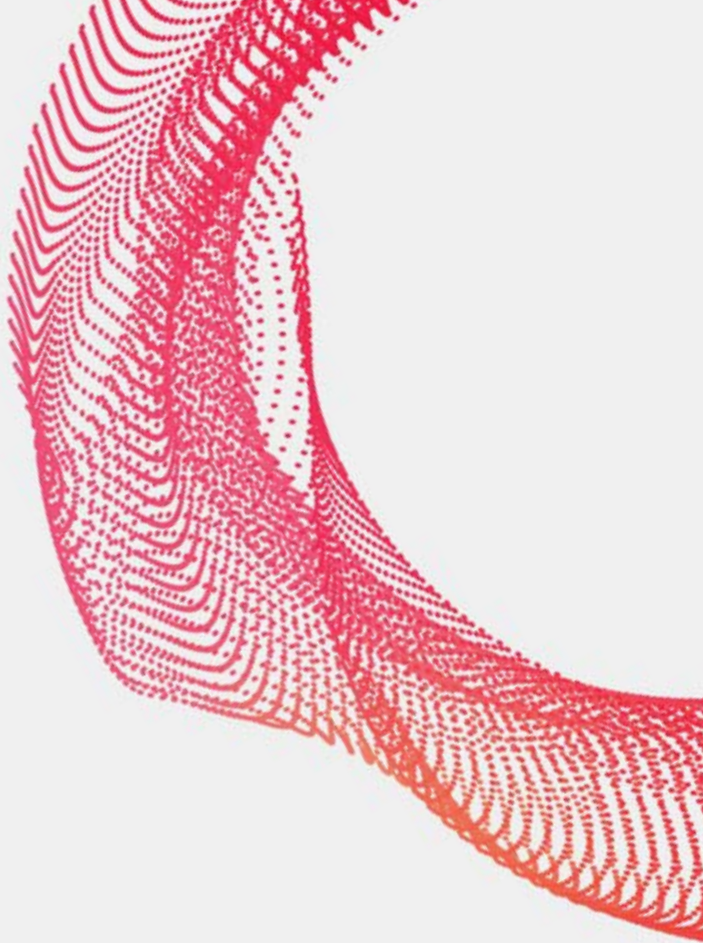






**Key principle**

Security



**The amount of data we each generate is rapidly increasing and this trend has accelerated during the COVID-19 pandemic due to the rapid uptake of remote working. Unlike conventional software solutions, intelligent algorithms at the heart of AI systems introduce vulnerabilities that require special consideration. Stakeholders need to be confident the integrity of the system will be kept at a high standard both by people within the organisation and protected from potential external malicious activities.**

## Ensure robust information security

Robust and clear information security and access protocols can help increase confidence in protecting confidentiality, integrity and availability of data (the 'CIA triad') throughout the data and AI lifecycle.

### Key considerations

- Proactively determine and document the locations, confidentiality, integrity and availability requirements of the systems and information.
- Consider preventative controls such as network segmentation and additional access controls.
- Use encryption for the data at rest and in transit.
- Back up all important data on a regular and proven basis.
- Avoid keeping unnecessary sensitive data for longer than the required period.



Of the several societal challenges of AI, Australians believe data security challenges such as fake online content (70 per cent) and cyber attacks (67 per cent) are most likely to impact large numbers of Australians over the next 10 years.

Australians have a clear expectation that companies and government will carefully manage and prevent cyber attacks and data breaches<sup>20</sup>.

## Protect against adversarial attacks

Cyber security is now one of the most pressing concerns for business and the public<sup>17</sup>. Of the reported data breaches in 2020, 61 per cent were caused by malicious or criminal attack.

The action taken against Microsoft's intelligent chatbot, Tay, is a good example of a malicious attack<sup>18</sup>. A systematic attack by a subset of people feeding it targeted messages and material changed its behaviour to tweet inappropriate and reprehensible words and images. In a more recent example, through an evasion attack last year, researchers identified a way to trick Cylance AI antivirus into accepting malware as safe files<sup>19</sup>.

Robust cyber security measures need to be 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.

### Key considerations

- Create close collaboration between cyber security professionals, AI and machine learning experts.
- Constantly monitor and periodically reassess algorithms – AI systems are not designed to set-and-forget.
- The more attackers know about your AI system design and logic the better they can plan an attack, so the trade-off between transparency and protecting against adversarial attacks needs to be considered and balanced.
- Minimise the chance of unauthorised access to your training data and always reconfirm its integrity before feeding it into models.
- Be conscious of fake and engineered data and design and use as many filters as possible, especially when the training or online learning is based on public data.
- Consider the potential adversarial vulnerabilities during the design stage and build solutions when developing the algorithms.



