AI ASIA PACIFIC
INSTITUTE

Trustworthy Artificial Intelligence

in the Asia-Pacific Region

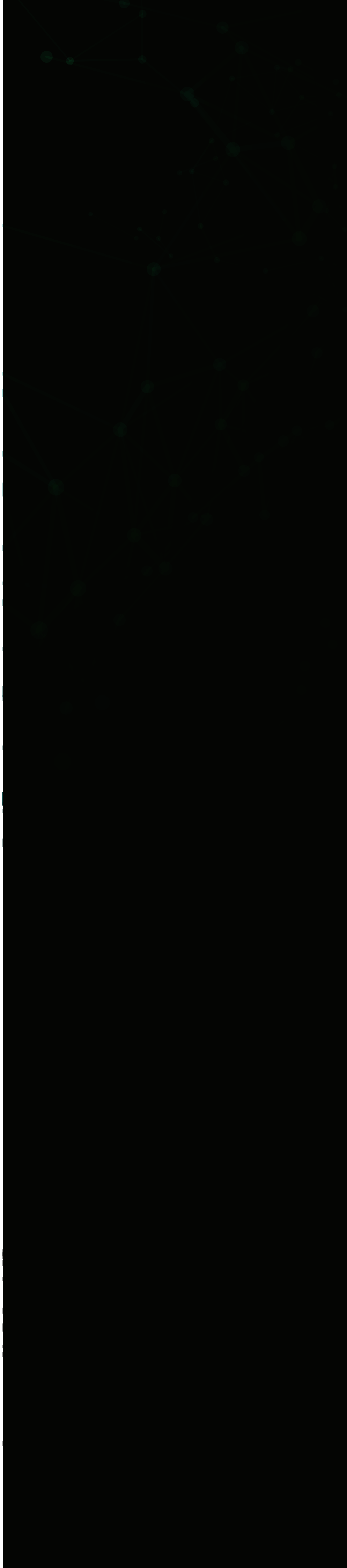
July 2021

About AI Asia Pacific Institute

The AI Asia Pacific Institute connects the Asia Pacific region in order to facilitate dialogue on AI. This collaboration is key to ensure the development and progress of Trustworthy AI.

> It is change, continuing change, inevitable change, that is the dominant factor in society today. No sensible decision can be made any longer without taking into account not only the world as it is, but the world as it will be...
>
> — Isaac Asimov, Asimov on Science Fiction

Table of Contents



Acknowledgements

We would like to thank our interns, Callum Hamilton, Samantha Jack and Ross Burrell for their time spent researching and contributing to this report; and Bond University for their support facilitating these internships.

Thank you to Richard Kelly (Senior Manager, Digital and Technology Risk at Deloitte) and Louisa Baczor (Founder and Research Consultant, Elbe Consulting) for their contributions to the editing process.

We are grateful for the insightful expertise and feedback provided by Professor Monique Mann (Senior Lecturer, Deakin University), Mark Manantan (Lloyd and Lilian Vasey Fellow at Pacific Forum and concurrently a non-resident fellow at the Center for Southeast Asian Studies at the National Chengchi University in Taiwan) and Hilary Sutcliffe (Director, SocietyInside).

Finally, we are grateful for the ongoing support of our advisory board members who help to shape and strengthen the AI Asia Pacific Institute: Philippa Penfold, Vincent Vuillard, Daniel Spencer, Antonio Feraco, Richard Kelly, Byoungchan (Ben) Eum; and to our founding and leadership team: William Chen and Luis Gonzalez.

This report has also benefited from the insights of a range of individuals quoted throughout.

Copyright

Important disclaimer

The reader is advised and needs to be aware that the information contained in this publication may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, the AI Asia Pacific Institute (including its members and advisors) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Foreword

We are experiencing a time of unprecedented technological transformation. AI is already transforming every aspect of our lives, and the COVID-19 pandemic has drawn its potential into light even more prominently. We sense AI's presence through an unspoken efficiency, and an ability to connect and work in ways previously unimaginable. But AI also poses complex questions that will require careful consideration. As we embark on building this new world, it's important to lay strong foundations that will not be shaken. The reality is that many of the pre-existing safety net structures were not designed for the transition to AI.

The challenges involving this technology are often not black-and-white. The work of technologists in building these tools deserves much respect, but they cannot design this future alone. This industrial revolution we are experiencing requires a multidisciplinary engagement. On that, I am personally thankful for all the support and input from various stakeholders in our work at the AI Asia Pacific Institute. Their involvement has been crucial in identifying problems and solutions to help guide our progress.

The implications of AI are transnational. Thus they cannot be resolved as an isolated issue, but must be a result of much cooperation and international collaboration. There is so much potential to work on our biggest problems, if we can come together as a society to design our future.

I believe we can draw a roadmap that will take us to where we want to go. Many questions still remain and much work lies ahead, including the urgent need for modernising laws, government policies, education, and resourcing. This report proposes a series of recommendations that can assist industry stakeholders in working towards the common goal of enabling AI to create a better future.

AI is set to continue to drive a vast range of efficiency optimisations, but it's our job to ensure that any legal, ethical and social implications are thought through at the same pace. I hope that this report and our work can contribute towards this goal.

Kelly Forbes
Director, AI Asia Pacific Institute
July 2021

Executive Summary

We stand at the dawn of a new era. As we enter a Fourth Industrial Revolution[1] characterised by a fusion of technologies, Artificial Intelligence (AI)[2] is influencing almost all areas of human life. This report aims to lay out a pathway for strengthening trust and acceptance of AI systems, therefore encouraging innovation and empowering stakeholders in the Asia-Pacific (APAC) region to move beyond a role as mere spectators of this technology revolution to active participants. Collectively the research insights provide an evidence-based pathway for building and maintaining the trust and acceptance of AI systems by the public. The insights are relevant for informing policy and practice across all three sectors of government, business, and non-profits.

Context

The APAC region is experiencing exponential growth in the development, integration, and normalisation of AI across sectors. Although many of these AI-based systems excel in performance and reliability, hesitance in implementation remains a challenge. This owes primarily to a lack of trust in AI, in large part due to the opaqueness of existing systems.

The implications of AI are now undeniable, with risks ranging from human rights infringements to the growth of the digital divide (i.e. a gap in terms of access and usage of AI). These risks can not only lower people's trust in AI systems, hindering innovation, but could impact the quality of the future being shaped by this technology.

Trust is the critical ingredient that can transform this reality and ensure that AI's full potential is realised. Thus, the AI Asia Pacific Institute's approach is founded upon ensuring that AI systems not only function, but are trustworthy. This is based on three pillars: Lawful, Ethical, and Robust[3].

Research

Our research is an attempt to take a deep dive into assessing the current state of Trustworthy AI in the region and exploring the level of industry understanding in relation to Trustworthy AI principles.

We hope our findings provide important and timely research insights into industry's adherence and understanding of Trustworthy AI. Through a multidisciplinary and collaborative approach, we have developed guiding recommendations for stakeholders in APAC to consider in the development, implementation and use of AI systems:

1. Build in processes to continually challenge whether AI systems adhere to the three pillars of Trustworthy AI (Lawful, Ethical, and Robust). Evolve these processes in line with industry developments.
2. Consider how the principles of Trustworthy AI can be applied in the given context to ensure economic and social benefits.
3. Governments and private sector organisations should prioritise rebuilding trust, building human capacity, international cooperation, and creating a global ethical framework around AI.

I. Why Trustworthy AI

Many different terms have been suggested in the industry, from ethical AI to responsible AI. In this chapter, we define what we consider to be Trustworthy AI and why we choose to work with this definition.

AI is driving the Fourth Industrial Revolution, and while efficiency and competition are primary drivers for innovation, trust is also evolving as a core foundation driver. It has been said that trust is the currency of the future; potentially becoming the cornerstone of the economy and a determining factor for organisations to remain relevant in the years to come.

For the AI Asia Pacific Institute, enabling trust in AI systems needs to go beyond adhering to merely ethical guidelines. Therefore we follow the approach put forward by the European Commission[4]:

'Trustworthy AI has three components, which should be met throughout the system's entire life cycle: (1) it should be lawful, complying with all applicable laws and regulations (2) it should be ethical, ensuring adherence to ethical principles and values and (3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm.'

Our model is based on the proposition that an AI system is deemed to be *trustworthy* if it complies with these three components:

LAWFUL **ETHICAL** **ROBUST**

To ensure alignment with each of these components, the following questions should be continually asked:

1. Does the AI system respect existing local and international laws? Such laws might encompass data-sharing legislation, Privacy Acts (if available), or/and any international agreement the country might be a signatory of.
2. Does the AI system respect principles of Trustworthy AI? In this context, we refer to the unified principles of Trustworthy AI within the APAC region (see chapter III). These are: Human-Centricity; Fairness; Transparency and Explainability; Privacy; and Accountability.
3. Is the AI system robust? While some frameworks treat this criterion as a principle, we view it as a separate requirement. That is, throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose. As an example, a self-driving car that can only operate in optimal weather conditions does not have sufficient robustness for deployment.

Each of these three components is necessary for the achievement of Trustworthy AI, but they are not self-sufficient. Given that this is an evolving field, organisations must continue to mature their processes in accordance with industry developments.

To further these standards, Trustworthy AI research has merged with the goal of creating appropriate legal and ethical frameworks with sound compliance mechanisms. A general understanding is emerging within the field that Trustworthy AI must also account for and improve interactions between humans and AI systems. Together, these factors ensure AI will continue to pose a novel challenge for policymakers and developers alike in the foreseeable future. With multiple usages, AI might be deployed successfully in one sector, while in another it may incur unanticipated consequences for stakeholders. Evidently, this can undo progress towards developing Trustworthy AI.

The dominant role of the private sector in AI technology development may necessitate new models of public–private partnerships and multistakeholder governance. With high demand for technical and legal expertise, the growing field of AI regulation is ripe for opaqueness and ambiguity, particularly for the everyday person. This may further a current trend towards democratic disillusionment and a general environment of diminished trust.

It follows that strategies geared towards creating and maintaining Trustworthy AI must be a high priority of all organisations engaging AI. To do so, AI systems must be considered in light of factors including their economic and social benefits, investment considerations, and compliance obligations. Organisations that take effective steps to account for these factors gain what is referred to as the 'Trustworthy Advantage'.

II. National AI Strategies

Overall, national AI strategies are becoming more prevalent in the APAC region. Their role is to provide national direction with respect to development and deployment of AI. It follows that these strategies are largely country-specific, accounting for factors such as economic specialisation, national demand, and cultural factors. In the following chapter, we set out an overview of these strategies at a country level before describing each in detail.

We acknowledge that there has been significant contribution to the AI industry from other countries in the region, such as Japan, South Korea, Taiwan and India, amongst others. While our research has focused on China, Australia, Singapore and New Zealand at this stage, we will continue to evolve and expand our work throughout the region. A revised version of this report reflecting this expanded research will be published in 2022.

