A PROPOSED MODEL ARTIFICIAL INTELLIGENCE GOVERNANCE FRAMEWORK

January 2019

TABLE OF CONTENTS

FOREWORD

From the well-publicised achievements of Google's DeepMind, SenseTime's technologies on facial recognition, to the ubiquitous presence of virtual assistants like Apple's *Siri* or Amazon's *Alexa*, Artificial Intelligence ("AI") is now a growing part of our lives. AI has delivered many benefits, from saving time to diagnosing hitherto unknown medical conditions, but it has also been accompanied by new concerns such as over personal privacy and algorithmic biases.

Amid such rapid technological advances and evolutions in business models, policy makers and regulators must embrace innovation in equal measure. The genesis of this Model AI Governance Framework ("Model Framework") can be traced to efforts by policy makers and regulators in Singapore to articulate a common AI governance approach and a set of consistent definitions and principles relating to the responsible use of AI, so as to provide greater certainty to industry players and promote the adoption of AI while ensuring that regulatory imperatives are met. This Model Framework is adapted from a discussion paper issued by the Personal Data Protection Commission (PDPC) in June 2018.

The first edition of this accountability-based Model Framework aims to frame the discussions around the challenges and possible solutions to harnessing AI in a responsible way. The Model Framework aims to collect a set of principles, organise them around key unifying themes, and compile them into an easily understandable and applicable structure. It seeks to equip its user with the tools to anticipate and eventually overcome these potential challenges in a practical way.

The Model Framework is Singapore's attempt to contribute to the global discussion on the ethics of AI by providing a framework that helps translate ethical principles into pragmatic measures that businesses can adopt. The Model Framework has been developed in consultation with academics, industry leaders and technologists from different backgrounds and jurisdictions. This diversity of views reflects the desire of the PDPC, the Info-communications Media Development Authority (IMDA), and the Advisory Council on the Ethical Use of AI and Data, to shape plans for Singapore's AI ecosystem in a collaborative and inclusive manner.

Where AI is concerned, there are big questions to be answered, and even bigger ones yet to be asked. The Model Framework may not have all the answers, but it represents a firm start and provides an opportunity for all – individuals and organisations alike – to grapple with fundamental ideas and practices that may prove to be key in determining the development of AI in the years to come.

S Iswaran
Minister for Communication and Information
Singapore
January 2019

1. PREAMBLE

1.1 The Model AI Governance Framework ("Model Framework") focuses primarily on four broad areas: internal governance, decision-making models, operations management and customer relationship management. While the Model Framework is certainly not limited in ambition, it is ultimately limited by form, purpose and practical considerations of scope. With that in mind, several caveats bear mentioning: the Model Framework is —

a. Algorithm-agnostic. It *does not* focus on specific AI or data analytics methodology. It applies to the design, application and use of AI in general;

b. Technology-agnostic. It *does not* focus on specific systems, software or technology, and will apply regardless of development language and data storage method; and

c. Sector-agnostic. It *serves as a baseline set* of considerations and measures for organisations operating in any sector to adopt. Specific sectors or organisations may choose to include additional considerations and measures or adapt this baseline set to meet their needs.

1.2 It is recognised that there are a number of issues that are closely interrelated to the ethical use and deployment of AI. This Model Framework *does not* focus on these specific issues, which are often sufficient in scope to warrant separate study and treatment. Examples of these issues include:

a. Articulating a set of ethical principles for AI. There are a number of attempts globally in establishing a set of principles. While there is a consistent core set of ethical principles, there is also a penumbra of variation across cultures, jurisdictions and industry sectors. The Model Framework does not set out to propose another set of such principles although it compiles a glossary from existing literature.

b. Providing Model Frameworks and addressing issues around data sharing, whether between the public and private sectors or between organisations or within consortia. There are a number of guides that are relevant, i.e. the PDPC Guide to Data Sharing and the Guide to Data Valuation for Data Sharing.

c. Discussing issues relating to the legal liabilities associated with AI, intellectual property rights and societal impacts of AI, e.g. on employment, competition, unequal access to AI products and services by different segments of society, AI technologies falling into hands of wrong people, etc. These issues are nevertheless pertinent and will be explored separately through the Centre for AI and Data Governance established in the Singapore Management University School of Law or other relevant forums.

2. INTRODUCTION

Objectives

2.1 The exponential growth in data and computing power has fuelled the advancement of data-driven technologies such as Artificial Intelligence ("AI"). AI can be used by organisations to provide new goods and services, boost productivity, enhance competitiveness, ultimately leading to economic growth and better quality of life. As with any new technologies, however, AI also introduces new ethical, legal and governance challenges. These include risks of unintended discrimination potentially leading to unfair outcomes, as well as issues relating to consumers' knowledge about how AI is involved in making significant or sensitive decisions about them.

2.2 The Personal Data Protection Commission (PDPC), Infocomm Media Development Authority (IMDA), with the advice from the Advisory Council on the Ethical Use of AI and Data ("Advisory Council"), proposes for consultation this first edition of a voluntary Model Framework as a general, ready-to-use tool to enable organisations that are deploying AI solutions at scale to do so in a responsible manner. This Model Framework is not intended for organisations that are deploying updated commercial off-the-shelf software packages that happen to now incorporate AI in their feature set.

2.3 This voluntary Model Framework provides guidance on the key issues to be considered and measures that can be implemented. Adopting this Model Framework entails tailoring the measures to address the risks identified for the implementing organisation. The Model Framework is intended to assist organisations to achieve the following objectives:

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

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) and OECD Privacy Principles.

2.4 The extent to which organisations adopt the recommendations in this Model Framework depends on several factors, including the nature and complexity of the AI used by the organisations; the extent to which AI is employed in the organisations' decision-making; and the severity and probability of the impact of the autonomous decision on the individuals. To elaborate: AI may be used to augment a human decision-maker or to autonomously make a decision. The impact on an individual of an autonomous decision in, for example, medical diagnosis will be greater than in processing a bank loan. The commercial risks of AI deployment would therefore be proportional to the impact on individuals. It is also recognised that where the cost of implementing AI technologies in an ethical manner outweighs the expected benefits, organisations should consider whether alternative non-AI solutions should be adopted.

