Introduction of a Model Artificial Intelligence Governance Framework

Background

On 23 January 2019, Singapore released its Model Artificial Intelligence Governance Framework ("**Model Framework**") at the World Economic Forum Annual Meeting in Davos, Switzerland for public consultation, pilot adoption, and feedback. The Model Framework builds on the earlier guidelines detailed in a paper released on 5 June 2018 by the Personal Data Protection Commission ("**PDPC**") and the Infocomm Media Development Authority ("**IMDA**").

The Model Framework is a voluntary accountability-based general framework that is algorithm-agnostic, technology-agnostic and sector-agnostic. It complements other discussion papers on the topic, such as the Monetary Authority of Singapore's "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector" released on 12 November 2018.

The Model Framework seeks to provide guidance on the key issues to be considered when deploying artificial intelligence ("**AI**") solutions at scale, and the measures that can be implemented to manage corresponding risks. The Model Framework is intended to assist organisations to:

(a) Build consumer confidence in AI through organisations' responsible use of such technologies to mitigate different types of risks in AI deployment; and

(b) Demonstrate reasonable efforts to align internal policies, structures and processes with relevant accountability-based practices in data management and protection, e.g. the Personal Data Protection Act 2012 ("**PDPA**") and Organisation for Economic Cooperation and Development (OECD) Privacy Principles.

The Model Framework

Overview

The Model Framework is based on two high-level guiding principles that seek to promote trust in and understanding of the use of AI technologies:

(a) Organisations using AI in decision-making should ensure that the decision-making process is explainable, transparent and fair. Although perfect explainability, transparency and fairness are impossible to attain, organisations should strive to ensure that their use or application of AI is undertaken in a manner that reflects the objectives of these principles. This helps in building trust and confidence in AI; and

(b) AI solutions should be human-centric. As AI is used to amplify human capabilities, the protection of the interests of human beings, including their well-being and safety, should be primary considerations in the design, development and deployment of AI.

Building on these guiding principles, the Model Framework comprises guidance on measures in four key areas promoting the responsible use of AI that organisations should adopt: (a) internal governance structures and measures; (b) determining the AI decision-making model; (c) operations management; and (d) customer relationship management.

Internal governance structure and measures

First, the Model Framework provides that organisations should have internal governance structures and measures to ensure robust oversight of the organisation's use of AI. In determining the appropriate features that the organisation's internal governance structures ought to have, organisations may adapt existing or set up new internal governance structures to incorporate corporate values, enterprise risks and responsibilities relating to algorithmic decision-making.

For example, organisations may consider allocating clear roles and responsibilities of the various stages and activities involved in AI deployment to appropriate personnel and/or departments in the organisation. The organisation may also wish to consider establishing a coordinating body to oversee the AI deployment in question and ensure the personnel involved are adequately trained, and provided with the requisite resources and guidance for them to discharge their allocated duties.

As part of the internal governance structure, organisations may also consider implementing a sound system of risk management and internal controls to address the risks involved in the deployment of AI solutions. This risk management system and internal controls may include efforts to ensure the quality of the datasets (i.e. addressing the risks of inaccuracy or bias in datasets) used for AI model training and the establishment of monitoring and reporting systems of performance and other issues pertaining to the deployed AI.

Determining AI decision-making model

Secondly, the Model Framework seeks to provide a risk-management methodology to aid organisations in setting its risk appetite for the use of AI. It requires organisations to weigh their commercial objectives of using AI against its risks, using their organisations' corporate values to guide this assessment. The documentation of this process through a periodic risk impact assessment will, therefore, help organisations develop clarity and confidence in using the AI solutions, and better respond to potential challenges from individuals, other organisations and regulators.

The Model Framework identifies three broad AI decision-making models with varying degrees of human oversight in the decision-making process:

(a) **Human-in-the-loop**: This model refers to the situation where human oversight is active and involved, with the human retaining full control and the AI only providing recommendations or input. Hence, the human will make informed decisions based on the information provided by AI (e.g. factors that are used in the decision, their value and weighting, and any correlation). An example of such a model is the usage of AI to identify potential diagnoses and treatments for an unfamiliar medical condition, with the doctor retaining the ultimate final decision on the diagnosis and corresponding treatment;

(b) **Human-out-of-loop:** This model refers to the situation where there is no human oversight over the execution of decisions. AI, therefore, has full control without the option of a human override.

An example would be the usage of a product recommendation solution which automatically suggest products and services to individuals based on pre-determined demographic and behavioural profiles; and

(c) **Human-over-the-loop:** This model refers to the situation where the human is still allowed to adjust parameters during the execution of the algorithm. An example would be the case where a GPS navigation system plans and offers the routes for the driver to pick, but the driver can still alter parameters during the trip without having to re-programme the route.