## Mitigate re-identification risks

The risk of malicious attacks trying to re-identify individuals by combining anonymised data and other sources should be taken seriously and managed efficiently. This is one of the most common issues damaging the public's trust in organisations collecting personal data or the systems that process them.

For example, when myki (the Victorian public transport payment card system) shared people's anonymised travel data they didn't realise that, in conjunction with social media data, it might reveal identities and cause a privacy breach<sup>21</sup>.

## Key considerations

- Minimise the sharing of personal data even if de-identified, and have appropriate policies to ensure safe use and access.
- The OAIC guidelines<sup>22</sup> stress “there is no one right way to de-identify data” so the best techniques and processes should be carefully chosen each time based on the context.
- The risk of reidentification should be actively assessed and managed for each case (the De-Identification Decision-Making Framework<sup>23</sup> is a helpful tool to assess and manage the risk of re-identification).
- Be conscious of the trade-off between the re-identification risk and the level of utility of the data.
- Try to ensure people with access to the de-identified data have not previously had access to other subsets of the identified data.



## Using AI to create meaningfully connected research databases and maintain data security and privacy

Medical practices hold datasets that are relevant to medical research, but they are not large or diverse enough to be useful by themselves. Australian company Presagen uses AI to efficiently and safely create the large databases required to advance medical research.


Presagen has developed a federated learning technique which allows the AI to train on data stored on various computers around the world, instead of needing to pool all the data together in a centralised database. With Presagen's decentralised system, the AI travels to the data, meaning it can remain private and secure on its home computer.



**Key principle**

Legal



An abstract graphic in the top right corner consisting of a dense, teal-colored dot pattern that forms a swirling, organic shape, resembling a stylized 'Q' or a wave.

**Believing that AI regulation  
and laws are sufficient to make AI  
safe and protect affected stakeholders  
from the risks, is a key determinant  
of Australians' trust in AI systems<sup>24</sup>.**



## Legal

The June 2020 launch of the Global Partnership on AI<sup>25</sup> – sponsored by various governments<sup>26</sup> including Australia – represents an important milestone in the journey to introduce AI regulations which are practical, sustainable, and grounded in human rights.

As the regulatory environment continues to evolve, leading organisations are playing a key role in driving trust in AI through the adoption of practices that proactively anticipate areas that will be in scope of upcoming AI regulations and address the limitations of the current legislative frameworks.

“While the law has historically lagged behind technological advancements, the scale and severity of the threats posed by uncontrolled Artificial Intelligence represent an opportunity for regulators, policy makers and the broader AI eco-system to collaborate and rethink the approach to developing and enforcing laws in the data and technology space.”

**Rossana Bianchi**  
KPMG Strategy, Growth & Digital



Almost all Australians (96 per cent) expect AI to be regulated, but most either disagree (45 per cent) or are ambivalent (20 per cent) that current regulations and laws are sufficient to make the use of AI safe and protect people from the risks.

Most Australians expect external regulatory oversight by the government or regulatory bodies, with co-regulation by government and industry also popular<sup>27</sup>.

## Develop a regulatory compass

Understanding current and upcoming AI and data regulations will prove challenging. In this context, organisations will want to invest in targeted areas.

### Key considerations

- Breaking silos to bring together compliance, governance, risk, data and technology specialists to develop a comprehensive and streamlined view of existing regulations’ relevance and applicability throughout the data and AI lifecycle.
- Launching pilots to assess the impact, practicability and sustainability of new principles and guidance for responsible AI (e.g. AI ethics framework published by the Australian Government in 2019).
- Contributing to future policies through participation in working groups coordinated by policy makers and other relevant Institutions (e.g. Standards Australia; Human Rights Commission).

