WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

Companion to the Model AI Governance Framework

– Implementation and Self-Assessment Guide for Organizations

Prepared in collaboration with the
Info-communications Media Development Authority of Singapore

January 2020

SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

pdpc PERSONAL DATA PROTECTION COMMISSION SINGAPORE

World Economic Forum
91-93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland
Tel.: +41 (0)22 869 1212
Fax: +41 (0)22 786 2744
Email: contact@weforum.org
www.weforum.org

Contents

Foreword

Tan Kiat How
Chief Executive,
Infocomm Media Development Authority
Commissioner, Personal Data Protection
Commission Singapore

At the brink of entering into a new decade, Artificial Intelligence (AI) seems to be shrouded under a cloud of ambivalence. On one hand, we expect advances in AI to bring significant benefits to businesses and citizens. On the other hand, there is increasing anxiety over the impact of AI on the workplace and in our societies, particularly around the implications for ethics and accountability.

There are no easy answers. Getting the balance right will be a crucial challenge that our generation will need to tackle and one that requires the collaboration and support of multiple stakeholders - enterprises, governments, civil society organizations and academics. Importantly, we believe that it is about adopting a pragmatic approach, putting principles into operations and taking concrete steps to build trust in AI deployments.

In January 2019, Singapore released Asia's first Model AI Governance Framework (Model Framework). The Model Framework translates AI ethical principles into practical measures for organizations to adopt voluntarily, such as in internal governance structures, decision-making models and operations management practices.

This year, we have taken another step forward by publishing the Implementation and Self-Assessment Guide for Organizations (ISAGO). This Guide provides a set of questions and practical examples to enable organizations to assess the alignment of their AI governance practices with the Model Framework. Professionals who are proficient in AI governance could use ISAGO to help organizations implement the Model Framework or assess an organisation's implementation.

We are pleased to have partnered with the World Economic Forum Centre for the Fourth Industrial Revolution to co-develop ISAGO, in close consultation with the industry. We are also grateful for the support and contributions by over 60 local and international organizations to ISAGO.

We are still very much at the beginning of this long journey, with many more questions than answers. But with the willingness to ask the right questions, work collaboratively with multiple stakeholders and take a pragmatic approach to problem-solving, we believe that it is possible to nurture a safe and trusted environment for AI innovation. The Model Framework and Guide are Singapore's contribution to this important global discussion, and we welcome views and fellow travellers.

Foreword

Murat Sonmez
Managing Director, the World Economic Forum Centre for the Fourth Industrial Revolution Network

The Fourth Industrial Revolution's advancements in Artificial Intelligence (AI) have spurred the global economy, starting a conversation on the role technology plays in our society. Companies and governments alike have embraced innovation as a way to help create more inclusive and responsible communities.

However, AI and technologies in the Fourth Industrial Revolution have created unique challenges that require new frameworks and pragmatic solutions to ensure an equitable, ethical and fair future for our society. Maximizing the benefits of new technologies while mitigating the unintended consequences will safeguard the positive impact possible with these technologies. In the Fourth Industrial Revolution, we see first movers capturing the advancements of AI, but it will be paramount that these technologies are used responsibly.

The first edition of the Model AI Governance Framework built the principles of what responsible AI would look like and allowed Singapore to contribute to the global discussion on the ethics of AI. Over the past year, the World Economic Forum's Centre for the Fourth Industrial Revolution and Singapore's Personal Data Protection Commission have co-developed the Implementation and Self-Assessment Guide for Organizations to complement the Model AI Governance Framework. The Implementation and Self-Assessment Guide aims to help organizations assess their AI governance processes. In doing so, organizations can identify potential gaps in their AI governance processes and address them accordingly. The Guide also provides examples of how organizations could implement the considerations and practices set out in the Model AI Governance Framework.

The project will be released in tandem with work completed by the Singapore Government to expound on the resources for companies to apply the responsible use of AI. First, a second edition of the Model AI Governance Framework will be released with new considerations brought about by advancement in the field and includes illustrations from companies on how to apply these practices. Second, the release of a Compendium of Use Cases will outline use case examples showing organizations how companies have operationalized the principles from the Model AI Governance Framework.

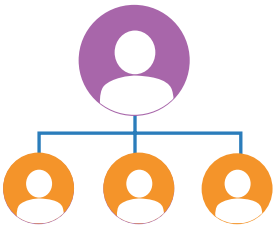
Inclusive and accountable policies like the Model AI Governance Framework will be vital to addressing these new challenges brought about by the Fourth Industrial Revolution.

Introduction

In collaboration with the World Economic Forum Centre for the Fourth Industrial Revolution, the Info-communications Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC) have developed this Implementation and Self-Assessment Guide for Organizations (ISAGO), a companion to complement the voluntary Model AI Governance Framework[1] (Model Framework). This Guide is meant to be a living document and aims to help organizations assess the alignment of their AI governance processes with the Model Framework, identify potential gaps in their existing processes and address them accordingly.