Table 1. Overview of some of the APAC national AI strategies

Areas	China	Australia	Singapore	New Zealand
National AI Strategy	Next Generation Artificial Intelligence Development Plan (2017)	AI Technology Roadmap[5]	National AI Strategy (2019)	In progress
Industry-Specific AI Initiatives	China Artificial Intelligence Industry Innovation Alliance (CAIIIA); Three-Year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry (2018-2020)	Australian Technology and Science Growth Plan; Cooperative Research Centres (CRC) Programme; Data61's PhD Scholarship program; Digital Technologies Hub	Veritas (2019); Principles to Promote FEAT in the Use of AI and Data Analytics in Singapore's Financial Sector; Digital Economy Framework for Action (2018); Autonomous Vehicle Rules	Algorithm Assessment Report; The Centre for AI and Public Policy, Otago University report; Government Use of AI in New Zealand; Reimagining Regulation for the Age of AI: New Zealand Pilot Project.
Ethical Framework	Beijing AI Principles (2019)	Artificial Intelligence: Australia's Ethics Framework (2019)	Model AI Governance Framework (2019)	Algorithm Charter for Aotearoa New Zealand (2020)
Future of Work Initiatives	AI Innovation Action Plan for Colleges and Universities	-	AI Singapore (2017); SkillsFuture	The Future of Work Tripartite Forum (2018)
Current number of AI start-ups	1392 [6]	528 [7]	537 [8]	80 [9]

Australia

The 'Artificial Intelligence Roadmap' (AIR)[10] and 'Ethical Framework Discussion Paper'[11] jointly serve to establish the foundation of Australia's national AI strategy. The common thread between the documents is an emphasis on the ways in which AI can boost Australia's industrial growth and productivity, create new jobs and economic opportunities, and consequently enhance the quality of life for Australia. Published by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and the Department of Industry Innovation and Science in 2019, the AIR highlights how Australia can utilise AI to develop exportable solutions in healthcare, city building and infrastructure, and natural resources and the environment. The AIR highlights the positive economic impact of AI, noting the CSIRO's estimate that digital technologies could be worth $315 billion to the Australian economy by 2028.[12]

The AIR sets out a number of key objectives relevant to AI in Australia. These objectives are aligned with the following factors: developing an AI workforce, minimising the negative effects of automation, improving data governance, building trust in AI, increasing the activity in research and development (R&D), improving digital infrastructure and cybersecurity, and ensuring AI is safe and ethical. Overall, it could be said that Australia's focus on AI relates to employment, testing, and security.

According to the AIR, as the AI industry grows, Australia will need to expand its existing workforce of 6,600 AI specialists. Future projections indicate that by 2030, Australia's digital technologies and data science industry will require 32,000 and 161,000 employees respectively. Like other countries, Australia has realised the importance of upskilling to meet current and future operational needs. According to the World Economic Forum, digital technologies will globally displace 75 million jobs, but create 133 million new jobs.[13]

The use of AI is dependent on the existence of strong digital infrastructure. This includes access to secure and high-performance computing, in addition to fast internet connectivity. To enable effective AI, Australia will need to build upon initiatives such as the 'Mobile Black Spot Program' and 'Regional Connectivity Program', while ensuring high levels of cybersecurity. The country must also improve its support on R&D (which currently accounts for about 1.9 percent of GDP) to ensure competitiveness in the digital economy.

Australia is also uniquely positioned to engage the developments in Trustworthy AI on an international level. While these efforts are still in discussion with the publication of 'Artificial Intelligence: Australia's Ethics Framework',[14] the country can ensure that its early advancements in the industry are aligned with solid ethical foundations. This cautious approach can serve as a model to the rest of world.

China

China's 'Next Generation Artificial Intelligence Development Plan' (NGAIDP) outlines the nation's strategy to use innovation in AI to serve socioeconomic development, national security, and enhance national competitiveness.[15] The national AI strategy was first published in 2017 by the State Council (China's highest governing body), demonstrating a recognition of the importance of a top-down strategy to the Chinese national agenda.

China sees AI as an integral component of its national agenda, addressing challenges such as an aging population and environmental constraints. It has been utilising the technology across many different sectors, including education, medical care, and judicial service. The NGAIDP represents China's ambitious goal towards AI: to build a domestic AI industry worth nearly US$150 billion. The nation intends to be a global centre for AI innovation, propelling China's transition towards an innovation-driven and internationally dominant economy by 2030.

To achieve this goal, China sees four major tasks. Firstly, the nation must establish a system for open and coordinated use of AI in science and technology. Secondly, China aims to integrate AI within the digital infrastructure of civil society. Thirdly, China will abide by the principle of 'Three in One', emphasising R&D, product application, and industrial cultivation. Lastly, AI will be engaged as a key mechanism to support the growth of other technologies, the economy, social development, and national security.

The Plan establishes six actionable steps to fulfill China's objective. Step six is of particular importance within the overall strategy, guiding China's efforts to support AI development. The NGAIDP describes this step as '1+N'. In this regard, China's plan for next generation AI is twofold: '1' refers to China's support for major technology projects in next generation AI; 'N' refers to AI R&D projects deployed in accordance with national plans, designed to link major technology projects through the application of AI.

China has several measures supporting the NGAIDP and the development of AI more broadly, which include:

- Formulating laws and regulations and ethical norms related to AI development
- Improving major policies for AI development
- Setting up an AI technology standard and intellectual property system
- Setting up AI safety regulation and assessment system
- Enhancing AI labour force training
- Carrying out extensive activities to popularise AI.

To support this strategy, in 2019, China published the 'Beijing AI Principles'[16] outlining eight principles for AI governance and responsible use of AI. These include a) harmony and friendliness; b) fairness and justice; c) inclusiveness and sharing; d) respect for privacy; e) security and controllability; f) shared responsibility; g) open cooperation; and h) agile governance. China's ambitious AI plan is documented as evinced not only in the NGAIDP but alongside other initiatives such as the 'Made in China 2025' and the '2015 Military Strategy White Paper'. It is important to highlight that the Chinese government is aware of the potential benefits and implications of AI, and that the interplay of these factors will largely define the direction of China's AI strategy. Furthermore, on a brief comparative analysis, the principles emerging from China place a greater emphasis on social responsibility and group and community relations, with relatively less focus on individualistic rights.[17]

China is a central figure in the development of AI, and hence its actions will have far-reaching implications in shaping the contours of the Fourth Industrial Revolution. To this end, a nuanced understanding of China's domestic and international interests is paramount to internalise how it formulates its policies, especially from an ethical and legal standpoint. There is no shortage of academic discussions and political commentary on how China's conception and deployment of AI diverges from the West. But to effectively arrive at a holistic analysis of China's high-level AI principles, it is important to assess the structural, cultural, and political context that shapes its approach to and development of AI.

New Zealand

New Zealand has been a pioneer in public technology, leading in the use of electronic medical records and electronic funds transfer at point of sale (EFTPOS) systems. Naturally, the country is looking for ways to advance its AI strategy.

Though not an official national AI strategy, the AI Forum's (AIF) report 'Artificial Intelligence: Shaping a Future New Zealand'[18] provides an overview of AI engagement and recommendations for future AI development in the country. The report finds that AI has the potential to boost New Zealand's GDP by up to $54 billion by 2035 across 18 industry classifications. The New Zealand government is encouraging the growth of information and communications technologies, as this sector is set to become the second largest contributor to GDP by 2025.

As of 2017, there were 2,166 postgraduate students in Computer Science and 1,405 studying at a Master's or PhD level.[19] Most research in the field is conducted through the ICT Graduate Schools at the University of Auckland, the University of Waikato, Victoria University of Wellington with Weltec and Whitireia Polytechnic, and the Southern Tertiary Alliance. Realising the importance of strengthening New Zealand's innovation ecosystem, the government launched a $35 million investment to attract world-leading entrepreneurial researchers and their teams to the country.[20] Further investment under the Strategic Science Investment Fund is under consideration to develop national research capabilities in big data and analytics.[21] Another proposal is in development among the aforementioned universities to launch a National Centre for Data Technologies and Artificial Intelligence, aiming to address the lack of national coordination and devise ways to attract investment in AI. To this end, the AIF report identifies the need for national dialogues in several key areas, such as the impact of AI on employment, economy, and society.

The AIF report sets out several recommendations targeting both government and industry leaders for the future of AI in New Zealand. These include:

1. Forging a coordinated New Zealand AI strategy
2. Creating AI awareness and understanding by supporting research and public dialogue
3. Assisting AI adoption with a focus on SMEs (small to medium-sized enterprises)
4. Increasing trusted data accessibility through the publication of public data
5. Growing the AI talent pool through funding education and promoting diversity
6. Adapting to the effects of AI on law, ethics, and society.

The New Zealand government released an 'Algorithm Charter'[22] - one of the very first AI-related standards intended to apply to all branches of government regarding the use of algorithms by government agencies. The World Economic Forum is also leading a multistakeholder, evidence-based policy project in partnership with the government of New Zealand. The project aims at co-designing actionable governance frameworks for AI regulation.[23] Critical questions such as how to ensure the responsible use of AI by governments, and how to ensure that AI systems currently deployed are effectively compliant with existing regulations, are part of its agenda.[24]

Singapore

Singapore's National Artificial Intelligence Strategy (NAIS)[25] spells out its plans to deepen the use of AI technologies and transform its economy, going beyond just adopting technology and moving towards fundamentally rethinking business models, reaping productivity gains, and creating new areas of growth.

The NAIS outlines Singapore's vision for an AI-enabled economy in three respects:

a. Singapore will be a global hub for developing, test-bedding, deploying, and scaling AI solutions. This includes learning how to govern and manage the impact of AI
b. Governments and businesses will use AI to generate economic gains and improve lives. AI will raise the government's capability to deliver anticipatory and personalised services and will also be a strong driver of growth in key sectors of Singapore's economy
c. Singaporeans will understand AI technologies and the benefits it can bring; the workforce will be equipped with the necessary competencies to participate in the AI economy.

In working towards this vision, Singapore is supporting five 'enablers' in the AI ecosystem: Triple Helix Partnership, AI Talent and Education, Data Architecture, Progressive and Trusted Environment, and International Collaboration. To manage the growth of AI technologies, Singapore prioritises trust and capacity building. The nation has served as a model for the rest of the world in adapting AI talent and education for the future of work. It is investing heavily into training more Singaporeans for high-quality AI jobs, teaching basic computing skills and computation thinking, and attracting top-tier AI talent through various local initiatives.

Singapore has also been one of the first countries to lead conversations and specific initiatives in the space of Trustworthy AI. In 2019, it published the 'Model AI Governance Framework' [26], providing detailed and readily implementable guidance relating to ethical AI and AI governance, to private sector organisations that deploy AI solutions. Singapore has also enabled industry-specific initiatives, such as 'Veritas' for the financial industry. These developments are influential in guiding initiatives across the rest of the APAC region.