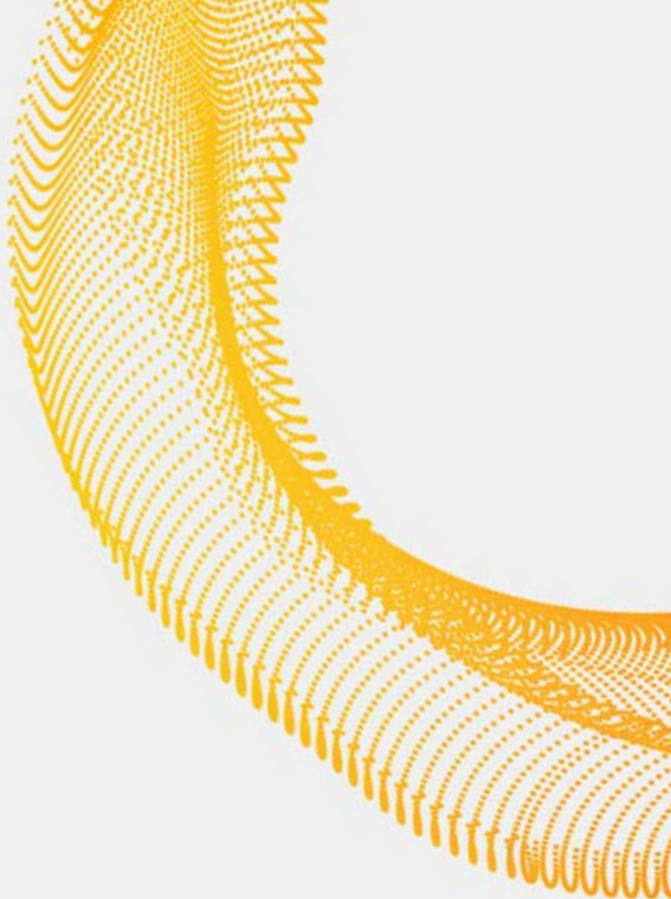




**Key principle**

# Ethics



An abstract graphic in the top right corner consisting of a series of concentric, wavy lines in a golden-orange color, creating a sense of motion and depth.

**To be trusted, AI systems need to be consciously developed and implemented to align with ethical norms and human rights. This underpins stakeholders' beliefs that the AI systems will operate with humanity and integrity.**

Internationally, over 80 reports outlining principles for trustworthy and ethical AI have been published<sup>29</sup>, which converge on a common set of principles<sup>30</sup>.

**These principles include:**



Australians have high expectations that organisations will adhere to the principles of trustworthy and ethical AI, such as those outlined by the Australian Department of Industry, Science, Energy, and Resources in their AI Ethics Principles Guidelines.



### Do no harm

AI systems do not cause physical, psychological or environmental harm, and preserve human autonomy and dignity by respecting, protecting and promoting human rights and agency.

It is important to give specific attention to the impact of AI systems on vulnerable stakeholders and populations, as harm to vulnerable people is particularly damaging to stakeholder trust<sup>28</sup>.



### Shared benefits

An AI system that benefits society – or a range of stakeholders – is likely to be more trusted than systems designed only to benefit the organisation implementing the AI.



### Contestability

Processes are in place for stakeholders to appeal, contest and challenge the outcomes of AI systems. If something goes wrong, a fair and accessible human review process exists with clear mechanisms for remediation where appropriate.



### Fairness

AI systems are fair, free of unfair bias, and designed to be inclusive for diverse stakeholders.



### Explainability

The purpose of an AI system, how it functions and arrives at its solutions, and how data is set and managed, should be transparently explained and understandable to a variety of stakeholders.



### Human oversight

AI systems should be designed and implemented in a way that retains human control, with appropriate human oversight including the capacity to decide how and when to use the AI system.

---

“Trust is a proxy for harm. We want to know when we lose trust. And if we lose trust, then almost certainly it means that people are being harmed, or that it’s not being used in a way that’s broadly acceptable.”

**Richard Boele**  
KPMG Global

---

“AI needs to be understood by the stakeholders that are making decisions, so that they’re comfortable that end consumers will receive the right outcomes. Having the right intent is not enough. We need to have the right governance and the right conduct to ensure AI systems don’t let us down.”

**Robert Warren**  
KPMG National Leader,  
Risk Strategy & Technology



To address the ethical challenges and risks posed by AI systems organisations can incorporate the following assessment and oversight processes into their AI governance and risk frameworks.

### Set ethical guardrails for your organisation

Ethical guidelines often fall into a form of self-governance, which relies on organisations and individuals to do the right thing. Yet there are many practical steps organisations can take to assure themselves, their stakeholders and the community at large, that the use of AI adheres to key ethical principles.

#### Key considerations

- Establish a set of standards the organisation commits to – an AI ethics codes of conduct – relevant to employees, customers and communities.
- Establish internal or external ethics boards to provide independent oversight, advice, assessment and monitoring of the ethics of AI systems throughout its lifespan<sup>31</sup>.