Guiding Principles

2.5 The Model Framework is based on two high-level guiding principles that promote trust in AI 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 build trust and confidence in AI.

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 the primary considerations in the design, development and deployment of AI.

2.6 AI technology joins a line of technologies whose purpose is to increase the productivity of humankind. Unlike earlier technologies, there are some aspects of autonomous predictions that may not be fully explainable. This Model Framework should be used by organisations that rely on AI's autonomous predictions to make decisions that affect individuals, or have significant impact on society, markets or economies.

2.7 Organisations should detail a set of ethical principles when they embark on deployment of AI at scale within their processes or to empower their products and/or services. As far as possible, organisations should also review their existing corporate values and incorporate the ethical principles that they have articulated. Some of the ethical principles may be articulated as risks that can be incorporated into the corporate risk management framework. The Model Framework is designed to assist organisations by incorporating ethical principles into familiar, pre-existing corporate governance structures and thereby aid in guiding the adoption of AI in an organisation. Where necessary, organisations may wish to refer to the **Glossary** of AI ethical values included at the end of the Model Framework (See Annex B).

Assumptions

2.8 The Model Framework aims to discuss good data management practices in general. They may be more applicable to big data AI models than pure decision tree driven AI models or small data set AI methods such as transfer learning, or use of synthetic data.

2.9 The Model Framework does not address the risk of catastrophic failure due to cyber-attacks on an organisation heavily dependent on AI. Organisations remain responsible for ensuring the availability, reliability, quality and safety of their products and services, regardless of whether AI technologies are used.

2.10 Adopting this voluntary Model Framework will not absolve organisations from compliance with current laws and regulations. However, as this is an accountability-based framework, adopting it will assist in demonstrating that they had implemented accountability-based practices in data management and protection, e.g. the Personal Data Protection Act (2012) and OECD Privacy Principles.

Definitions

2.11 The following simplified diagram depicts the key stakeholders in an AI adoption process discussed in the Model Framework:

AI Solution Provider Organisation Individuals

2.12 Some terms used in AI may have different definitions depending on context and use. The definitions of some key terms used in this Model Framework are as follows:

"Artificial Intelligence (AI)" refers to a set of technologies that seek to simulate human traits such as knowledge, reasoning, problem solving, perception, learning and planning. AI technologies rely on AI algorithms to generate models. The most appropriate model(s) is/are selected and deployed in a production system.

"AI Solution Providers" develop AI solutions or application systems that make use of AI technology. These include not just commercial off-the-shelf products, online services, mobile applications, and other software that consumers can use directly, but also business-to-business-to-consumer applications, e.g. AI-powered fraud detection software sold to financial institutions. They also include device and equipment manufacturers that integrate AI-powered features into their products, and those whose solutions are not standalone products but are meant to be integrated into a final product. Some organisations develop their own AI solutions and can be their own solution providers.

"Organisations" refers to companies or other entities that adopt or deploy AI solutions in their operations, such as backroom operations (e.g. processing applications for loans), front-of-house services (e.g. e-commerce portal or ride-hailing app), or the sale or distribution of devices that provide AI-powered features (e.g. smart home appliances).

"Individuals", depending on the context, can refer to persons to whom organisations intend to supply AI products and/or services, or persons who have already purchased the AI products and/or services. These may be referred to as "consumers" or "customers" as well.

3. MODEL AI GOVERNANCE FRAMEWORK

3.1 This Model Framework comprises guidance on measures promoting the responsible use of AI that organisations should adopt in the following key areas:

a. **Internal Governance Structures and Measures:** Adapting existing or setting up internal governance structure and measures to incorporate values, risks, and responsibilities relating to algorithmic decision-making.

b. **Determining AI Decision-Making Model:** A methodology to aid organisations in setting its risk appetite for use of AI, i.e. determining acceptable risks and identifying an appropriate decision-making model for implementing AI.

c. **Operations Management:** Issues to be considered when developing, selecting and maintaining AI models, including data management.

d. **Customer Relationship Management:** Strategies for communicating to consumers and customers, and the management of relationships with them.

3.2 Where not all elements of this Model Framework apply, organisations should adopt the relevant elements. An illustration of how this Model Framework can be adopted by an organisation is in Annex C.

Internal Governance Structures and Measures

3.3 Organisations should have internal governance structures and measures to ensure robust oversight of the organisation's use of AI. The organisation's existing internal governance structures can be adapted, and/or new structures can be implemented if necessary. For example, risks associated with the use of AI can be managed within the enterprise risk management structure; ethical considerations can be introduced as corporate values and managed through ethics review boards or similar structures. Organisations should also determine the appropriate features in their internal governance structures. For example, when relying completely on a centralised governance mechanism is not optimal, a de-centralised one could be considered to incorporate ethical considerations into day-to-day decision-making at operational level, if necessary. The sponsorship, support and participation of the organisation's top management and its Board in the organisation's AI governance are crucial.

3.4 Organisations should include some or all of the following features in their internal governance structure:

1. <u>Clear roles and responsibilities for the ethical deployment of AI</u>
   a. Responsibility for and oversight of the various stages and activities involved in AI deployment should be allocated to the appropriate personnel and/or

departments. If necessary and possible, consider establishing a coordinating body, having relevant expertise and proper representation from across the organisation.

b. Personnel and/or departments having internal AI governance functions should be fully aware of their roles and responsibilities, be properly trained, and be provided with the resources and guidance needed for them to discharge their duties.

c. Key roles and responsibilities that should be allocated include:
   i. Using any existing risk management framework and applying risk control measures (See further "Risk management and internal controls" below) to
      - Assess and manage the risks of deploying AI (including any potential adverse impact on the individuals, e.g. who are most vulnerable, how are they impacted, how to assess the scale of the impact, how to get feedback from those impacted, etc.)
      - Decide on appropriate AI decision-making models.
      - Manage the AI model training and selection process.

   ii. Maintenance, monitoring and review of the AI models that have been deployed, with a view to taking remediation measures where needed.

   iii. Reviewing communications channels and interactions with consumers and customers with a view to providing disclosure and effective feedback channels.

   iv. Ensuring relevant staff dealing with AI systems are trained in interpreting AI model output and decisions.