Singapore has also been a leader in international collaboration, recognising the importance of cooperation for the sustainable development of AI. The nation aims to promote an open and neutral environment for international researchers and businesses, which includes contributing to global standards for AI-related policy guidelines, as well as collaborating on multinational AI projects.

III. Principles of Trustworthy AI

The proliferation of AI technology has sparked numerous debates surrounding the principles that ought to guide its development, implementation, and use. As a result, studies have delved into what Trustworthy AI looks like, notably in meta-assessments or in relation to systemic risks and inadvertent adverse consequences of the technology, such as discrimination or algorithmic bias.

International organisations have reacted to these concerns by developing expert committees on AI, created to draft policy documents (or ethical guidelines) to address these issues. These committees include, but are not limited to, the High-Level Expert Group on Artificial Intelligence appointed by the European Commission, the expert group on AI in Society of the Organisation for Economic Co-operation and Development (OECD), and the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore. As part of their institutional appointments, these committees have produced reports and guidelines to support the development of Trustworthy AI. Similar efforts are taking place in the private sector, especially among organisations that rely on AI for their business operations.

While there are no universally accepted principles to inform the development and use of AI, the above initiatives provide guidance within the context in which they are implemented. In this section of the report, we aim to unify the principles of AI addressed within context of the APAC region, noting a degree of coherence and overlap between the principles in the region. We have created a synthesis of existing sets of principles produced by various reputable, multistakeholder organisations and initiatives in different countries. Our research outcomes have been tested with industry and are detailed in chapter IV.

The following AI Principles Comparison table compares existing developments across four countries in the APAC region. We have analysed and compared the existing developments in the principles of Trustworthy AI in these four countries to measure how fully have these countries have defined or provided guiding mechanisms in respect to the implementation of each principle. We acknowledge that these developments are continually evolving and a revised version of this table will be provided as part of the 2022 report.

Based on our research findings (see chapter IV), we propose the following principles to unify and encourage the development of Trustworthy AI in the region. In chapter V, we provide insight into the application of these principles by referring to use cases.

Table 2. Comparison of AI principles in APAC

Key[27]

🟢 Fully Considered
◐ Partially Considered
⚫ Not Presently Considered or in Development

Principles	China[28]	Australia[29]	Singapore[30]	New Zealand[31]
Human-Centricity	◐	🟢	🟢	🟢
Fairness	◐	🟢	🟢	⚫
Explainability	◐	🟢	🟢	⚫
Transparency	◐	🟢	🟢	🟢
Privacy[32]	◐	🟢	🟢	🟢
Accountability	◐	🟢	🟢	⚫

1. Human-Centricity

Human-Centricity represents a new paradigm for thinking about AI and its potential. The main objective of this principle is to encourage personal autonomy and self-determination in the development of AI and data-related decision-making.[33] As explained by Luciano Floridi,[34] the principle of creating AI technology that is beneficial to humanity is expressed in different ways.

These expressions are derived from the internationally recognised concept of human rights.[35] Essentially, a human-centric approach as outlined by the World Economic Forum is:

> One that makes central the following: that people have the right to determine, without any kind of coercion or compulsion, what happens to them. In the digital age, it can be compellingly argued that the data generated about us (including our social existence and our community at large) is deeply connected with the lived personhood of any and every human being.[36]

Put simply, a human-centric approach to AI is placing humans and the human experience at the centre of design considerations and intended outcomes of AI technologies. It is crucial to have a human-centric approach to AI in order to foster trust and confidence whilst respecting internationally accepted human rights.

> We need to realize that the current public dialog on AI — which focuses on a narrow subset of industry and a narrow subset of academia — risks blinding us to the challenges and opportunities that are presented by the full scope of AI, IA and II. In the current era, we have a real opportunity to conceive of something historically new — a human-centric engineering discipline.[37]
>
> – Michael I. Jordan

2. Fairness

The principle of Fairness of AI has been the topic of discussion among many scholars within the AI field. The proliferation of definitions in this area represents an attempt to make technical sense of the complex, shifting social understanding of fairness.[38] These discussions arise from the general need for all decisions to be free from unfair bias and discrimination.[39] In order to be fair, an AI system must avoid the creation and reinforcement of unfair bias, independent of given variables such as personal traits.[40]

Bias and fairness are human notions, as discussed in the McKinsey Global Institute's paper, 'Notes from the AI frontier: Tackling bias in AI (and in humans)'[41]. This paper outlines the distinctions between bias and fairness within the AI context. 'Bias' refers to 'unfair, unwanted or undesirable bias – that is, systematic, discrimination against certain individuals or groups of individuals based on the inappropriate use of certain traits or characteristics.'[42] The most common forms of unfair bias often relate to age, gender, disability and race. There have been several examples of this kind of discrimination with the use of facial recognition AI, where the system has made wrongful predictions against individuals based on their race or economic status.[43] Notwithstanding this, the absence of bias in AI systems does not necessarily indicate that a system has made a fair conclusion.

Bias influences people's everyday judgements. Once those innate feelings are removed from an AI system, what is left is a cold moral agent that passes judgement based on data points.[44] This may be beneficial to combat allegations of racism and sexism, but has the unintended effect of reducing certain processes to inhumane exercises. Nevertheless, fairness in AI is highlighted in other ways, such as fair access to AI and the fair distribution of its related benefits. The opposite of that signifies a growing phenomenon known as the 'digital divide' - a gap in terms of access and usage of AI. Further, stakeholders from the public sector place particular emphasis on AI's impact on the labour market and the need to fairly address democratic or societal issues.[45]

The development or use of the AI system should not result in unfair discrimination or disproportionate negative impact on individuals or groups. This principle requires great attention specifically during the AI development process to ensure data fairness as the foundation for fair results.[46]

3. Explainability and Transparency

Explainability refers to the ability of the AI system to explain its decisions in a way that can be understood by humans. To enable this, systems must offer a significant degree of transparency, hence why both principles often go hand-in-hand and can be described as improving AI through the minimisation of harm and development of trust. However, implementing these principles in the context of AI can be a complex process.

A group of scholars for the ScienceDirect journal thought it necessary to define explainability in terms of AI. They defined the term as 'given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.'[47] In general, this principle guides us in answering a simple question: 'how does the system work?'. By failing to answer this question, we are arguably building a 'black box' – a system that can only be viewed in terms of inputs and outputs, without knowledge of its internal workings. By definition,[48] black boxes lack explanation.

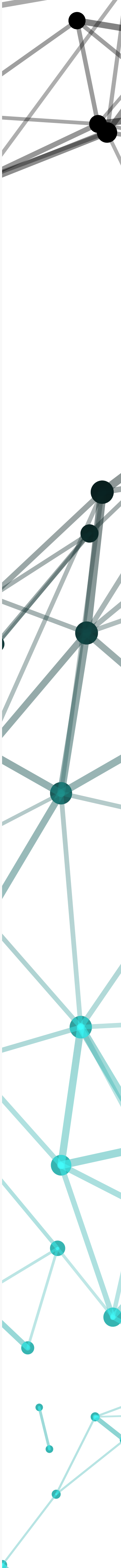
Explainability also assists with the identification of potential privacy breaches, causality, fairness, and trust.[49] These are all pertinent elements of Trustworthy AI. More importantly, they are intrinsically related to accountability, where we move beyond answering how the system works, towards placing responsibility on developers for its creation, deployment, and outcomes.

4. Privacy

Privacy and data protection is a ubiquitous issue in the AI space. Privacy has become one of the most important issues of this century and AI is at the forefront of these concerns, due to its capacity to amplify or rectify results. At a baseline level, humans expect a high level of privacy supported by legitimate means such as legislation and policy. In the effort to support Trustworthy AI, systems must ensure that not only personal data is protected and kept confidential, but that control remains in the individuals' hands.

While traditional conceptions of AI (such as surveillance systems and smartphone AI) challenge the notion of privacy, it is possible to foreshadow a future where privacy will be enabled by AI. For example, AI-enabled privacy may mean that fewer people will need access to raw data to utilise it, thus minimising the risk of privacy breaches due to human error.[50] In the digital world, it is important to remember that although privacy and technology have changed and evolved, the founding principles upon which they are based are not redundant. Rather, the increased use of AI may require the original definitions of privacy and specifically its implementation to be revisited.

In the context of this report, this principle occupies a significant role in ensuring that systems reflect trust and serves as a strong foundation for innovation.



5. Accountability

The principle of Accountability is arguably the most important principle to achieve Trustworthy AI. As previously stated, there are many recent developments encouraging the development of Trustworthy AI; often these initiatives are successful in setting strong foundations but are still on a theoretical level. An accountability principle can support the implementation of these ideas.

While accountability might initially relate to law and policies governing AI and its usage, compliance with the law is merely a step towards complete accountability. More crucially, individuals and organisations who are responsible for the creation and implementation of AI systems should be identifiable and accountable for any negative impacts. This has been suggested even if the impacts are unintended, to ensure the highest protection against unnecessary harm.