The Model Framework is published by the PDPC and provides guidance to private sector organizations deploying AI at scale on how to do so in a responsible manner. The Model Framework translates ethical principles into implementable practices, applicable to a common AI deployment process. It covers four key areas:

A
Internal governance structures and measures



C
Operations management

B
Determining the level of human involvement in AI-augmented decision-making

D
Stakeholder interaction and communication

Who should use this Guide?

The Model Framework and this Guide are intended to guide organizations that procure and deploy AI solutions and use them to offer products and/or services to their customers or consumers. For example, a retail company may use the Guide when deploying AI to make product recommendations to consumers based on their profiles. An insurance company may use the Guide when deploying AI to determine the premium and approve the application for an insurance product.

The Model Framework and this Guide can also be used by organizations using AI to improve their operational efficiency or to collaborate with other organizations. For example, an organization may use AI to identify anomalies in its transactions and flag them for the relevant department's attention. This Guide is not intended for organizations that are deploying updated commercial off-the-shelf software packages that happen to incorporate AI in their feature set.

How should this Guide be used?

The Guide sets out a list of questions, based on and organized according to the four key areas described in the Model Framework, for organizations to consider in a systematic manner. Hence, **this Guide should be read in conjunction with the Model Framework**. Organizations should refer to the Model Framework for definitions of terms and explanations of concepts used in this Guide.

The Guide also provides references and examples on how organizations could implement the considerations and practices set out in the Model Framework. These references and examples include publications by the PDPC (e.g. advisory guidelines and guides), and industry use cases and practices that have been shared with the PDPC. We have also included a list of international AI standards that are being developed (Annex). Organizations are free to implement other measures that best fit the purpose and context of their AI deployment, as appropriate.

Organizations should not attempt to implement all the practices and considerations in this Guide because **not all practices and considerations may be applicable in their context**. Additionally, a number of considerations may only be applicable in specific scenarios. These have been marked with the label "**Relevant only in limited scenarios**". This would help organizations prioritize their implementation of AI governance measures.

When using the Guide, organizations should consider whether the questions and practices are relevant to their unique business context and industry. Organizations would also need to consider their business needs, resource constraints, regulatory requirements and specific use cases. Generally, an organization should consider adopting a risk-based approach to AI governance that is commensurate with the potential harm of the AI solution deployed. The scope of the questions in the Guide may overlap and could reinforce concepts that are important in ensuring responsible deployment of AI. Last but not least, organizations are encouraged to document the development of their governance process as a matter of good practice.

Section 1: Objectives of deploying AI

To guide organizations on how to include ethical considerations in developing their business case to deploy AI

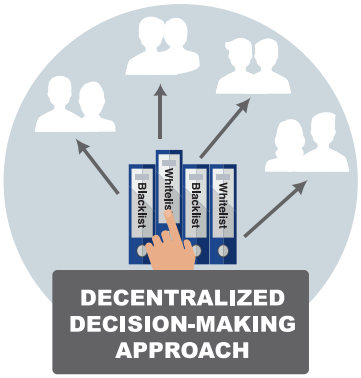
Guiding questions		Useful industry examples, practices and guides for consideration
Considerations prior to deployment of AI:		
1.1	Has your organization defined a clear purpose in using the identified AI solution (e.g. operational efficiency and cost reduction)?	– Consider whether AI is able to address the identified problem or issue
1.2	Has your organization considered conducting an assessment on whether the expected benefits of implementing the identified AI solution in a responsible manner (as described in the Model Framework) outweighs the expected costs?	– Consider whether to conduct a cost-benefit analysis – Consider whether it is useful to leverage benchmarks and case studies for similar AI solutions (e.g. PDPC's Compendium of Use Cases[2]), and adopt the AI governance practices for your organization's identified AI solution, where applicable. These could be case studies applied in other geographies, industries or domains, but with similarities to your organization's use case

1.3	Did your organization consider whether the decision to use AI for a specific application/use case is consistent with its core values and/or societal expectations?	– Consider developing a set of ethical principles that is in line with or can be incorporated into the organization's mission statement. In addition, it would be useful to outline how to adopt (e.g. contextualise) them in practice
		– Consider developing a Code of Ethics for the use of AI. Relevant areas to consider include:
		  – Regulatory risks (e.g. compliance with Singapore's Personal Data Protection Act 2012 (PDPA) and sectoral regulations)
		  – Public relations risks (e.g. public perception towards the organization's AI practices)
		  – Costs (e.g. impact of incorporating governance practices into the organization's current legacy business models and organizational structure)
		  – Resources and internal champions to drive responsible implementation of AI