2. <u>Risk management and internal controls</u>

a. A sound system of risk management and internal controls, specifically addressing the risks involved in the deployment of the selected AI model, should be implemented.

b. Such measures include:
   i. Using reasonable efforts to ensure that the datasets used for AI model training are adequate for the intended purpose, and to assess and manage the risks of inaccuracy or bias, as well as reviewing exceptions identified during model training. Virtually, no dataset is completely unbiased. Organisations should strive to understand the ways in which datasets may be biased and address this in their safety measures and deployment strategies.

   ii. Establishing monitoring and reporting systems as well as processes to ensure that the appropriate level of management is aware of the performance of and other issues relating to the deployed AI. Where appropriate, the monitoring

can include autonomous monitoring to effectively scale human oversight. AI systems can be designed to report on the confidence level of their predictions, and explainability features can focus on why the AI model had a certain level of confidence, rather than why a prediction was made.

iii. Ensuring proper knowledge transfer whenever there are changes in key personnel involved in AI activities. This will reduce the risk of staff movement creating a gap in internal governance.

iv. Reviewing the internal governance structure and measures when there are significant changes to organisational structure or key personnel involved.

v. Periodically reviewing the internal governance structure and measures to ensure their continued relevance and effectiveness.

Determining AI Decision-Making Model

3.5 Prior to deploying AI solutions, organisations should decide on their commercial objectives of using AI, e.g. ensuring consistency in decision-making, improving operational efficiency and reducing costs, or introducing new product features to increase consumer choice. Organisations then weigh them against the risks of using AI in the organisation's decision-making. This assessment should be guided by organisations' corporate values, which in turn, could reflect the societal norms of the territories in which the organisations operate.

3.6 Organisations operating in multiple countries should consider the differences in societal norms and values, where possible. For example, gaming advertisement may be acceptable in one country but not in the other. Even within a country, risks may vary significantly depending on where AI is deployed. For example, risks to individuals associated with recommendation engines that promote products in an online mall or automating the approval of online applications for travel insurance may be lower than those associated with algorithmic trading facilities offered to sophisticated investors.

3.7 Some risks to individuals may only manifest at group level. For example, widespread adoption of a stock recommendation algorithm might cause herding behaviour, increasing overall market volatility if sufficiently large numbers of individuals make similar decisions at the same time. In addition to risks to individuals, other types of risks may also be identified, e.g. risk to an organisation's commercial reputation.

3.8 Organisations' weighing of their commercial objectives against the risks of using AI should be guided by their corporate values. Organisations can assess if the intended AI deployment and the selected model for algorithmic decision-making are consistent with their own core values. Any inconsistencies and deviations should be conscious decisions made by the organisations with a clearly defined and documented rationale.

3.9 As identifying commercial objectives, risks and selection of an appropriate decision-making model is an iterative and ongoing process, organisations should continually identify and review risks relevant to their technology solutions, mitigate those risks, and maintain a response plan should mitigation fail. Documenting this process through a periodically reviewed **risk impact assessment** helps organisations develop clarity and confidence in using the AI solutions. It will also help organisations respond to potential challenges from individuals, other organisations or businesses and regulators.

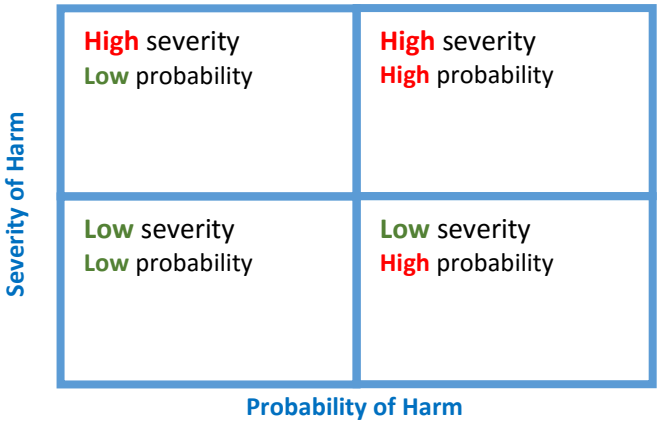
3.10 Based on the risk management approach described above, the Model Framework identifies three broad decision-making models with varying degrees of human oversight in the decision-making process:

a. **Human-in-the-loop.** This model suggests that human oversight is active and involved, with the human retaining full control and the AI only providing recommendations or input. Decisions cannot be exercised without affirmative actions by the human, such as a human command to proceed with a given decision. For example, a doctor may use AI to identify possible diagnoses of and treatments for an unfamiliar medical condition. However, the doctor will make the final decision on the diagnosis and the corresponding treatment. This model requires AI to provide enough information for the human to make an informed decision (e.g. factors that are used in the decision, their value and weighting, correlations).

b. **Human-out-of-the-loop.** This model suggests that there is no human oversight over the execution of decisions. AI has full control without the option of human override. For example, a product recommendation solution may automatically suggest products and services to individuals based on pre-determined demographic and behavioural profiles. AI can also dynamically create new profiles, then make product and service suggestions rather than relying on predetermined categories. A machine learning model might also be used by an airline to forecast demand or likely disruptions, and the outputs of this model are used by a solver module to optimise the airline's scheduling, without a human in the loop.

c. **Human-over-the-loop.** This model allows humans to adjust parameters during the execution of the algorithm. For example, a GPS navigation system plans the route from Point A to Point B, offering several possible routes for the driver to pick. The driver can alter parameters (e.g. due to unforeseen road congestions) during the trip without having to re-programme the route.

3.11 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 an organisation about that individual. The definition of harm and the computation of probability and severity depend on the context and vary from sector to sector. For example, the harm associated with a wrong diagnosis of a patient's medical condition will differ from that associated with a wrong product recommendation for apparels.