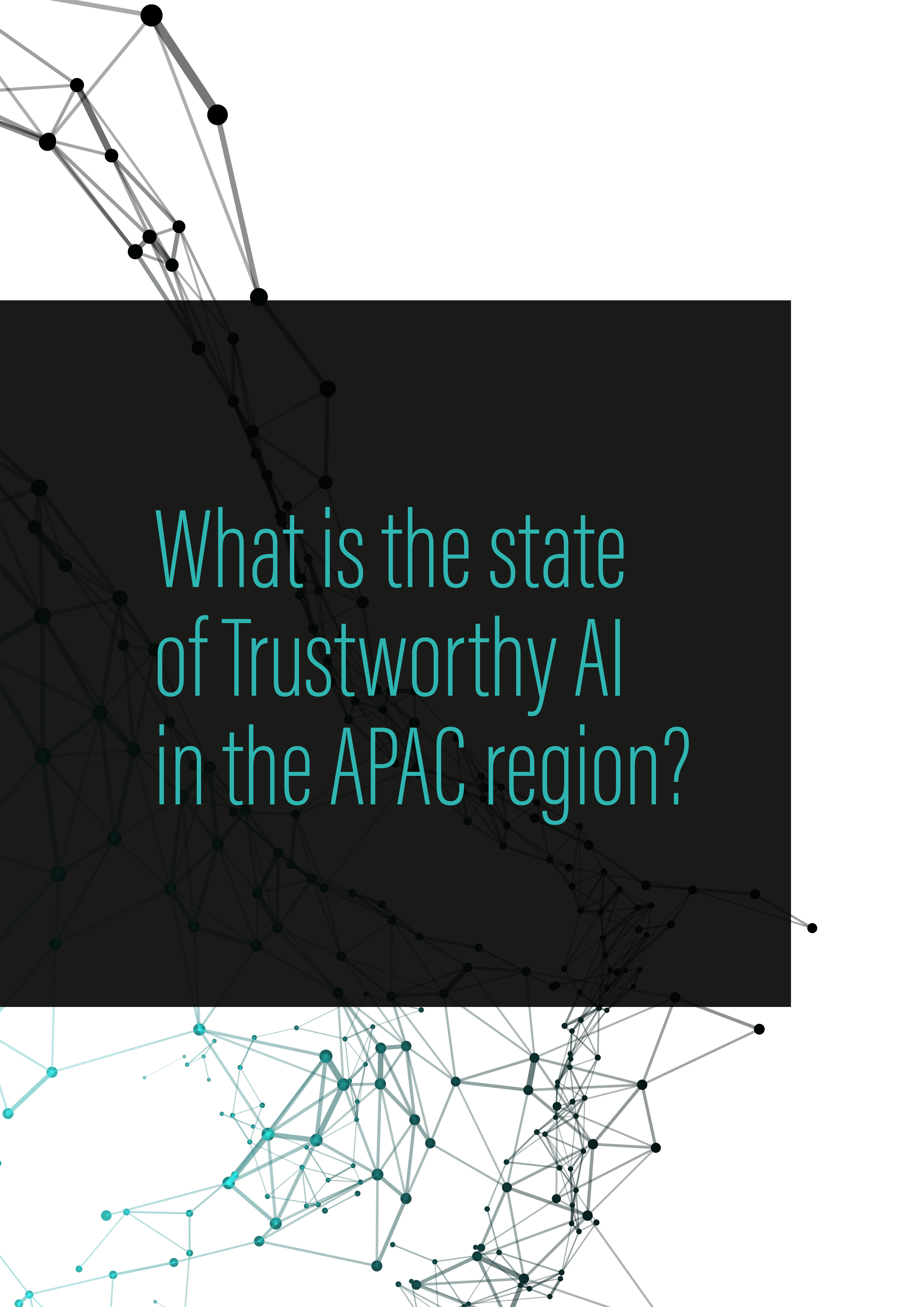
Accountability also refers to the mechanisms in place to ensure adequate redress when inevitable and preventable adverse impacts occur. By securing pathways for redress, scenarios which have the potential to break the chain of trust can be mended. As we will see in chapter VI, one of the most difficult issues pertaining to AI and its governance is that the technology is not confined to a single state or jurisdiction. This makes it problematic to create and maintain adequate practices and governance across state lines. While this is a global challenge, in practice, the implementation of accountability mechanisms by organisations (such as external auditing) can mitigate many AI risks.

IV. The Research

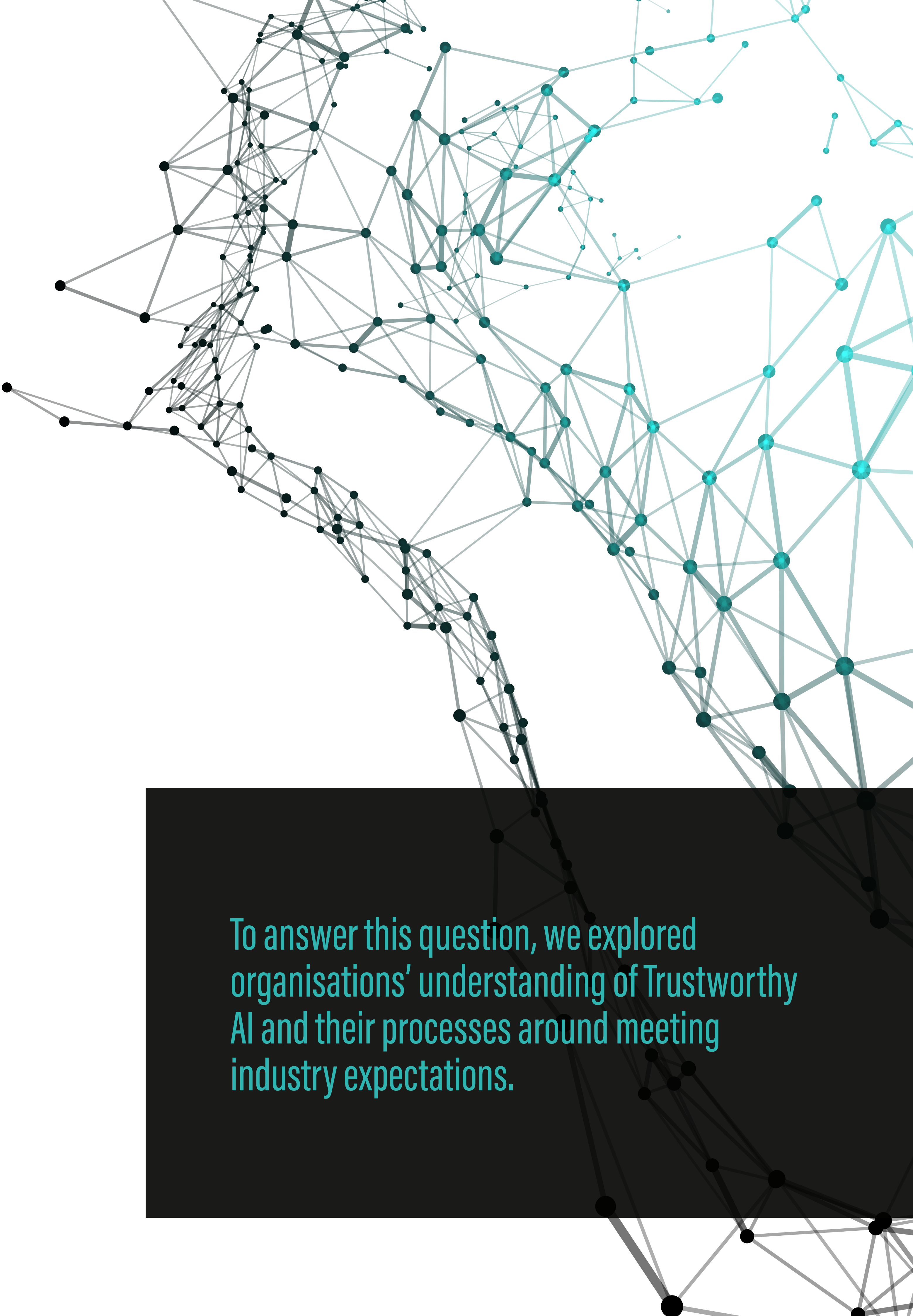
As AI has evolved, a culmination of frameworks has been published outlining key practices and principles for the development and deployment of Trustworthy AI. These are often theoretical ideals which remain a challenge for industry implementation.

One of the goals of our research was to determine not only what practices and principles are crucial for industry in the development and implementation of AI to foster trust, but also to identify current inconsistencies and propose measures of addressing these.

To explore these questions, we interviewed organisations that have a presence in the region, as well as individuals experts outside of the region.[51] In the following section we share the key findings from this research.



What is the state of Trustworthy AI in the APAC region?

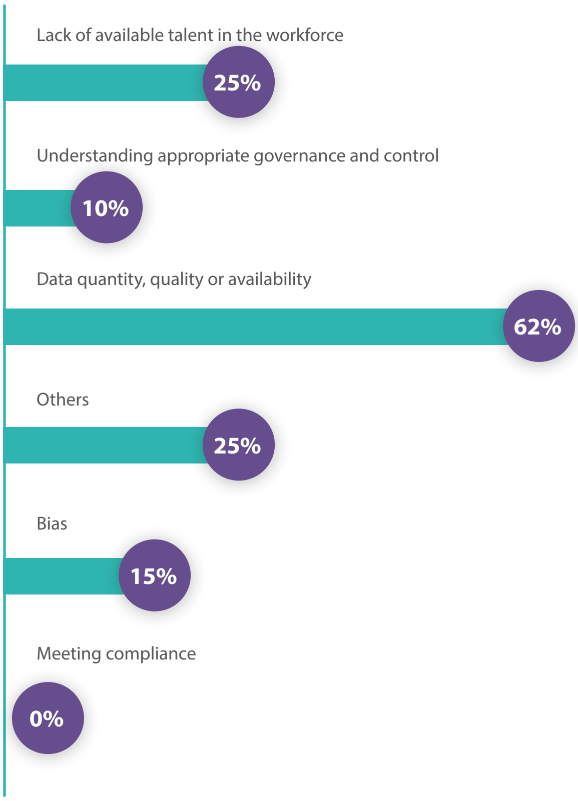


To answer this question, we explored organisations' understanding of Trustworthy AI and their processes around meeting industry expectations.

What are the biggest data challenges your company has encountered in the process of developing or deploying AI?

A staggering 62% say data quantity, quality or availability is the main challenge in developing or deploying AI. Some stakeholders note that often there is a misconception in the industry with respect to the collection of data. They added that without proper access to the correct level and quality of data, improving fairness in AI becomes more challenging. If we are working towards avoiding bias, for example, data on diversity can be collected to understand whether an algorithm is biased. In industries such as healthcare, where data is very sensitive, one way to approach this challenge is through the collection of indirect data. On this issue, some stakeholders also point out on the importance of balancing principles of Trustworthy AI (see chapter III) in each context. These principles cannot be applied in a generic way but remain adaptable to each product or situation.[52] Given the complexity of many algorithms, companies are advised to go above and beyond in explaining what data is being used and for what purpose, to build transparency and trust.

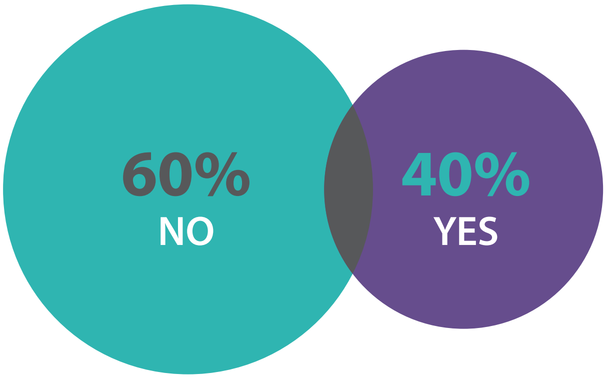
Figure 1. Current organisational challenges in the use of AI



Does your organisation's board or leadership team have access to regular industry discussion on topics related to the ethics of AI?

A high proportion of companies (60%) suggest that they lack engagement in Trustworthy AI industry conversations – for example, they are not part of a relevant body or do not follow industry updates. This is partly due to the lack of resources available to make this an internal priority or that they are simply very new to this conversation. Industry engagement is important to highlight, accelerate and shape the future of AI, helping develop an industry-led Trustworthy AI strategy.

