Deloitte.

Perspective | 11 Apr 2023 | 10 minute read

A practical approach to building trust in Artificial Intelligence (Trustworthy AI)

By Dr. Maryam Fanaeepour and Richard Kelly

The benefits of AI and automated systems are being harnessed by organisations and citizens, however poor governance and controls for these technologies present a risk to society and threaten to undermine the trust consumers and business have in AI technology. Deloitte's *Trustworthy AI framework* provides an approach to overcoming these challenges and build trust in AI, ensuring that risks are avoided, mitigated or controlled.

The use of AI technology and automated systems has been exponentially increasing. From self-driving cars and predictive policing to loan approvals and hiring algorithms, this new technology provides unique and effective solutions to organisational and societal challenges. Although the benefits are plentiful, there are risks if AI is not governed and controlled effectively. Citizens could face unintended harm or discrimination when using the technology, including compromised data privacy, embedded bias and errors in diagnosis systems.

The result of these impacts over time could erode the collective trust organisations and society have in AI technologies. The loss of trust introduces underuse - one of the greatest risks to AI technology benefits and progress.

To ensure human rights are respected, societal values are upheld and organisational risks are mitigated there is an urgent need to build Trustworthy AI. While existing regulations mandate organisations put in place necessary controls relating to AI technology, on their own they are insufficient. It is the combination of 'hard' or mandated regulation and 'soft' or self-regulation approaches, such as ethical or responsible AI principles and frameworks, that will create and maintain trust in AI. Fortunately, there is a growing understanding of the need within

industry for the union of ‘hard’ and ‘soft’ regulation to mitigate the unique risks of AI.

Deloitte’s Trustworthy AI framework

Deloitte’s Trustworthy AI framework supports ‘soft’ regulation of AI by:

- Providing a bridge of responsible practice and use of AI until legislation evolves to meet the need
- Engendering trust between AI technology and its end users
- Helping organisations adopt new technologies which are aligned to their business values
- Protecting business and society from the unique and inherent risks of AI technologies.

The framework is underpinned by six principles: Transparent and Explainable, Fair and Impartial, Robust and Reliable, Privacy, Safe and Secure, Responsible, Accountable and Contestable (Figure 1).