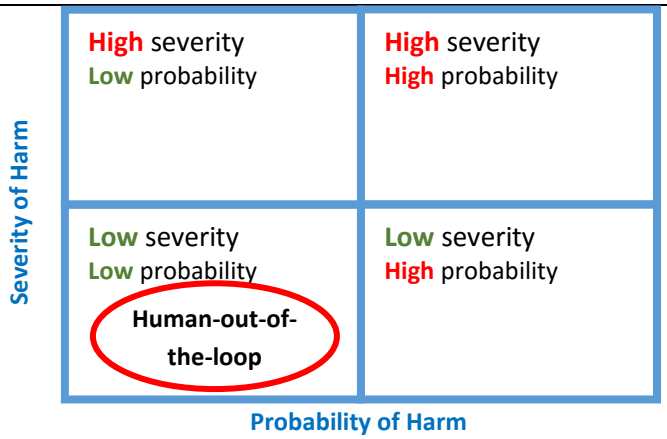
3.12 In determining the level of human oversight in an organisation's decision-making process involving AI, the organisation should consider the impact of such a decision on the individual using the probability-severity of harm matrix. On that basis, the organisation identifies the required level of human involvement in the decision-making. For safety-critical systems, organisations should ensure that a person be allowed to assume control, with the AI providing sufficient information for that person to make meaningful decisions or to safely shut down the system where control is not available.

Illustration:

An online retail store wishes to use AI to fully automate the recommendation of food products to individuals based on their browsing behaviours and purchase history. The automation will meet the organisation's commercial objective of operational efficiency.

Severity-Probability Assessment: The definition of *harm* can be the impact of making product recommendations that do not address the perceived needs of the individuals. The *severity* of making the wrong product recommendations to individuals may be low since individuals ultimately decide whether to make the purchase. The *probability of harm* may be high or low depending on the efficiency and efficacy of the AI solution.

Degree of human intervention in decision-making process: Given the low severity of harm, the assessment points to an approach that requires no human intervention. Hence, human-out-of-the loop model is adopted.

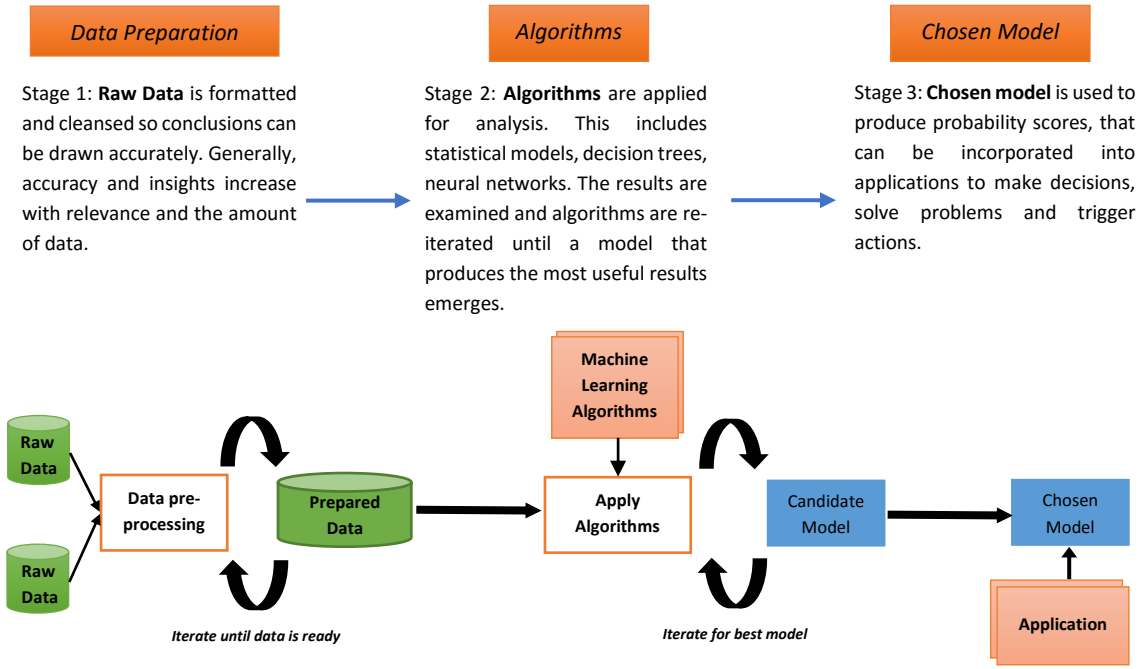


Regular review: The organisation can review this approach regularly to assess the *severity of harm* and as societal norms and values evolve. For example, the product recommendation solution may consistently promote sugary drinks to certain individuals. With heightened concerns about diabetes, the organisation should consider fine-tuning the models to reduce the promotion of sugary drinks.

Note: This is a simple illustration using bright-line norms and values. Organisations can consider testing this method of determining AI decision-making model against cases with more challenging and complex ethical dilemmas.

Operations Management

3.13 The Model Framework uses the following generalised AI adoption process[1] to describe phases in the deployment of an AI solution by an organisation. Organisations should be aware that the AI adoption process is not always uni-directional; it is a continuous process of learning.



[1] Adapted from Azure

3.14 During deployment, algorithms such as decision trees or neural networks are applied for analysis on training datasets. The resultant algorithmic models are examined and algorithms are iterated until a model that produces the most useful results for the use case emerges. This model and its results are then incorporated into applications to offer predictions, make decisions, and trigger actions. The intimate interaction between data and algorithm/model is the focus of this part of the Model Framework.

Data for Model Development

3.15 Datasets used for building models may come from multiple sources. The quality and selection of data are critical to the success of an AI solution. If a model is built using biased, inaccurate or non-representative data, the risks of unintended discriminatory decisions from the model will increase.

3.16 The persons who are involved in training and in selecting models for deployment may be internal staff or external service providers. The models deployed in an intelligent system should have an internal departmental owner, who will be the one making decisions on which models to deploy. To ensure the effectiveness of an AI solution, relevant departments within the organisation with responsibilities over quality of data, model training and model selection must work together to put in place **good data accountability practices**. These may include the following:

a. **Understanding the lineage of data**. This means knowing where the data originally came from, how it was collected, curated and moved within the organisation, and how its accuracy is maintained over time. Data lineage can be represented visually to trace how the data moves from its source to its destination, how the data gets transformed along the way, where it interacts with other data, and how the representations change. There are three types of data lineage:

i. Backward data lineage looks at the data from its end-use and backdating it to its source.

ii. Forward data lineage begins at the data's source and follows it through to its end-use.

iii. End-to-end data lineage combines the two and looks at the entire solution from both the data's source to its end-use and from its end-use to its source.