Figure 2. Access to regular industry discussion

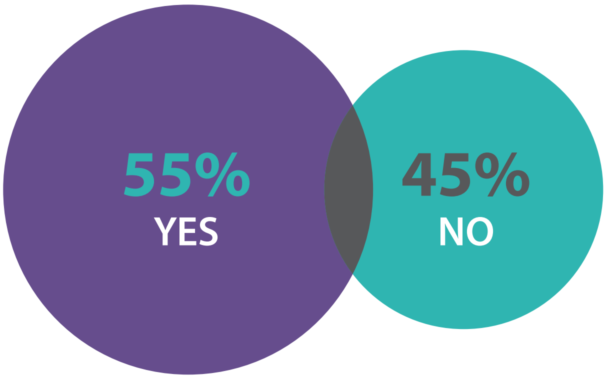


Would you say the principles Human-Centricity, Fairness, Explainability and Transparency, Accountability, and Privacy are sufficient to ensure that AI systems are trustworthy?

We consulted stakeholders on the principles of Trustworthy AI (described in chapter III). We recognise that these definitions and implementation processes are often very subjective. Specifically, we tested proposed definitions and how well stakeholders understand each one.

While most stakeholders agree that the principles of Human-Centricity, Fairness, Explainability and Transparency, Accountability, and Privacy form a solid foundation towards Trustworthy AI, 55% say additional attention around regulation or other mechanisms enforcing these principles should be considered as a matter of urgency. The need for clear standards and certification mechanisms was also highlighted as one of the approaches that industry can consider.

Figure 3. Principles



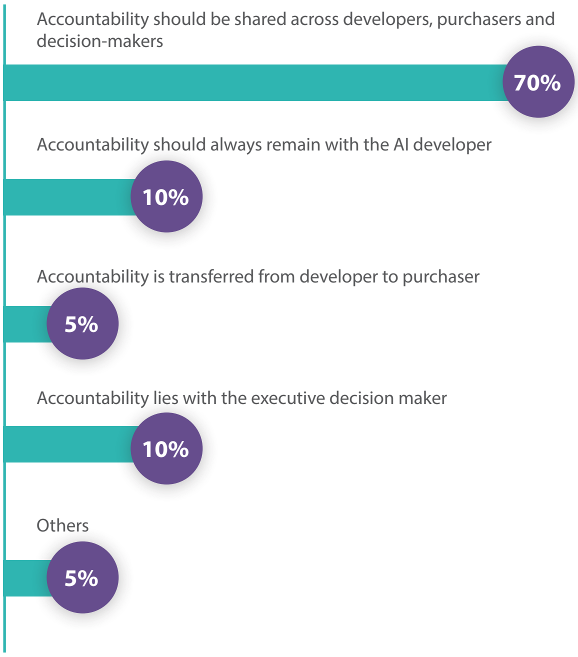
> Many companies are releasing high-level principles about their approach to designing and deploying AI products. But principles are only valuable if they actually get implemented.
>
> – Barbara Cosgrove
> *Vice President, Chief Privacy at Workday*

In respect to the Accountability principle, which statement do you agree with the most?

70% of stakeholders believe that Accountability should be shared across developers, purchasers, and decision-makers. Stakeholders also note the importance of external Accountability, as opposed to merely internal processes which can often result in the observance of principles without implementation. To this end, stakeholders encourage engaging an external company for auditing processes.

While most companies are still in the infancy stage of understanding Trustworthy AI and its guiding principles, the biggest challenge to overcome is around their capacity to operationalise the principles. To enable this, more education needs to be made available so that companies can move beyond being mere spectators of this evolving industry and towards inputting to how these decisions are unfolding.

Figure 4. Accountability



V. Case Studies

As AI adoption accelerates throughout the APAC region, it is foreseeable that challenges in the development and deployment of AI could become more prevalent. In the following section we share a compilation of case studies where the deployment or usage of AI has encountered such challenges or backlash owing to technical defects in the system or governance-related implications. Evidently, in many of the scenarios these challenges have caused adverse effects for the commercial feasibility of an AI system or the reputation and credibility of its deployer.

Through the analysis of use cases, we seek to provide guidance in avoiding similar pitfalls in the future by observing the lessons drawn from the scenarios.

AI in Social Services

Overview

The AI system known as 'Robodebt' used an algorithm to identify inconsistencies between individuals' declared income to different agencies within the Australian government. Where a discrepancy was identified, an automated notice of debt was generated and sent to the individual. It was later found that significant errors and discrepancies were being generated by the system, resulting in it wrongly pursuing thousands of welfare clients for debt they did not owe.

Impact

The algorithm-driven system has been accused of not only breaching Australian law on several counts, but that core principles for AI such as 'Transparency and Explainability' which are currently in discussion in Australia have also been breached. The affected parties were not only unaware that an algorithm was being used to make decisions that impacted them, but they did not know what information or data had been used by the algorithm in reaching these conclusions.

A proposed settlement where Commonwealth has agreed to pay $112 million in compensation to affected individuals, in addition to repaying more than $720 million in debts collected unlawfully, has been agreed.

Recommendation

It is recommended to invest in a more sensitive process for deployment of AI systems. Such a process should concentrate on the accessibility, usability, and transparency of the technology, including quality of service delivery and procedural fairness. In accordance with Transparency and Explainability principles, the responsible party should make public the use of the AI system, and ensure that the algorithms are explainable and its decisions able to be disputed by the affected parties.

AI in Education

Overview

Ofqual, England's regulatory agency for exams and qualification, used an algorithm that weighted scores based on the historic performance of individual secondary schools.

Impact

The algorithm incorrectly scored nearly 40 percent of the students lower than expected based on their previous grades. For some students, this scoring meant that they were now ineligible for the university programs they were expected to attend.

The testing algorithm was said to reinforce the existing social bias built into the UK's education system and disproportionally impact students from working-class or disadvantaged communities.

Recommendation

One of the unique recommendations realised by this debacle is that an analysis on the timing and appropriateness of such algorithm must be conducted prior to deployment. In this scenario, the system was deployed during an unprecedented time - a pandemic - to replace a complex exam process.

Additionally, it is important to acknowledge the limiting role algorithms play in fixing complex structural problems, such as bias. Beyond the timing of algorithmic deployment, investing in a process of more inclusivity and data diversity would have contributed to the avoidance of such results.

AI in Criminal Justice

Overview

The NSW Police Suspect Target Management Plan (STMP) is a New South Wales Police Force initiative deployed to prevent and reduce crimes. The system is based on a predictive style method of policing and uses an algorithm to assess how likely individuals are of committing a crime. It categorises people as high, medium or low risk of offending.

Impact

Evidence has shown that the programme has a disproportionate impact on Aboriginal and Torres Strait Islander peoples. The use of such AI system engages several principles of Trustworthy AI, such as the lack of Transparency in how the algorithm works. Although discriminatory claims exist, there is no explanation of how the algorithm operates and arrives at such decisions.

The affected parties were also not aware that an algorithm was being used to make decisions that impacted them, and did not know what information or data had been used by the algorithm in reaching these conclusions.

Recommendation

The STMP should be paused and reviewed for the venerable groups which are likely to be affected by the system.

In this particular example, the risks are extremely important as the effects of such a system do not stop when the accused is arrested, but can be carried throughout the justice system. As noted by Cathy O'Neil in 'Weapons of Math Destruction', 'the biased data from uneven policing funnels right into this model. Judges then look to this supposedly scientific analysis, crystallised into a single risk score. And those who take this score seriously have reason to give longer sentences to prisoners who appear to pose a higher risk of committing other crimes.'

AI in Hiring and Landing Credit

Overview

ChoicePoint was a background check service that collected information from public and private databases to put together reports on individuals that companies wished to enquire about.

An Arkansas resident, Catherine Taylor, was denied a job at Red Cross. Attached to her rejection letter was a report from ChoicePoint. The report detailed criminal charges linked to her, which belonged to another Catherine Taylor born on the same day. Such incorrect history was later found to be also linked to Catherine's inability to acquire a loan.

Impact

Catherine was victim of what we might call a 'black box' system. It made decisions about her job applications and credit scores without prior disclosure of the data used or how it arrived at decisions. Individuals subject to such systems are often unaware that a system is being deployed and making decisions which might impact their lives.

ChoicePoint was involved in several lawsuits and consumer complaints of providing inaccurate and out-of-date information in its criminal background reports, resulting in unfair job losses for applicants.

Recommendation

While such systems are built to run automatically and errors can often be inevitable, such errors can be mitigated by applying the principle of Human-Centricity. By keeping a human in the loop in such situations to verify sources and datasets, risks can be identified and reduced in a timely way.

Additionally, companies providing such services should ensure that people are aware when their personal data is being collected to inform conclusions by a system. Consent in such cases allows for clarity on the ownership of data, and provides the ability to question decisions and control misuse.

An analysis of the criticality of risk of a false positive or false negative versus accuracy of the technology should be conducted.

> One of the key benefits of AI for hiring is that automated procedures can be evaluated for bias in a way that human decision-making processes cannot. This starts with examining the data that is used to train a model to avoid proxy variables, or inputs that are strongly correlated with demographic identity. Some proxy variables are easy to identify, like a person's name or alma mater, but seemingly innocuous variables can also entrench bias depending on the underlying sample. Developers of AI systems need to be especially aware of hidden sources of bias in contexts like employment; technology can either reproduce historical inequalities or mitigate them in an unprecedented manner.
>
> — Sara Kassir, Manager, Public Policy and Research at Pymetrics

VI. International Recommendations

As we have seen in chapter III, while there are abounding principles guiding organisations using or developing AI systems to strive for the best outcomes, there are no universally accepted recommendations that aim to achieve Trustworthy AI. Rather, governmental institutions and international forums have provided recommendations for the safe development and implementation of AI. These recommendations are not legally binding, but they are highly influential in being capable of setting the international standard in a wide range of areas and informing national policy decisions. Below we have listed the key international recommendations we encourage governments, as well as the private sector, to consider adopting. We note that this is not an in-depth exploration of the complexities of the AI ecosystem, but is merely a starting point in encouraging the creation of Trustworthy AI.

Rebuild Trust

We cannot foster innovation without trust. Following the theme proposed in the recent Davos Agenda by the World Economic Forum, this is a *crucial year to rebuild trust*.[53] This international recognition builds on many recent and historic studies which uphold trust as a key element for economic growth and social prosperity. In the context of AI, trust can foster the adoption of technology, and enable communities to evolve from being mere spectators of this Fourth Industrial Revolution to actively contributing to it. AI has inspired both fascination and fear, leading to much resistance. By enabling trust within communities, we are opening the path to innovation as society is more likely to welcome new technologies.