The Model Framework also proposes a matrix to classify the probability and severity of harm to an individual as a result of the decision made by the organisation about that individual. In determining the appropriate decision-making model for the AI solution, the organisation should, therefore, consider the impact of such a decision on the individual using the probability-severity of harm matrix. For example, in situations involving safety-critical systems, organisations may wish to employ a human-in-the-loop decision-making model to ensure that a person is still allowed to make meaningful decisions or to shut down the system where control is not safely available. Conversely, a human-out-of-loop decision-making model may be appropriate for situations where there is a low severity and probability of harm.

	Probability of harm	
Probability of harm	**High** severity **Low** probability	**High** severity **High** probability
	Low severity **Low** probability	**Low** severity **High** probability

Table 1: Probability-Severity of Harm Matrix

Operations management

Thirdly, the Model Framework provides some guidance in respect of the deployment process of an AI solution, focusing in particular on the interaction between data and the algorithm.

The Model Framework highlights the need to put in place good data accountability practices to ensure the effectiveness of an AI solution, including:

(a) **Understanding the data lineage:** Understanding where the data originally came from, how it was collected, curated and moved within the organisation, and how its accuracy is maintained over time. Keeping a data provenance record allows an organisation to ascertain the quality of the data based on its origin and subsequent transformation, trace potential sources of errors, update data, and attribute data to their sources.

(b) **Ensuring data quality:** Understanding and addressing factors that may affect data quality, such as the accuracy, completeness, veracity, recency, relevance, integrity and usability of the dataset.

(c) **Minimising inherent bias:** Identifying and addressing inherent biases in the dataset (i.e. selection bias and measurement bias). The steps that organisations can take to mitigate the risk of inherent bias include the usage of a heterogeneous dataset from a variety of reliable sources, and ensuring the completeness of dataset used (in terms of data attributes and data items).

(d) **Different data sets:** Using different datasets for training, testing and validation of the AI model where applicable, be cognisant of the risks of systematic bias and put in place appropriate safeguards.

(e) **Periodic reviewing and updating of datasets:** Reviewing datasets periodically to ensure accuracy, quality, currency, relevance and reliability. Where necessary, the datasets should also be updated with new input data obtained from actual use of the deployed AI models.

The Model Framework also highlights that organisations should consider measures to enhance the transparency of algorithms found in AI models through concepts of explainability, repeatability and traceability:

(a) **Explainable:** Explainable AI can be achieved by explaining how deployed AI models' algorithms function and/or how the decision-making process incorporates model predictions. Organisations should, therefore, provide descriptions of the AI solutions' design and expected behaviour to demonstrate accountability to individuals and/or regulators. Nevertheless, it is recognised that there may be scenarios where it might not be practical or reasonable to provide information concerning an algorithm (i.e. contexts of proprietary information, intellectual property, information security, etc.). To further demonstrate accountability, it is also provided that organisations should document the model training and selection processes, the reasons for which decisions are made, and measures taken to address identified risks. In particular, organisations seeking to use "auto-machine learning" which automates the iterative process for the best model ought to consider the transparency, explainability, and traceability of these algorithms. In the right circumstances, algorithm audits can also be carried out to discover the actual operations of algorithms comprised in models.

(b) **Repeatability:** Where explainability cannot practically be achieved (e.g. black box), organisations can then consider documenting the repeatability of results produced by the AI model. This ensures that the algorithm can consistently perform an action or make a decision given the same scenario thereby giving AI users a certain degree of confidence.

(c) **Traceability:** Organisations should also ensure that their AI model is traceable (i.e. its decision-making processes are documented in an easily understandable way). The information recorded may then be used as a source of input data in the future as a training dataset, for troubleshooting, or an investigation into how the model was functioning or why a particular prediction was made.

Organisations should also ensure active monitoring, review and tuning of their AI models even after deployment in the real-world environment so as to cater for changes to customer behaviour over time and to refresh models based on updated training datasets that incorporate new input data.

Customer relationship management

Finally, the Model Framework lays out various communication strategies in the deployment of AI to inspire trust in the building and maintenance of open relationships between organisations and individuals. In particular:

(a) Organisations should have a developed policy on what explanations to provide to individuals as part of general communication or in respect of specific decision upon request. Organisations should adopt greater transparency in their communication strategy. This may manifest in the form of a general disclosure by way of provision of general information in an easy-to-understand language on whether AI is used in their products and/or services and information on how AI is used in decision-making about individuals, and the role and extent that AI plays in that decision-making process;

(b) Where there are user interfaces, organisations ought to ensure that the user interface serves its intended purposes. In the case of a chatbot, organisations should manage individuals' expectations by informing that they are interacting with a chatbot as opposed to an individual being. The organisation should also inform individuals if their replies to the chatbot would be used to train the AI system;

(c) Organisations should decide whether to allow individuals to opt-out, either by default or upon request only, depending on the various considerations including the degree of harm to individuals, technical feasibility etc.; and

(d) Organisations should put in place communication channels (i.e. feedback channel and/or decision review channel) for their customers, to enable individuals to raise feedback or queries and also to provide an avenue for individuals to request a review of material AI decisions that have affected them.