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. The Model Framework recognises that in some instances, the origin of data could be difficult to establish. One example could be datasets obtained from a trusted third-party who may have

commingled data from multiple sources. Organisations should assess the risks of using such data and manage them accordingly.

b. **Ensuring data quality**. This means understanding and addressing factors that may affect the quality of data, such as:

   i. The accuracy of the dataset, in terms of how well the values in the dataset match the true characteristics of the entities described by the dataset.

   ii. The completeness of the dataset, both in terms of attributes and items.

   iii. The veracity of the dataset, which refers to how credible the data is, including whether the data originated from a reliable source.

   iv. How recently the dataset was compiled or updated.

   v. The relevance of the dataset and the context for data collection, as it may affect the interpretation of and reliance on the data for the intended purpose.

   vi. The integrity of the dataset that has been joined from multiple datasets, which refers to how well extraction and transformation have been performed.

   vii. The usability of the dataset, including how well the dataset is structured in a machine-understandable form.

   viii. Human interventions, e.g. if any human has filtered, applied labels, or edited the data.

c. **Minimising inherent bias.** This Model Framework recognises that there are many types of bias relevant to AI. The Model Framework focuses on inherent bias in datasets, which may lead to undesired outcomes such as unintended discriminatory decisions. Organisations should be aware that the data which they provide to AI systems could be inherently biased and should take steps to mitigate such bias. The two common types of bias in data include:

   i. **Selection bias.** This bias occurs when the data used to produce the model are not fully representative of the actual data or environment that the model may receive or function in. Common examples of selection bias in datasets are *omission bias* and *stereotype bias*. Omission bias describes the omission of certain characteristics from the dataset, e.g. a dataset of Asian faces only will exhibit omission bias if it is used for facial recognition training for a population that includes non-Asians. A dataset of vehicle types within the central business district on a weekday may exhibit stereotype bias weighted in favour of cars, buses and motorcycles but under-represent bicycles if it is used to model the types of transportation available in Singapore.

ii. **Measurement bias.** This bias occurs when the data collection device causes the data to be systematically skewed in a particular direction. For example, the training data could be obtained using a camera with a colour filter that has been turned off, thereby skewing the machine learning result.

Identifying and addressing inherent bias in datasets is not easy. One way to mitigate the risk of inherent bias is to have a heterogeneous dataset, i.e. collecting data from a variety of reliable sources. Another way is to ensure the dataset is as complete as possible, both from the perspective of data attributes and data items. Premature removal of data attributes can make it difficult to identify and address inherent bias.

d. **Different datasets for training, testing, and validation.** Different datasets are required for training, testing, and validation. The model is trained using the training data, while the model's accuracy is determined using the test data. Where applicable, the model could also be checked for systematic bias by testing it on different demographic groups to observe whether any groups are being systematically advantaged or disadvantaged. Finally, the trained model can be validated using the validation dataset. It is considered good practice to split a large dataset into subsets for these purposes. However, where this is not possible if organisations are not working with large dataset AI models or are using pre-trained model as in the case of transfer learning, organisations should be cognisant of the risks of systematic bias and put in place appropriate safeguards.

e. **Periodic reviewing and updating of datasets.** Datasets (including training, testing, and validation datasets) should be reviewed periodically to ensure accuracy, quality, currency, relevance, and reliability. Where necessary, the datasets should be updated with new input data that is obtained from actual use of the AI models deployed in production. When such new input data is used, organisations need to be aware of potential bias as using new input data that has already gone through a model once could create a reinforcement bias.

<u>Algorithm and Model</u>

3.17 Organisations should consider measures to enhance the transparency of algorithms found in AI models through concepts of explainability, repeatability and traceability. An algorithm deployed in an AI solution is said to be **explainable** if how it functions and how it arrives at a particular prediction can be explained. The purpose of being able to explain predictions made by AI is to build understanding and trust. Organisations deploying AI solutions should also incorporate descriptions of the solutions' design and expected behaviour into their product or service description and system technical specifications documentation to demonstrate accountability to individuals and/or regulators. This could also include design decisions in relation to why certain features, attributes or models are selected in place of others. Where necessary, organisations

should request assistance from AI Solution Providers as they may be better placed to explain how the solutions function.

3.18 The Model Framework sets out that explainable AI can be achieved through explaining how deployed AI models' algorithms function and/or how the decision-making process incorporates model predictions. Organisations implementing the Model Framework may provide different levels of detail in their explanations depending on the technical sophistication of the intended recipient (e.g. individuals, other businesses or organisations, and regulators) and the type of AI solution that is used (e.g. statistical model).

3.19 Model training and selection are necessary for developing an intelligent system (system that contains AI technologies). Organisations using intelligent systems should document how the model training and selection processes are conducted, the reasons for which decisions are made, and measures taken to address identified risks. The field of "Auto-Machine Learning" aims to automate the iterative process of the search for the best model (as well as other meta-variables such as training procedures). Organisations using these types of tools should consider the transparency, explainability, and traceability of the higher-order algorithms, as well as the child-models selected. Algorithm audits can also be carried out in certain circumstances (See Annex A).

3.20 It should be noted that technical explainability may not always be enlightening, especially to the man in the street. Implicit explanations of how the AI models' algorithms function may be more useful than explicit descriptions of the models' logic. For example, providing an individual with counterfactuals (such as "you would have been approved if your average debt was 15% lower" or "these are users with similar profiles to yours that received a different decision") can be a powerful type of explanation that organisations could consider.

3.21 There could also be scenarios where it might not be practical or reasonable to provide information in relation to an algorithm. This is especially so in the contexts of proprietary information, intellectual property, anti-money laundering detection, information security, and fraud prevention where providing detailed information about or reviews of the algorithms or the decisions made by the algorithms may expose confidential business information and/or inadvertently allow bad actors to avoid detection.