Figure 1- Deloitte’s Trustworthy AI Principles

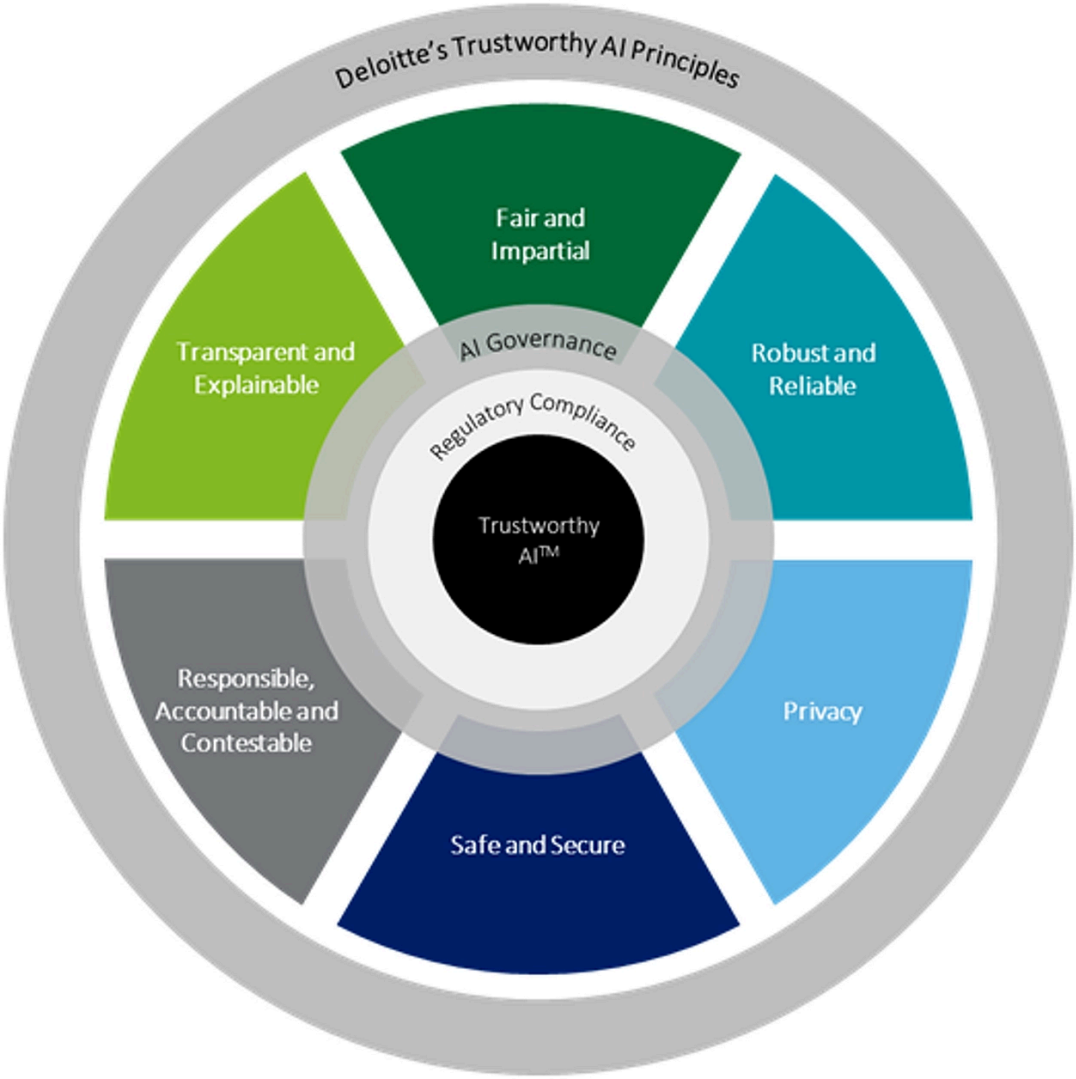

Figure 2- Deloitte's Trustworthy AI Lifecycle

It takes a lifecycle approach to overcoming AI technology challenges and identifies where in an organisation Trustworthy AI would deliver the greatest value and benefits, within the context of an organisation's industry, strategy, maturity, and objectives. The methods and tools used are customisable, practical and outcome focused, making the results unique to each user (Figure 2).

Figure 3- AI Development Lifecycle

For organisations with their own data science and AI development capabilities, all the elements of Deloitte's Trustworthy AI framework can be distilled into components that can be applied practically in the AI and machine learning development lifecycle, (Figure 3). From the first step to the point of delivery, there are practical processes and governance tools that provide continued alignment to organisational values and Trustworthy AI principles, safeguarding against risk.

If you would like to know more about Deloitte's approach to Trustworthy AI, please contact our team.

Authors :

Maryam Fanaeepour - Specialist Manager, Risk Advisory

Richard Kelly - Senior Manager, Risk Advisory

Contact us

Dr. Elea Wurth

Partner, Risk Advisory

ewurth@deloitte.com.au

+61 3 8486 1732

Recommendations

Deloitte's Cyber Team at the AISA Cyber Conference in Melbourne

Deloitte was recently a Diamond Sponsor of the AISA Cyber Conference held in Melbourne from 11-13 October at the Melbourne Convention and...

Perspective | 5 minute read

Challenge and choices - Risk Advisory Blog

Regulators and organisations alike recognise the importance of mature risk decision-making processes and the embedding of effective challenge as a...

Perspective | 5 minute read

Australia's response to modern slavery

Modern slavery is an umbrella term that encompasses a range of illegal and unethical workplace practices that exploit people for personal or...

Perspective | 10 minute read

Beyond passwords: The future of identity and access management - Risk Adviso...

New technology is both challenging identity and access management and providing new ways to address it.

Perspective | 10 minute read