Comment

The Model Framework represents the culmination of efforts by policy makers and regulators in Singapore to articulate a common AI governance approach, and a set of consistent principles relating to the responsible use of AI. This serves to provide clarity and certainty to industry players, while promoting the responsible adoption of AI and harnessing the benefits from technological advancements.

Similar efforts have also been undertaken in other jurisdictions to seek to regulate AI activities. For example, under Article 22 of the European Union's General Data Protection Regulation ("**GDPR**"), individuals have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning the data subject or similarly significantly affects him or her, unless the decision is authorised by law of the EU member states, necessary for the preparation and execution of a contract, or done with the individual's explicit consent.

Moving forward, the Model Framework remains a 'living document' that will continue to evolve with industry feedback. IMDA and the World Economic Forum's Centre for the Fourth Industrial Revolution will be collaborating to discuss the Model Framework in further detail and to facilitate its adoption. In particular, both parties will develop a measurement matrix for the Model Framework which regulators and certification bodies globally can adopt and adapt for their use in assessing whether organisations are responsibly deploying AI.

A copy of the Model Framework can be accessed at:
https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/A-Proposed-Model-AI-Governance-Framework-January-2019.pdf

If you would like information on this or any other area of law, you may wish to contact the partner at WongPartnership that you normally deal with or any of the following partners:

LAM Chung Nian
Head – Intellectual Property, Technology and Media, Telecommunications and Data Protection Practices
d +65 6416 8271
e chungnian.lam@wongpartnership.com
Click here to view Chung Nian's CV.

Kylie PEH
Partner – Intellectual Property, Technology and Media, Telecommunications and Data Protection Practices
d +65 6416 8259
e kylie.peh@wongpartnership.com
Click here to view Kylie's CV.

WPG MEMBERS AND OFFICES

- contactus@wongpartnership.com

SINGAPORE

-
WongPartnership LLP
12 Marina Boulevard Level 28
Marina Bay Financial Centre Tower 3
Singapore 018982
t +65 6416 8000
f +65 6532 5711/5722

CHINA

-
WongPartnership LLP
Beijing Representative Office
Unit 3111 China World Office 2
1 Jianguomenwai Avenue, Chaoyang District
Beijing 100004, PRC
t +86 10 6505 6900
f +86 10 6505 2562

-
WongPartnership LLP
Shanghai Representative Office
Unit 1015 Corporate Avenue 1
222 Hubin Road
Shanghai 200021, PRC
t +86 21 6340 3131
f +86 21 6340 3315

MYANMAR

-
WongPartnership Myanmar Ltd.
Junction City Tower, #09-03
Bogyoke Aung San Road
Pabedan Township, Yangon
Myanmar
t +95 1 925 3737
f +95 1 925 3742

INDONESIA

-
Makes & Partners Law Firm
Menara Batavia, 7th Floor
Jl. KH. Mas Mansyur Kav. 126
Jakarta 10220, Indonesia
t +62 21 574 7181
f +62 21 574 7180
w makeslaw.com

MALAYSIA

-
Foong & Partners
Advocates & Solicitors
13-1, Menara 1MK, Kompleks 1 Mont' Kiara
No 1 Jalan Kiara, Mont' Kiara
50480 Kuala Lumpur, Malaysia
t +60 3 6419 0822
f +60 3 6419 0823
w foongpartners.com

MIDDLE EAST

-
Al Aidarous International Legal Practice
Abdullah Al Mulla Building, Mezzanine Suite 02
39 Hameem Street (side street of Al Murroor Street)
Al Nahyan Camp Area
P.O. Box No. 71284
Abu Dhabi, UAE
t +971 2 6439 222
f +971 2 6349 229
w aidarous.com

-
Al Aidarous International Legal Practice
Zalfa Building, Suite 101 - 102
Sh. Rashid Road
Garhoud
P.O. Box No. 33299
Dubai, UAE
t +971 4 2828 000
f +971 4 2828 011

PHILIPPINES

-
ZGLaw
27/F 88 Corporate Center
141 Sedeño Street, Salcedo Village
Makati City 1227, Philippines
t +63 2 889 6060
f +63 2 889 6066
w zglaw.com/~zglaw

wongpartnership.com