3.22 Where explainability cannot be practicably achieved (e.g. black box) given the current state of technology, organisations can consider documenting the **repeatability** of results produced by the AI model. It should be noted that documentation of repeatability is not an equivalent alternative to explainability. Repeatability refers to the ability to consistently perform an action or make a decision, given the same scenario. The consistency in performance could provide AI users with a certain degree of confidence. Helpful practices include:

a. Conducting **repeatability assessments** for commercial deployments in live environments to ensure that deployments are repeatable.

b. Perform **counterfactual fairness testing**. A decision is fair towards an individual if it is the same in the actual world and a counterfactual world where the individual belonged to a different demographic group.

c. Assessing how **exceptions** can be identified and handled when decisions are not repeatable, e.g. when randomness has been introduced by design.

d. Ensuring **exception handling** is in line with organisations' policies.

e. Identifying and accounting for changes over time to ensure that models trained on time-sensitive data remain relevant.

3.23 An AI model is considered to be **traceable** if its decision-making processes are documented in an easily understandable way. Traceability is important for various reasons: the traceability record in the form of an audit log can be a source of input data that can in future be used as a training dataset; the information is also useful for troubleshooting, and in an investigation into how the model was functioning or why a particular prediction was made.

3.24 Practices that promote traceability include:

a. Building an **audit trail** to document the decision-making process.

b. Implementing a **black box recorder** that captures all input data streams. For example, a black box recorder in a self-driving car tracks the vehicle's position and records when and where the self-driving system takes control of the vehicle, suffers a technical problem or requests the driver to take over the control of the vehicle.

c. Ensuring that data relevant to traceability are **stored appropriately** to avoid degradation or alteration, and **retained for durations** relevant to the industry.

3.25 Organisations should establish an internal policy and process to perform **regular model tuning** to cater for changes to customer behaviour over time and to refresh models based on updated training datasets that incorporate new input data. Model tuning may also be necessary when commercial objectives, risks, or corporate values change.

3.26 Wherever possible, testing should reflect the dynamism of the planned production environment. To ensure safety, testing may need to assess the degree to which an AI solution generalises well and fails gracefully. For example, a warehouse robot tasked with avoiding obstacles to complete a task (e.g. picking packages) should be tested with different types of obstacles and realistically varied internal environments (e.g. workers wearing a variety of different coloured shirts). Otherwise, models risk learning regularities in the environment which do not reflect actual conditions (e.g. assuming that all humans that it must avoid will be wearing white lab coats). Once AI models are

deployed in the real-world environment, **active monitoring, review and tuning** are advisable.

Customer Relationship Management

3.27 Appropriate communication inspires trust as it builds and maintains open relationships between organisations and individuals (including employees). Organisations should incorporate the following factors to effectively implement and manage their communication strategies when deploying AI.

3.28 **General disclosure**. Organisations should provide general information on whether AI is used in their products and/or services. Where appropriate, this could include information on how AI is used in decision-making about individuals, and the role and extent that AI plays in the decision-making process. For example, the manufacturer of a GPS navigation system may inform its users that AI is used to automatically generate possible routes from point A to point B. However, the user of the navigation system makes the decision on which route to take. An online portal may inform its users that the chatbot they are interacting with is AI-powered.

3.29 **Increased transparency** contributes to building greater confidence in and acceptance of AI by increasing the openness in customer relationships. To do so, organisations can consider disclosing the manner in which an AI decision may affect the individuals, and if the decision is reversible. For example, an organisation may inform the individuals of how their credit ratings may lead to refusal of loan not only from this organisation but also from other similar organisations; but such a decision is reversible if individuals can provide more evidence on their credit worthiness.

3.30 Organisations should use easy-to-understand language in their communications to increase transparency. There are existing tools to measure readability, such as the Fry readability graph, the Gunning Fog Index, the Flesh-Kincaid readability tests, etc. Decisions with higher impact should be communicated in an easy-to-understand manner, with the need to be transparent about the technology being used.

3.31 As ethical standards governing the use and building of AI evolve, organisations could also carry out their **ethical evaluations** and make meaningful summaries of these evaluations available.

3.32 **Policy for explanation.** Organisations should develop a policy on what explanations to provide to individuals. These can include explanations on how AI works in a decision-making process, how a specific decision was made and the reasons behind that decision, and the impact and consequence of the decision. The explanation can be provided as part of general communication. It can also be information in respect of a specific decision upon request.

3.33 **Human-AI interface**. Organisations should test user interfaces and address usability problems before deployment, so that the user interface serves its intended purposes.

Individuals' expectations can also be managed by informing them that they are interacting with a chatbot rather than a human being. If applicable, organisations should also inform individuals that their replies would be used to train the AI system. Organisations should be aware of the risks of using such replies as some individuals may intentionally use "bad language" or "random replies" which would affect the training of the AI system.

3.34 **Option to opt-out.** Organisations should consider carefully when deciding whether to provide individuals the option to opt-out and whether this option should be offered by default or only upon request. The considerations should include:

a. Degree of risk/harm to the individuals.

b. Reversibility of harm to the individual should risk actualise.

c. Availability of alternative decision-making mechanisms.

d. Cost or trade-offs of alternative mechanisms.

e. Complexity and inefficiency of maintaining parallel systems.

f. Technical feasibility.

3.35 Where an organisation has weighed the factors above and decided not to provide an option to opt-out, it should then consider other modes of providing recourse to the individual such as providing a channel for reviewing the decision. Where appropriate, organisations should also keep a history of chatbot conversation when facing complaints or seeking recourse from consumers.

3.36 Organisations should put in place the following communications channels for their customers:

a. **Feedback channel.** This channel could be used for individuals to raise feedback or raise queries. It could be managed by an organisation's Data Protection Officer ("DPO") if this is appropriate. Where individuals find inaccuracies in their personal data which has been used for decisions affecting them, this channel can also allow them to correct their data. Such correction and feedback, in turn, maintain data veracity. It could also be managed by an organisation's Quality Service Manager (QSM) if individuals wish to raise feedback and queries on material inferences made about them.

b. **Decision review channel**. Apart from existing review obligations, organisations can consider providing an avenue for individuals to request a review of material AI decisions that have affected them. Where a decision is fully automated, it is reasonable to provide an individual review by a human agent upon request, if the

impact of the decision on the individual is material. However, should it be partially automated with review prior to confirming the decision, the decision has already been reviewed by a human agent. In the latter scenario, this would be no different than a non-AI decision.