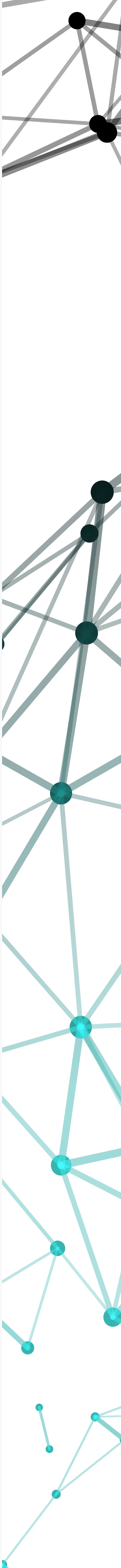
The '2020 Edelman Trust Barometer Special Report'[54] reveals that brand trust (53 percent) is the second *most important* purchasing factor for brands across most geographies, age groups, gender, and income levels - trailing slightly behind price (64 percent). Brand trust was found to be more imperative than all other factors, such as reputation or performance.

For governments and organisations engaging with the trust agenda, TIGTech's 'Trust and Tech Governance'[55] report provides practical guidelines which firstly focus on being trustworthy. It is not enough, however, to simply be trustworthy: it is necessary to progress towards a position where evidence of trustworthiness can be shown. Applying this in the context of AI, the focus turns towards several key principles, as discussed in chapter III. In ethics, principles help determine right from wrong. They provide guidance on building and deploying AI. Identifying these principles is a good start, but we need to move beyond theory and reflect upon the relevance of these principles in our work.

Building Human Capacity

Building human capacity for using AI is crucial to seamlessly integrate AI technologies into the world. The G20[56] and OECD[57] outline some key recommendations to prepare the labour market for transition into the world of AI: firstly, governments should invest in collaboration and work closely with stakeholders; and secondly, governments should invest in empowering people to effectively use AI technologies, facilitating their skill development.

One of the ways to empower the younger generation is by opening opportunities for the development of skills necessary for work in the twenty-first century. Singapore has demonstrated excellence in this area with the creation of many initiatives, including SkillsFuture.[58] This recommendation has also been supported



by the UN System Chief Board for Coordination, which deliberated a system-wide strategic approach and roadmap for supporting capacity development for AI.[59] Its main focus is to support developing countries, with an emphasis on the bottom one billion in the context of achieving the Global Sustainable Development Goals.[60] This approach will address the structural and dynamic consequences that may arise from AI discrimination, marginalisation, and bias on factors such as sex, age, disability, religion, race, ethnicity, and class.[61]

Overall, building human capacity requires the support and assistance of state and federal governments to provide fair and equal opportunity to people and businesses, as well as private enterprises and corporates to contribute to these efforts. Further, its overarching goal is working for the public good, hence proactive support will be pertinent to the seamless transition of AI into the labour market and preparing communities for the future of work.

International Cooperation

Finally, in order to achieve Trustworthy AI, international cooperation is crucial. In March 2020, participants from the UN Educational, Scientific and Cultural Organization (UNESCO) came together to discuss ways to steer the use of AI towards inclusion and equity. At the event, Regional Coordinator for the AI for Development, Kathleen Siminyu stated:

> International initiatives can support the work of redesigning tech for communal control and decentralisation which would counter the development. They can also advocate for discussions around tech that are holistic and address structural inequality, identity culture and politics, as well as support capacity building of technical expertise in the developing world.[62]

AI has potential benefits for all nations, but distribution of this technology across all regions remains a key challenge. Research indicates that China's and North America's economic gains are expected to represent 70 percent of AI's global economic impact by 2030. It is important to ensure that developing states are not left behind in terms of access to knowledge, data, education, training, and human resources.

Cooperation includes the sharing of knowledge between states and institutions. UNESCO's recent 2020 report on AI and gender equality highlighted as one of its key recommendations to 'establish a whole society view and mapping of the broader goals we seek to achieve.'[63] To achieve this goal, governments should employ cross-border evidence gathering measures to collate internationally comparable metrics measuring AI innovation. In this way, AI development gaps between states can be outlined and principles applied accordingly.

We recognise that deep and meaningful cooperation is not always a given. To achieve this, we must rely on alternatives that can open doors for larger global conversations. The World Economic Forum has recommended the creation of Centres of Excellence as an option to foster collaboration in AI.[64] The creation of these centres can be perceived as evidence of countries' efforts in building a safe and responsible AI ecosystem. Additionally, enabling them to start these efforts locally but through a Centre of Excellence, they are able to join forces and have a global network. A number of countries already has Centres of Excellence, such as Singapore; those which do not are encouraged to develop one:

> We need to start thinking about how we can use AI to break down some of these borders, to break down some of these jurisdictions and to make sure that we work together as humans, not just as different countries[65]
>
> — Kate MacDonald, World Economic Forum

Global Ethical Framework

Principles alone cannot guarantee Trustworthy AI. Hence, we need to globally develop a multicultural system which sets the foundation for AI developments.

AI does not recognise borders; it is a global phenomenon. Yet, today, no global ethical framework for AI developments exists. Building on the international cooperation recommendation above, UNESCO and other international organisations have called for a global approach to creating such a framework.

> An ethical framework allows us to pursue excellence. By basing our thoughts, decisions and actions on a clear statement of why we're here, what we stand for and where we draw a line in the sand, we go far beyond a 'do no harm' approach to ethics. Instead, we're able to imagine the best version of something – in this case, the best kind of technology.
>
> — The Ethics Centre[66]

There is growing support from different stakeholders - businesses, research centres, science academies, governments, international organisations, and civil society associations - for a global ethical framework for AI development.

While we may need to allow time for a legal framework to guide future research on ethics on a global scale, we can start regionally. This process could be ignited with local investments in its consideration and agreement involving existing principles. We also cannot underestimate the need for every country to work on publishing their own national 'AI Strategy'.

VII. The Future and AI: Predictions

Data Protection

There is no question that data usage and its accompanying security issues are some of the biggest challenges in modern society. Industry requires access to comprehensive datasets to ensure algorithms are both accurate and reliable, and to tap into the potential that AI can offer. This raises several questions: what dataset should be used to train an algorithm? How should that data be collected? What measures should be taken to store data securely? Who should be able to access secured data? And perhaps most importantly, what limitations should be placed on the use of personal or sensitive data?

Two points on the usage of user data are now clear. On the one hand, to create reliable, accurate, and commercially feasible algorithms, developers need access to comprehensive and unbiased datasets. On the other hand, users must have some degree of protection from the exploitation of their data. While currently actions on user data disputes in APAC must be brought under civil law, it seems likely that legislative frameworks for the protection of personal data will eventually come to balance these overlapping interests.

Both globally and regionally, regulatory schemes covering personal data protection are proliferating. The European Union's implementation of the General Data Protection Regulations (GDPR) in 2018 is one such landmark. According to European Research Federation's Director of Policy, Michelle Goddard, by requiring overseas-based organisations to comply with the EU's standards for data protection, the GDPR' marks a fundamental change in the balance of power between organisations and individuals in the collection, processing and storage of personal data'.[67] In Singapore, the Personal Data Protection Act launched in 2012 marks a similar push in APAC to balance the individual right to privacy with the need for organisations to access reliable datasets.[68] China is also seeking to expand the guiding principles contained in the Personal Information Security Specification by drafting a Personal Information Protection Law similar to the GDPR in 2020.[69] The draft significantly resembles some of the GDPR definitions, with terms such as 'personal data' and 'processing'. If in force, the legislation is likely to significantly impact companies with operations inside and outside of China.

This year we saw the private sector joining in this dialogue, with the release of a new policy by Apple. The new App Tracking Transparency feature will require apps to receive the user's permission before tracking their data on platforms owned by other companies.[70] This has been recognised in the industry as a step forward in enabling consumer awareness and control over their data.

As personal data protection develops, companies will need to ensure that existing internal policies account for the compliance requirements of emerging regulatory demands. In the APAC region, adaptation to personal data protection laws may take one of two forms that transnational organisations must be prepared for. On the one hand, states may develop isolated and exclusive frameworks that legislate domestic attitudes towards data protection. Under this scenario, there may be significant transnational barriers to the utilisation of data. Such a development might restrict the flow of digital technology capital as well as regional innovation in machine learning and AI more broadly. On the other hand, multilateral engagements on data protection such as the Asia-Pacific Economic Cooperation (APEC) Privacy Framework (developed by the APEC forum as the blueprint for greater regional cooperation on privacy rules and enforcement) may give rise to a regional framework resembling the GDPR.[71]

Multilateral engagements could promote knowledge and capital exchange within the APAC region by establishing necessary protections and removing barriers for cross-border information-sharing. In the short-term, relevant consideration must be given to the imminence of national data protection frameworks. Although such recommendations are not legally binding, we can anticipate significant changes in this field that must be attended to by researchers and policy writers. For example, the Organisation for Economic Co-operation and Development Privacy Guidelines, adopted in 1980, states that there should be limits to the collection of personal data.[72] The proposal embodies the principles underlying many privacy laws and frameworks in the United States,[73] Europe,[74] and Asia.[75] We can certainly expect to see more developments regarding protective obligations on personal data protection, as well as legal liability for violation of these new laws.

Human Impact

The continued growth of AI will enable new capabilities and empower society, but will also impose new challenges which need to be addressed at an organisational and governmental level. As with previous industrial revolutions which reduced the demand for physical labour, advances in AI pose new challenges for the cognitive labour workforce. In service-based economies, workers use learned skills to address a series of tasks within a given occupation. McKinsey Global estimates that AI has the capacity to automate half of all existing work activities and handle 30 percent of all activities performed in 60 percent of all occupations.[76] Further research indicates that in the coming decade, 20 percent of jobs in Asia will be affected by AI and 12 percent could be displaced.[77] This poses a significant challenge when addressing the human impact of AI.

Firstly, organisations must consider how AI will be integrated within the workplace. Apprehension towards this may pose a significant obstacle. This unease may arise from reasons that have frequently been cited as causes for concern regarding the implementation of AI; including automation and job replacement, digital ethics, and a lack of knowledge.[78] To mitigate these concerns, the Singapore Advisory Council of the Ethical

Use of AI and Data proposes that organisations can consider a two-step process:

1) Emphasise human-centred AI: engaging AI as a tool that supports the success of customers and employees, rather than using AI as a means to an end
2) Ensure effective communication: to build trust for AI systems within an organisation it will be necessary to convey to employees the why, what, and how of AI implementation. Proactive communication on these points serves as a crucial way to identify and resolve internal issues at an early stage.[79]

Secondly, governments will need to consider the broader implications of AI for the future of work and develop policies that reflect societal attitudes. To this end, there are two major challenges for governments. On the one hand, governments will need to equip future workers with the skills necessary to succeed in a digital labour market. Fundamentally, governments must ensure availability of educational opportunities - emphasising digital skills - to support a strong pool of AI talent. On the other hand, governments must be prepared for employment challenges in industries adversely affected by AI. Although governments often cannot prevent the obsolescence of individual enterprises, policy can be used to create employment pathways for labourers in affected industries.