### Understand and quantify the ethical impact of the AI system throughout the AI lifecycle

To help understand the impacts of an algorithm's outcomes on individuals, communities, society and the environment, various guidelines and toolkits<sup>32</sup> have been produced to support the performance of Algorithmic Impact Assessments. Although there is no international standard for Algorithmic Impact Assessments, there are proactive steps organisations can take while norms are being clarified.

#### Key considerations

- Proactively engage relevant stakeholders throughout the AI lifecycle to ensure the system is addressing the right needs, in the right way.
- Perform a holistic assessment of the AI system's impact, including:
  - AI business case and intent
  - complexity of the system
  - maturity of data management practices
  - research capabilities
  - socio-economic impact and fairness, measured via the combination of human rights, ethical, societal, environmental and data protection impact assessments
  - safety and security.



### Adopt a balanced and proportionate approach to ethical risk management and human oversight

The ethics of AI is contextual, because it is driven by cultural values, norms and the specific use of the system. As algorithms are designed to continuously learn from experience – the ethics of an AI system is also highly dynamic. This means that the ethical risks of AI need to be regularly and proactively monitored as well as subjected to targeted and proportionate oversight and due diligence.

#### Key considerations

- Adopt a proportionate approach to oversight of AI systems, with more stringent and frequent controls implemented for higher risk applications and use cases.
- Complement existing AI risk assessment processes to understand and quantify the ethical risk assessment and use case throughout the AI life cycle.
- Invest in continuous monitoring mechanisms to address changes in the behaviour of the system that may result in heightened ethical risks.

### Seek independent assurance of the ethics and broader trustworthiness of AI

Independent assurance of AI systems is one of the key methods to drive trust in the adoption of AI.

#### Key considerations

- Put in place regular reviews of the ethics of AI systems by an independent body including representation of communities and stakeholders impacted by the AI systems.
- Adhere to a certification system that confirms a minimum level of transparency, accountability and fairness to the broader public<sup>33</sup>.





# Conclusion

---

Deriving the business value of an AI system while meeting stakeholders' expectations of trustworthy AI is not something a single executive, or business function, can answer alone.

Rather, AI systems are developed and deployed in complex ecosystems, where cross disciplinary expertise and collaboration is critical.

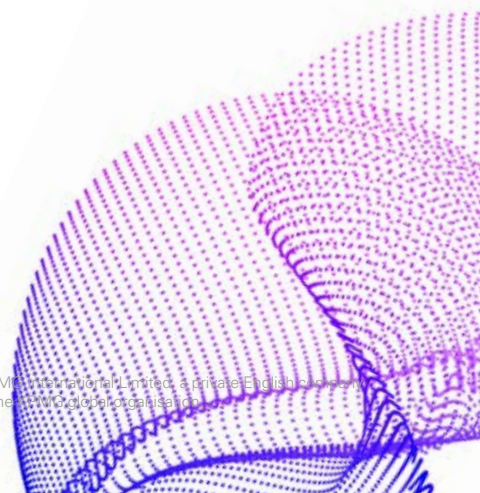
Adopting an integrated, organisation-wide approach – as presented in the model of Trustworthy AI – is necessary to effectively design, deploy, and govern AI systems that earn trust.

This trust will be critical to reaping the competitive and reputational benefits of AI.