Conclusion

3.37 This Model AI Governance Framework is by no means complete or exhaustive and remains a document open to feedback. As AI technologies evolve, so would the related ethical and governance issues. It is PDPC's aim to update this Framework periodically with the feedback received, to ensure that it remains relevant and useful to organisations deploying AI solutions.

ANNEX A

Algorithm Audits

4.1 Algorithm audits are conducted if it is necessary to discover the actual operations of algorithms comprised in models. This would have to be carried out at the request of a regulator having jurisdiction over the organisation or by an AI technology provider to assist its customer organisation which has to respond to a regulator's request. Conducting an algorithm audit requires technical expertise which may require engaging external experts. The audit report may be beyond the understanding of most individuals and organisations. The expense and time required to conduct an algorithm audit should be weighed against the expected benefits obtained from the audit report.

4.2 Organisations can consider the following factors when considering whether to conduct an algorithm audit:

a. The **purpose** for conducting an algorithm audit. The Model Framework promotes the provision of information about how AI models function as part of explainable AI. Before embarking on an algorithm audit, it is advisable to consider whether the information that has already been made available to individuals, other organisations or businesses, and regulators is sufficient and credible (e.g. product or service descriptions, system technical specifications, model training and selection records, data provenance record, audit trail).

b. Target **audience** of audit results. This refers to the **expertise** required of the target audience to effectively understand the data, algorithm and/or models. The information required by different audience varies. When the audience is **individuals**, providing information on the decision-making process and/or how the individuals' data is used in such process will achieve the objective of explainable AI more efficaciously. When the audience is **regulators**, information relating to data accountability and the functioning of algorithms should be examined first. An algorithm audit can prove how an AI model operates if there is reason to doubt the veracity or completeness of information about its operations.

c. General **data accountability**. Organisations can provide information on how general data accountability is achieved within the organisations. This includes all the good data practices described in the Model Framework under Data for Model Development section such as maintaining data lineage through keeping a data provenance record, ensuring data accuracy, minimising inherent bias in data, splitting data for different purposes, determining data veracity and reviewing and updating data regularly.

d. Algorithms in AI models can be **commercially valuable information** that can affect market competitiveness. If a technical audit is contemplated, corresponding mitigation measures should also be considered.

ANNEX B

Glossary

5.1 This glossary comprises a collection of foundational AI ethical principles, distilled from various sources.[2] Not all are included or addressed in the Model Framework. Organisations may consider to incorporate these principles into their own corporate principles, where relevant and desired.

5.2 On Accuracy:

a. Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst case implications can be understood and can inform mitigation procedures.

5.3 On Explainability:

a. Ensure that automated and algorithmic decisions and any associated data driving those decisions can be explained to end-users and other stakeholders in non-technical terms.

5.4 On Fairness:

a. Ensure that algorithmic decisions do not create discriminatory or unjust impacts across different demographic lines (e.g. race, sex, etc.).

b. To develop and include monitoring and accounting mechanisms to avoid unintentional discrimination when implementing decision-making systems.

c. To consult a diversity of voices and demographics when developing systems, applications and algorithms.

5.5 On Human Centricity and Well-Being:

a. To aim for an equitable distribution of the benefits of data practices and avoid data practices that disproportionately disadvantage vulnerable groups.

b. To aim to create the greatest possible benefit from the use of data and advanced modelling techniques.

[2] These include Institute of Electrical and Electronics Engineers (IEEE) Standards Association's *Ethically Aligned Design* (https://standards.ieee.org/industry-connections/ec/ead-v1.html), Software and Information Industry Association's *Ethical Principles for Artificial Intelligence and Data Analytics* (https://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2017-11-06-160346-990) and Fairness, Accountability and Transparency in Machine Learning's *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms* (http://www.fatml.org/resources/principles-for-accountable-algorithms). They also include feedback from the industry in previous rounds of consultation.

c. Engage in data practices that encourage the practice of virtues that contribute to human flourishing, human dignity and human autonomy.

d. To give weight to the considered judgments of people or communities affected by data practices and to be aligned with the values and ethical principles of the people or communities affected.

e. To make decisions that should cause no foreseeable harm to the individual, or should at least minimise such harm (in necessary circumstances, when weighed against the greater good).

f. To allow users to maintain control over the data being used, the context such data is being used in and the ability to modify that use and context.

5.6 On Responsibility, Accountability and Transparency:

a. Build trust by ensuring that designers and operators are responsible and accountable for their systems, applications and algorithms, and to ensure that such systems, applications and algorithms operate in a transparent and fair manner.

b. To make available externally visible and impartial avenues of redress for adverse individual or societal effects of an algorithmic decision system, and to designate a role to a person or office who is responsible for the timely remedy of such issues.

c. Incorporate downstream measures and processes for users or consumers to verify how and when AI technology is being applied.

d. To keep detailed records of design processes and decision-making.

5.7 On Human Rights

a. Ensure that the design, development and implementation of technologies do not infringe on internationally recognised human rights.

5.8 On being Sustainable

a. Favour implementations that effectively predict future behaviour and generate beneficial insights over a reasonable period of time.

5.9 On being Progressive

a. Favour implementations where the value created is materially better than not engaging in that project.

5.10 On Auditability

a. Enable interested third parties to probe, understand, and review the behaviour of the algorithm through disclosure of information that enables monitoring, checking, or criticism.

5.11 On Robustness and Security

a. AI systems should be safe and secure, not vulnerable to tampering or compromising the data they are trained on.

5.12 On Inclusivity

a. Ensure that AI is accessible to all.

ANNEX C

Use Case in Healthcare – UCARE.AI

UCARE.AI (https://www.ucare.ai) is an artificial intelligence and machine learning company on a scientific mission to solve healthcare problems and advance humankind through the ethical and responsible use of data. UCARE.AI deploys a suite of AI and machine learning algorithms, including proprietary deep learning and neural network algorithms, built on a cloud-based microservices architecture to provide sustainable and customisable healthcare solutions for doctors, hospitals, patients, insurers and pharmaceutical companies.