Ethical Advantage

Industry has shown that ethics in AI is moving from a 'nice to have' to a critical element; providing a foundation for innovative developments. The Ethics Centre recently commissioned Deloitte Access Economics to develop a framework to quantify the societal benefits of ethical AI. The result is the report 'The Ethical Advantage: The Economic and Social Benefits of Ethics to Australia'[80], the first to quantify the benefits of ethics. Although the study focuses on Australia, it summarises the average GDP per capita against the average trust level from a sample of 59 countries, between 2010 and 2016. According to this analysis, countries with higher levels of trust tended to have higher income levels. Approximately one-fifth of the cross-country variation in GDP per capita is related to the differences in trust levels.

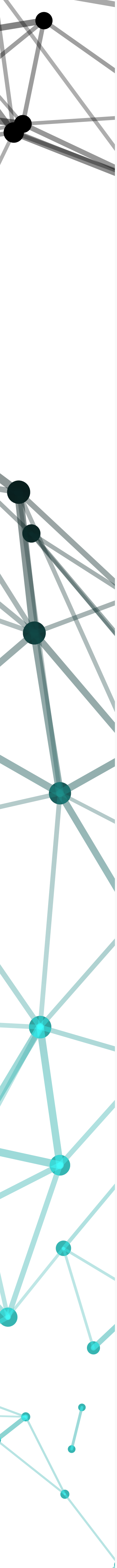
Many other studies point to similar results: ethics leads to more economic growth and social benefits. As Professor Klaus Schwab, Founder and Executive Chairman of the World Economic Forum, stated, 'we are witnessing a mindset shift from short-term profit maximisation to a world that is much more characterised by stakeholder responsibility'.

In this mindset shift, there are many advantages awarded to organisations prioritising ethics, including: minimising and preventing costly risks, earning public and governmental trust, and investment preferences.

According to Janet Wong CFA, Investment Stewardship Specialist of J.P. Morgan Asset Management[81], at minimum investors are expecting to see that the following elements can be demonstrated:

1. Evidence of AI governance and oversight within the company, including clear responsibility on the board level to oversee AI-related issues
2. Evidence of public commitment to trust the AI
3. Evidence of how the company is operationalising these ethical principles.

Whether for investors promoting social change or executives seeking to limit reputational and financial risk, this evaluation of AI systems is now critical. Understanding AI applications is a vital step in managing stakeholder expectations and integrating AI. To correctly conduct such assessments, stakeholders will be required to develop deeper layers of inquiry on risk and implications of such AI systems against ethical recommendations.



Compliance and Regulatory Considerations

The industry is constantly evolving, and many recent developments point to the increasing intention of shaping these developments and building trust in AI. What this will look like in the near future is a matter of debate. What we do know is that we can expect to see more initiatives as the industry continues to draw the attention of policy makers, legal professionals, and politicians. A recent study in Australia found that 96 percent of the community expect AI to be regulated.[82] The study also showed the increasing distrust in AI systems, and that the general population perceives the current regulatory and legal landscape to be insufficient to mitigate AI risks.

As discussed above, many international organisations are working towards building this awareness of the need to regulate AI developments so that they conform to the fundamental rights that frame society. This might take shape in the form of regulation or sector-specific initiatives. The latter was the approach taken by Singapore with the creation of 'Veritas', a series of financial institutions intended to promote the responsible adoption of Artificial Intelligence and Data Analytics (AIDA).

In April 2021, the European Commission proposed a regulation laying down harmonised rules on AI (the Artificial Intelligence Act).[83] The initiative is a result of a coordinated European approach on the human and ethical implications of AI which represents a major legislative global advancement in the industry. If integrated into the European Commission's legislative proposal – expected early next year – these recommendations will make the European Union the first region to create a structured legal framework on AI governance.

The European proposal contains a whole host of new duties for those who put into circulation what the proposal defines as 'high-risk AI'. Categories of high-risk AI include AI intended to be used in critical infrastructure, educational institutions (including access to and performance within such institutions), and employment. In this aspect, we can expect to see growing global interest to mitigate and potentially undercut technologies like facial recognition. The industry has received increasing support to regulate this technology, including from large Tech companies such as Microsoft and Amazon which agreed to pause their developments in this area. Others have stricter views which recommend the ban of these technologies.[84]

In chapter VI, we also saw that there are considerations on the development of a global framework on AI governance. We can probably expect to see new levels of development and cooperation in this area, with international organisations such as UNESCO encouraging a global approach to AI.

A clear pathway for enhancing trust in AI might be to strengthen the regulatory and legal framework governing it. Independent of how regulatory developments unfold in the next few years, organisations that can place these considerations as a high priority will unleash new potential. Due to the business risk attached to trust in AI, understanding AI applications is a vital step in managing stakeholder expectations and applying practice while integrating AI.

These efforts must be collaborative. For this reason, the AI Asia Pacific Institute approaches AI in a multidisciplinary way. We can create the future we want – if we all work together.

Conclusion and Recommendations

This report highlights important insights on how Trustworthy AI is understood, encouraging a pathway to ensure strong ethical standards and build public trust of AI systems.

To ensure the development, implementation and use of Trustworthy AI systems, we propose the following recommendations for stakeholders in the APAC region:

1. Build in processes to continually challenge whether AI systems adhere to the three pillars of Trustworthy AI (Lawful, Ethical, and Robust). Evolve these processes in line with industry developments
2. Consider how the five principles of Trustworthy AI can be applied in the given context to ensure economic and social benefits
3. Governments and private sector organisations should prioritise rebuilding trust, building human capacity, international cooperation, and creating a global ethical framework around AI.

Appendix

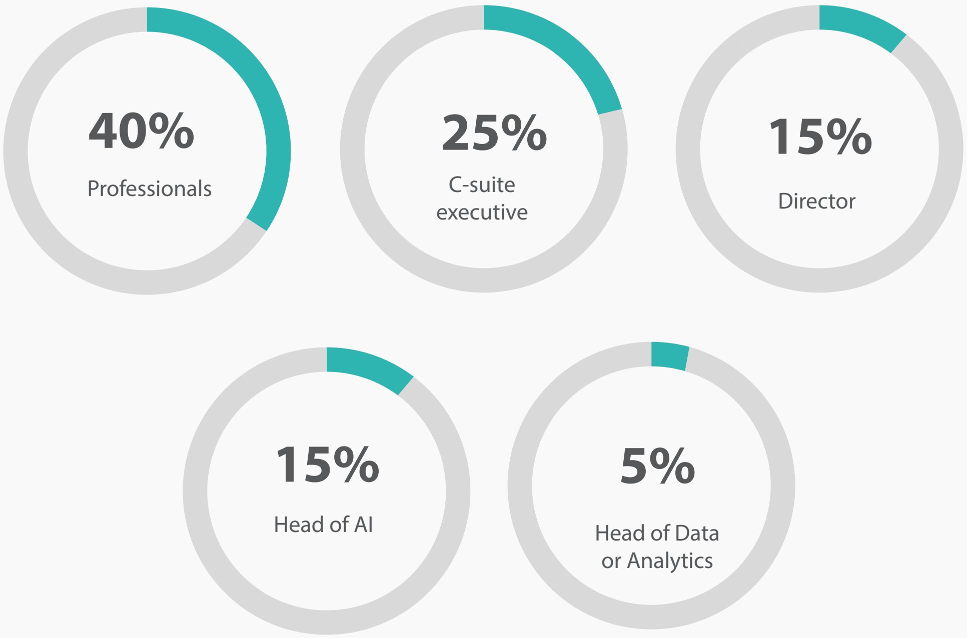
Stakeholder and Expert Consultation

Consultations with 20 representatives from industry were conducted across the region. The interviews were a key component to developing a cohesive and representative narrative that accurately captured the perspectives on trustworthy artificial intelligence. In addition to the these consultation sessions and our online survey, advisory and technical experts were engaged in the development of this report.

The research programme remains open and we will continue to collect data. A revised version of this report will be published in 2022 reflecting the new findings. If you would like to contribute to the research, please visit https://aiasiapacific.org/survey/.

What is the nature of your role within your organisation?

Figure 5: Participants



Most of the research participants hold leadership roles within their company, although not necessarily in the field of ethics or AI.

Endnotes

[1] Schwab, K. (2016). The Fourth Industrial Revolution. World Economic Forum.

[2] In 1956, John McCarthy, Minsky, Nathaniel Rochester and Claude E. Shannon coined the term "artificial intelligence" during the Dartmouth conference.

[3] AI HLEG. (2019). Ethics Guidelines for Trustworthy AI. European Commission. Retrieved from https://ec.europa.eu/

[4] AI HLEG. (2019). Ethics Guidelines for Trustworthy AI. European Commission. Retrieved from https://ec.europa.eu/

[5] Commonwealth Scientific and Industrial Research Organisation. (2019) Artificial Intelligence: Solving problems, growing the economy and improving our quality of life. Data16. Retrieved from https://data61.csiro.au/en/Our-Research/Our-Work/AI-Roadmap

[6] https://tracxn.com/explore/Artificial-Intelligence-Startups-in-China

[7] Tracxn, 2020. Artificial Intelligence Startups in Australia. Retrieved from https://tracxn.com/explore/Artificial-Intelligence-Startups-in-Australia

[8] Tracxn, 2020. Artificial Intelligence Startups in Singapore. Retrieved from https://tracxn.com/explore/Artificial-Intelligence-Startups-in-Singapore

[9] Tracxn, 2020. Artificial Intelligence Startups in New Zealand. Retrieved from https://tracxn.com/explore/Artificial-Intelligence-Startups-in-New-Zealand

[10] Hajkowicz SA1+, Karimi S1, Wark T1, Chen C1, Evans M1, Rens N3, Dawson D1, Charlton A2, Brennan T2, Moffatt C2, Srikumar S2, Tong KJ2 (2019) Artificial intelligence: Solving problems, growing the economy and improving our quality of life. CSIRO Data61, Australia.

[11] Dawson D and Schleiger E*, Horton J, McLaughlin J, Robinson C∞, Quezada G, Scowcroft J, and Hajkowicz S† (2019) Artificial Intelligence: Australia's Ethics Framework. Data61 CSIRO, Australia.

[12] AlphaBeta (2018) Digital Innovation: Australia's $315B Opportunity. CSIRO Data61, Australia. Retrieved from: https://data61.csiro.au/en/Our-Research/Our-Work/Future-Cities/Planning-sustainable-infrastructure/Digital-Innovation.

[13] Centre for the New Economy and Society. (2018). The Future of Jobs Report: 2018. World Economic Forum. Retrieved from https://www.weforum.org/reports/the-future-of-jobs-report-2018

[14] Dawson D and Schleiger E*, Horton J, McLaughlin J, Robinson C∞, Quezada G, Scowcroft J, and Hajkowicz S† (2019) Artificial Intelligence: Australia's Ethics Framework. Data61 CSIRO, Australia.