# Endnotes

- 1 OECD (2019), Artificial Intelligence in Society, OECD Publishing, Paris. <https://doi.org/10.1787/eedfee77-en>
- 2 International Data Corporation. (2019, 4 September). Worldwide Spending on Artificial Intelligence Systems Will Be Nearly \$98 Billion in 2023, According to New IDC Spending Guide. [www.idc.com](http://www.idc.com)
- 3 <https://en.unesco.org/artificial-intelligence/ethics>
- 4 OECD (2019), Artificial Intelligence in Society, OECD Publishing, Paris. <https://doi.org/10.1787/eedfee77-en>
- 5 <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>  
Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), DOI: 10.1126/sciadv.aao5580.
- 6 <https://www.nytimes.com/2018/11/29/style/cambridge-analytica-fashion-data.html>  
<https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie>  
Bechmann, A., & Bowker, G. C. (2019). Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, 6(1), 2053951718819569.
- 7 <https://www.afr.com/politics/federal/morrison-apologises-for-robodebt-saga-20200611-p551ps>  
<https://www.afr.com/politics/federal/latere-721m-to-be-refunded-after-robo-debt-backflip-20200529-p54xrn>  
<https://www.theguardian.com/australia-news/2019/nov/27/government-admits-robodebt-was-unlawful-as-it-settles-legal-challenge>  
<https://www.theguardian.com/australia-news/2019/feb/06/robodebt-faces-landmark-legal-challenge-over-crude-income-calculations>
- 8 Lockey, S., Gillespie, N., & Curtis, C. (2020). Trust in Artificial Intelligence: Australian Insights. The University of Queensland and KPMG Australia. [doi.org/10.14264/b32f129](https://doi.org/10.14264/b32f129)
- 9 Lockey, S., Gillespie, N., & Curtis, C. (2020). Trust in Artificial Intelligence: Australian Insights. The University of Queensland and KPMG Australia. [doi.org/10.14264/b32f129](https://doi.org/10.14264/b32f129)
- 10 Lockey, S., Gillespie, N., & Curtis, C. (2020). Trust in Artificial Intelligence: Australian Insights. The University of Queensland and KPMG Australia. [doi.org/10.14264/b32f129](https://doi.org/10.14264/b32f129)
- 11 OECD.AI Policy Observatory – Accountability (Principle 1.5). Available at <https://oecd.ai/dashboards/ai-principles/P9>
- 12 European Commission HLEG AI (2019). Ethics Guidelines for Trustworthy AI. European Commission. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- 13 <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- 14 <https://www.bbc.com/news/world-australia-51496206>
- 15 <https://www.nature.com/articles/d41586-019-03228-6>
- 16 Lockey, S., Gillespie, N., & Curtis, C. (2020). Trust in Artificial Intelligence: Australian Insights. The University of Queensland and KPMG Australia. [doi.org/10.14264/b32f129](https://doi.org/10.14264/b32f129)
- 17 <https://www.oaic.gov.au/privacy/notifiable-data-breaches/notifiable-data-breaches-report-january-june-2020/>
- 18 <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>
- 19 <https://securityboulevard.com/2019/07/a-universal-bypass-tricks-cylance-ai-antivirus-into-accepting-all-top-10-malware-revealing-a-new-attack-surface-for-machine-learning-based-security/>
- 20 Lockey, S., Gillespie, N., & Curtis, C. (2020). Trust in Artificial Intelligence: Australian Insights. The University of Queensland and KPMG Australia. [doi.org/10.14264/b32f129](https://doi.org/10.14264/b32f129)
- 21 <https://www.abc.net.au/news/2019-08-15/myki-data-spill-breaches-privacy-for-millions-of-users/11416616>
- 22 <https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-and-the-privacy-act/>
- 23 <https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-decision-making-framework/>
- 24 Lockey, S., Gillespie, N., & Curtis, C. (2020). Trust in Artificial Intelligence: Australian Insights. The University of Queensland and KPMG Australia. [doi.org/10.14264/b32f129](https://doi.org/10.14264/b32f129)
- 25 <https://www.canada.ca/en/innovation-science-economic-development/news/2020/06/joint-statement-from-founding-members-of-the-global-partnership-on-artificial-intelligence.html>
- 26 Current members include France, Canada, Australia, the European Union, Germany, India, Italy, Japan, Mexico, New Zealand, the Republic of Korea, Singapore, Slovenia, the United Kingdom and the United States of America.
- 27 Lockey, S., Gillespie, N., & Curtis, C. (2020). Trust in Artificial Intelligence: Australian Insights. The University of Queensland and KPMG Australia. [doi.org/10.14264/b32f129](https://doi.org/10.14264/b32f129)
- 28 Jobin, A., Ienca, M., & Vayena, E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 1, 389–399 (2019). <https://doi.org/10.1038/s42256-019-0088-2>
- 29 Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikanth, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center Research Publication No. 2020-1. [doi.org/10.2139/ssrn.3518482](https://doi.org/10.2139/ssrn.3518482)
- 30 <https://home.kpmg/au/en/home/insights/2019/11/organisational-trust-guide-trustworthy-by-design.html>
- 31 <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>
- 32 <https://ainowinstitute.org/aiareport2018.pdf>
- 33 Independent High-Level Expert Group On Artificial Intelligence Set Up By The European Commission, April 2019. [https://ai.bsa.org/wp-content/uploads/2019/09/AIHLEG\\_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf](https://ai.bsa.org/wp-content/uploads/2019/09/AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf)



# 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 Futures  
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

©2020 The University of Queensland

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).

©2020 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.

Liability limited by a scheme approved under Professional Standards Legislation.

November 2020. 573005795MC