A successful use case is the recent implementation of AI-Powered Pre-Admission Cost of Hospitalization Estimation (APACHE^{TM}) for four major hospitals, namely Mount Elizabeth, Mount Elizabeth Novena, Gleneagles and Parkway East hospitals; owned by Parkway Pantai. This study shares UCARE.AI's methodology for developing and deploying APACHE, a scalable plug-and-play system that provides high availability, fault-tolerance, and real-time processing of high-volume estimate requests. APACHE provides more accurate estimates, with a four-fold improvement in accuracy over Parkway Pantai's previous bill estimation system. This is done with the intent of achieving standardisation of healthcare cost estimation and provision of greater price transparency to facilitate the building and maintenance of trust between payers, providers, and patients. This is in line with UCARE.AI's commitment to ensure patients continue to make well-informed decisions on available medical treatment options.

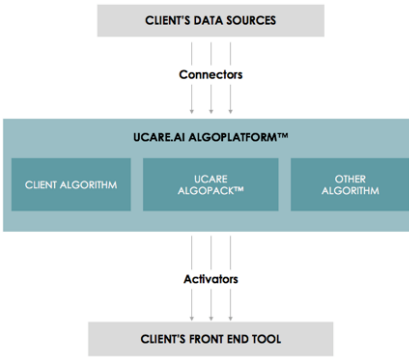
Background

Previous healthcare cost estimation methods involve traditional techniques such as (i) normal distribution-based techniques, (ii) parametric models based on skewed distributions, (iii) mixture models, (iv) survival analysis, etc. The existing approach used was via simple statistical aggregations based on the Table of Surgical Procedures quoted prices or ICD-10 diagnostic codes.

Challenges include relatively high error rates, high financial and human cost of updates, and low frequency of updates due to these high costs.

UCARE.AI worked with Parkway to resolve these issues with a multi-step process involving: (i) data exploration, (ii) data cleaning, (iii) feasibility assessment (iv) feature engineering, (v) machine learning, and (vi) presentation of results. With satisfactory results from the proof of concept, APACHE was then put into production.

High-Level Architecture of APACHE API



1. *Data Sources.* Relevant data is obtained from partner organisations for use. As the system is further improved upon, publicly available data sources as well as third-party data are used to generate predictions, thereby reducing the need for personal data collection.
2. *Connectors.* Basic data validation is conducted prior to being ingested into the data production warehouse.
3. *AlgoPlatform.* The data is processed by the algorithms, and encrypted for storage. The algorithms are integrated with reporting and monitoring systems for performance management and intervention to minimise downtime. Various machine learning models can be deployed to allow for model comparisons and can be hot-swapped in a live production environment.
4. *Activators.* These serve to assist with data authentication and verification, to send results to the client's chosen front end tool.

Aligning with PDPC's Model AI Governance Framework

UCARE.AI adopts a proactive approach that aligns with PDPC's Model AI Governance Framework.

Trustworthy and Verifiable

The proposed AI governing framework acknowledges that neural networks are inscrutable and verification of the results provided by such networks is required prior to putting them to use in human applications. UCARE.AI circumvents this problem by continuously validating the accuracy of its algorithms against the ground truth. Weekly check-ins with participating partners and domain experts are also employed to ensure quicker and more reliable iterations. Automated re-training of the data models ensure that the algorithms remain up-to-date. This methodology of continuous validation of its AI models with the help of experts from Parkway Pantai will help to boost confidence in the accuracy of its predictive insights and will help train algorithms to become even more precise with each amount of data inputted.

Accountability and Transparency

Prior to data collection, informed consent from stakeholders would have been obtained and approval of the use of data sought via open communication channels. The careful curating and conversion of data into usable format prior to building the models ensures the AI algorithm is kept accountable and coherent to users; this is done in conjunction with Parkway Pantai. The proper storage and repair of previously broken or missing data also serve to provide greater transparency and safety to users by minimising the influence of data gaps in the projection of the result. Careful monitoring of data is key in ensuring service reliability, and therefore detailed and consistent logging across the multiple components involved is also employed in APACHE, collected in a secure, centralised log storage that is made easily accessible to the development and operations team when required, allowing for prompt debugging and uptime tracking if necessary.

Fairness

The automated prediction of hospitalisation costs reduces the likelihood of human biases affecting the ultimate judgement of the data and provides an element of consistency across all predictions. Discrimination based on income levels and insurance coverage, for instance, would be effectively negated. Although there would be concerns about the use of a 'human-out-of-the-loop' system, the algorithm in question is designed to be human-centric.

Human-Centric

This use case highlights how artificial intelligence may be used in augmenting decision-making capabilities in a human-centric manner whilst minimising the potential risks of harm to involved parties. The automated process of bill estimation negates the need for tedious statistical calculations, thereby freeing up man-hours and effort to allow for the channeling of these into more creative pursuits. Furthermore, the information provided would serve to benefit patients and payers by allowing for more accurate cost forecasting, efficient allocation and distribution of healthcare resources, and guidance on new policy initiatives. Patients would be conferred greater peace of mind over their healthcare expenditure such that they may focus their energies on recovery instead.

To minimise the risk of harm, rigorous feasibility studies are conducted prior to using the data to focus on creating a valid and robust validation framework. This will be done in conjunction with partners and their feedback on the proposed framework obtained before proceeding. A human feedback loop with inputs from the client organisation (Parkway Pantai-owned hospitals) is also in-built into each algorithm to enhance sophistication, while a manual override protocol is also included to ensure that these algorithms can be safely terminated if deemed necessary. This ensures that the algorithm remains under human control and in line with the medical field's well-established ethical principles of beneficence, non-maleficence, and social justice.

For more information, please visit https://www.ucare.ai or contact hello@ucare.ai.

ACKNOWLEDGEMENTS

The Personal Data Protection Commission, Infocomm Media Development Authority, expresses its sincere appreciation to the following for their valuable feedback to this Model AI Governance Framework (in alphabetical order):

AIG Asia Pacific Insurance Pte. Ltd.
Asia Cloud Computing Association
AsiaDPO
BSA | The Software Alliance
DBS
Element AI
Facebook
Fullerton Systems and Services
Grab
IBM Asia Pacific
MasterCard
MSD
Microsoft Asia
OCBC Bank
Salesforce
Standard Chartered Bank
Telenor Group
Temasek International
Ucare.AI

END OF DOCUMENT