[15] State Council. (2020). Next Generation Artificial Intelligence Development Plan. Retrieved from: https://flia.org/notice-state-council-issuing-new-generation-artificial-intelligence-development-plan/

[16] Lorand Laskai and Graham Webster. (2019). Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI'. New America. Retrieved from https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/

[17] Roberts, H., Cowls, J., Morley, J. et al. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. AI & Soc 36, 59–77 (2021). https://doi.org/10.1007/s00146-020-00992-2

[18] Artificial Intelligence Forum New Zealand. (2018). Artificial Intelligence: Shaping a Future New Zealand. AI Forum New Zealand. Retrieved from https://aiforum.org.nz/2018/05/02/ai-forums-research-report-artificial-intelligence-shaping-a-future-new-zealand/

[19] Tertiary Education Commission. (2018). Enrolment. TEC. Retrieved from https://tec.govt.nz/funding/funding-and-performance/funding/fund-finder/student-achievement-component-provision-at-level-3-and-above-on-the-nzqf-fund/enrolment

[20] Morton, J. (2016). *Govt announces new $35 million fund to attract world-class researchers.* New Zealand Herald. Retrieved from https://www.nzherald.co.nz/nz/govt-announces-new-35-million-fund-to-attract-world-class-researchers/CCJMUOQMAN3CEIVXLMTY65YZ6U/.

[21] Ministry of Business, Innovation and Employment. (2017). *Strategic Science Investment Fund Investment Plan 2017-2024.* New Zealand Government. Retrieved from https://www.mbie.govt.nz/science-and-technology/science-and-innovation/funding-information-and-opportunities/investment-funds/strategic-science-investment-fund/

[22] New Zealand Government. (2020). Algorithm Charter for Aotearoa New Zealand. Statistics NZ. Retrieved from https://www.data.govt.nz/manage-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter

[23] World Economic Forum. (2020). Reimagining Regulation for the Age of AI: New Zealand Pilot Project, Retrieved from http://www3.weforum.org/docs/WEF_Reimagining_Regulation_Age_AI_2020.pdf

[24] Ministry of Business, Innovation and Employment. (2021). Moving Towards Responsible Government Use of AI in New Zealand. Digital Tech ITP. Retrieved from https://digitaltechitp.nz/2021/03/22/moving-towards-responsible-government-use-of-ai-in-new-zealand/

[25] Smart Nation and Digital Government Office. (2019). *National Artificial Intelligence Strategy: Advancing Our Smart Nation Journey.* Smart Nation Singapore. Retrieved from https://www.smartnation.gov.sg/why-Smart-Nation/NationalAIStrategy

[26] Released the Model AI Governance Framework (Second Edition) at the 2020 World Economic Forum Annual Meeting in Davos, Switzerland. Retrieved from https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

[27] We have analysed and compared the existing developments in these four countries as the methodology for measuring if they are satisfactory across the 6 different sources of information. We acknowledge that these principles and developments are constantly evolving.

[28] Zhiyuan News. (2019). Beijing AI Principles. BAAI. Retrieved from https://www.baai.ac.cn/news/beijing-ai-principles-en.html

[29] Dave Dawson et al. (2019). Artificial Intelligence: Australia's Ethics Framework (A Discussion Paper) Data 16 CSIRO, Australia. Retrieved from https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf

[30] Personal Data Protection Commission Singapore. (2020). Singapore's Approach to AI Governance. Retrieved from https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework

[31] New Zealand Government. (2020). Algorithm Charter for Aotearoa New Zealand. Statistics NZ. Retrieved from https://www.data.govt.nz/manage-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter/

[32] As detailed in chapter VII, China has been implementing robust privacy protection measures since 2018 with the Personal Information Security Specification. However, China's privacy legislation can be seen to contradict other initiatives such as the Social Credit System, which requires extensive personal data.

[33] Lisa LeVasseur et al. (2021). On the Importance of Human-Centricity and Data. World Economic Forum. Retrieved from http://www3.weforum.org/docs/WEF_On_the_Importance_of_Human_Centricity_2021.pdf

[34] Luciano Floridi et al. (2018). AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Minds and Machines

[35] United Nations. (1948). Universal Declaration of Human Rights, General Assembly Resolution 217A.

[36] LeVasseur et al (n 33).

37 https://rise.cs.berkeley.edu/blog/michael-i-jordan-artificial-intelligence%E2%80%8A-%E2%80%8Athe-revolution-hasnt-happened-yet/

38 Arvind Narayanan. (2018). 21 fairness definitions and their politics, FAT* Tutorial.

39 Mikael Munck and Agnes Antczak. (2020). Fairness in AI Retrieved from https://2021.ai/fairness-in-ai

40 Solon Barocas, Moritz Hardt, and Arvind Narayanan. (2019). Fairness and machine learning: Limitations and Opportunities. Retrieved from https://fairmlbook.org.

41 McKinsey Global Institute, Notes from the AI frontier: Tackling bias in AI (and in humans), June 2019. Available at < https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans> [Accessed 3 April 2021].

42 Jake Silberg and James Manyika. (2019). Notes from the AI frontier: Tackling bias in AI (and in humans) McKinsey Global Institute. Retrieved from https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.pdf

43 Amba Kak, ed., 'Regulating Biometrics: Global Approaches and Urgent Questions' AI Now Institute, September 1 2020, https://ainowinstitute.org/regulatingbiometrics.html.

44 Serafimova, Silviya. (2020). Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. Humanities and Social Sciences Communications. 7. 10.1057/s41599-020-00614-8.

45 Barton, D., Woetzel, J., Seong, J., & Tian, Q. (2017). Artificial Intelligence: Implications for China (Paper Presentation). 2017 China Development Forum. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/China/Artificial%20intelligence%20Implications%20for%20China/MGI-Artificial-intelligence-implications-for-China.ashx

46 Dave Dawson et al (n 29).

47 Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Volume 58, Pages 82-115, ISSN 1566-2535, Retrieved from https://doi.org/10.1016/j.inffus.2019.12.012

48 Ranulph Glanvile. (2009). Black Boxes. Cybernetics and Human Knowing. Volume 16, Numbers 1-2, 2009, pp. 153-167(15).

49 Alejandro Barredo Arrieta et al (n 47).

50 Office of the Victorian Commissioner. (2018). Artificial Intelligence and Privacy (Issues Paper). OVIC. Retrieved from https://ovic.vic.gov.au/blog/artificial-intelligence-and-privacy-issues-paper/

51 See Appendix for more information.

52 Oliver Smith, Koa Health.

53 World Economic Forum. (2021). World Leaders to Meet During Davos Agenda in a Crucial Year to Rebuild Trust. Retrieved from https://www.weforum.org/press/2021/01/world-leaders-to-meet-during-davos-agenda-in-a-crucial-year-to-rebuild-trust-51d7fa48d1

54 Edelman AI. (2019). 2019 Edelman AI Survey. Retrieved from https://www.edelman.com/

55 TigTech. (2021) Trust & Tech Governance, Retrieved from https://www.tigtech.org

56 G20. (2019). Ministerial Statement on Trade and Digital Economy. Osaka: G20.

57 OECD. (2021). Recommendation of the Council on Artificial Intelligence, Paris: Secretary-General of the OECD.

58 SkillsFuture is a national movement to provide Singaporeans with the opportunities to develop their fullest potential throughout life, regardless of their starting point. Retrieved from: https://www.skillsfuture.gov.sg.

59 United Nations. (2019). A United Nations system-wide strategic approach and road map for supporting capacity development on artificial intelligence, Geneva: Cief Executives Board for Coordination.

60 The 17 Sustainable Development Goals (SDGs) adopted by all United Nations Member States in 2015.

61 Ibid.

62 UNESCO. (2020). International cooperation is key to inclusive AI. Retrieved from https://en.unesco.org/news/international-cooperation-key-inclusive-ai

63 UNESCO. (2020). Artificial Intelligence and Gender Equality. France: UNESCO.

64 MacDonald, K. & Madzou, L. (2020). AI is here. This is how it can benefit everyone. Retrieved from https://www.weforum.org/agenda/2020/09/ai-is-here-this-is-how-it-can-benefit-everyone/

65 AI Asia Pacific Institute. (2020). AI Asia Pacific Institute Podcast. s.l.:AI Asia Pacific Institute. Retrieved from https://podcasts.apple.com/us/podcast/26-centre-excellence-to-champion-ethical-use-ai-kate/id1460402348?i=1000498926640.

66 Beard, M. & Longstaff, S. (2018). Ethical by Design: Principles for Good Technology, Sydney: The Ethics Centre.

67 Goddard, M. (2017). The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. International Journal of Market Research. 59(6).

68 Data Guidance. (2020). Singapore-Data Protection Overview. Data Guidance. Retrieved from https://www.dataguidance.com/notes/singapore-data-protection-overview

69 Zhang, G., Yin, K. (2020). A look at China's draft of Personal Information Protection Law. IAPP. Retrieved from https://iapp.org/news/a/a-look-at-chinas-draft-of-personal-data-protection-law/

70 https://www.apple.com/au/newsroom/2021/01/data-privacy-day-at-apple-improving-transparency-and-empowering-users/

71 https://www.ag.gov.au/rights-and-protections/privacy/asia-pacific-economic-cooperation-and-privacy

72 https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

73 https://www.varonis.com/blog/us-privacy-laws/

74 https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr_en

75 https://www2.deloitte.com/content/dam/Deloitte/sg/Documents/risk/sea-risk-unity-diversity-privacy-guide.pdf

76 Chitturu, S. (2017). Artificial Intelligence and Southeast Asia's Future. McKinsey Global Institute.

77 MIT Technology Review. (2019). Asia's AI Agenda: AI and human capital. MIT Technology Review. Retrieved from https://www.technologyreview.com/2019/05/10/135421/asias-ai-agenda-ai-and-human-capital/

78 Mathur, R. (2019). Creating human impact with artificial intelligence. Avanade Insights. Retrieved from https://www.avanade.com/en/blogs/avanade-insights/artificial-intelligence/human-impact-with-ai

79 Singapore Advisory Council of the Ethical Use of AI and Data. (2020). Job Redesign in the Age of AI.

80 The Ethics Centre and Deloitte Access Economics. The Ethical Advantage: The Economic and Social Benefits of Ethics to Australia. (2020). Retrieved from https://ethics.org.au/wp-content/uploads/2018/05/The-Ethical-Advantage-4.pdf/.

81 Formerly a Federated Hermes employee.

82 Lockey, S., Gillespie, N., & Curtis, C. (2020). Trust in Artificial Intelligence: Australian Insights. The University of Queensland and KPMG Australia. Retrieved from doi.org/10.14264/b32f129

83 https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence/

84 Amba Kak, ed., 'Regulating Biometrics: Global Approaches and Urgent Questions' AI Now Institute, September 1 2020, https://ainowinstitute.org/regulatingbiometrics.html.

AI ASIA PACIFIC
INSTITUTE

contact@aiasiapacific.org
aiasiapacific.org