Section 2: Internal governance structures and measures

To guide organizations to develop appropriate internal governance structures

Guiding questions	Useful industry examples, practices and guides for consideration
2.1 Does your organization have an existing governance structure that can be leveraged to oversee the organization's use of AI? **2.2** If your organization does not have an existing structure to tap on, has your organization put in place a governance structure to oversee the organization's use of AI?	– Consider whether it is useful to adapt existing governance, risk and compliance (GRC) structures to incorporate AI governance processes To provide oversight on the use of data and AI within an organization: – Consider a sandbox type of governance to test-bed and deploy AI solutions, before fully-fledged governance structures are put in place – Consider whether it is necessary to establish a committee comprising representatives from relevant departments (e.g. legal/compliance, technical and sales and communication) to oversee AI governance in the organization with proper terms of reference (e.g. refine organization's AI governance frameworks to ensure they meet the organization's commercial, legal, ethical and reputational requirements)  – Consider whether to implement a process where each department head develops and is accountable for the controls and policies that pertain to the respective areas, overseen by relevant subject matter experts such as chief security officer and data protection officer – Consider whether it is necessary to establish checks and balances: – An internal team consisting of relevant departments to oversee methodology, algorithms and deployment of AI – A separate team to conduct validation If there are strong concerns about how AI is being used for the project, neither of the teams will be able to one-sidedly terminate the project, but they can conduct further testing and validation. – Consider whether AI governance processes will ensure that the deployment of AI solutions complies with existing laws and regulations (e.g. trials for autonomous vehicles should comply with Road Traffic (Autonomous Motor Vehicles) Rules 2017; the deployment of AI solutions should comply with Singapore's Competition Act and should not result in collusive outcomes) – Consider developing a handbook that outlines the entire governance process for AI deployment and make the handbook available to all staff



In implementing the governance structure, organizations may consider determining appropriate features such as:

- Adopting a centralized or decentralized decision-making approach based on certain guidelines:
  - Take a centralized decision-making approach. For the deployment of an AI solution that is not determined to be low risk or could be potentially contentious, respective departments to bring the issue to the senior management or the AI ethics committee
  - Take a decentralized decision-making approach. Respective departments can make the decision on whether to deploy the AI solution based on a predetermined whitelist and/or blacklist. Considerations that could be included in a blacklist are AI applications that would likely cause overall harm and direct injury. Having clear policies that describe off-limits practices (i.e. blacklist) would be useful for organizations that adopt decentralized models where tracking AI is more challenging
- Exploring an approach that takes less time to review/recalibrate AI models to be more resource-efficient, if the model validation processes take a long period of time
- Conducting regular reviews of their governance processes and structures

---

2.3 Did your organization's board and/or senior management sponsor, support and participate in your organization's AI governance?

- Consider whether it is useful to form a committee/board that is chaired by the senior management and include senior leaders from the various teams (e.g. chief data officer, chief privacy officer and chief information security officer). Including key decision-makers is critical for efficiency and the credibility of the committee/board
- Consider having top management set clear expectations/directions for AI governance within the organization

Clear roles and responsibilities for the ethical deployment of AI

2.4 Are the responsibilities of the personnel involved in the various AI governance processes clearly defined?

- Consider whether it is useful or practical for the board and senior management to champion responsible AI deployment and ensure that all employees are committed to implementing the practices:
  - **Strategic level:** Board to be responsible for risk and corporate values, and C-suites translate them into strategies. Committee comprising senior management to approve the AI models
  - **Implementation level:** While there is oversight from the senior management, individual project team leads and officers should be held accountable for the AI projects. The roles, responsibilities for managing model risks and ensuring regulatory compliance should be clearly established and documented. For organizations that have the resources or sophistication to have a dedicated regulatory/compliance team, this team could check for relevant existing legal restrictions or compliance requirements for the deployment. At the same time, the regulatory team engages their clients to receive feedback on ethical issues as they implement AI-enabled solutions/services
- Consider defining separate responsibilities for business and technical staff:
  - Business staff responsible for defining business goals and business rules, and checking that an AI system behaves consistently with those goals and rules
  - Technical staff responsible for data practices, security, stability, error handling
- Consider conducting a review of job descriptions periodically for roles that involve AI deployment

BUSINESS STAFF

TECHNICAL STAFF

2.5	Are the personnel involved in various AI governance processes: A. Fully aware of their roles and responsibilities? B. Properly trained? C. Equipped with the necessary resources and guidance to perform their duties?	– Consider the importance and relevance of hiring talent with the right skillsets – Having a multi-disciplinary team to provide a broader lens on the impact of AI deployment on the organization and individuals – Creating a new and specialized role (e.g. data scientist) with specific responsibilities to examine ethical and data protection issues in the AI deployment process – Consider educating key internal stakeholders to increase awareness of the implications of AI development/deployment as well as the need for guidelines (e.g. AI engineering guidelines)
2.6	Are the relevant staff dealing with AI systems properly trained to interpret AI model output and decisions as well as to detect and manage bias in data?	– Consider whether it is useful to conduct general training for personnel involved in various AI governance processes. For staff dealing with AI systems, consider whether it is necessary to conduct specialized training – Considering developing or partnering with an education institution to create a suite of online learning modules to support AI skill development for employees
2.7	Are the other staff who interact with the AI system aware of and sensitive to the relevant risks when using AI? Do they know who to raise such issues to when they spot them (e.g. subject-matter experts within their organizations)?	– Consider educating employees at all levels, particularly those using the AI system or with customer-facing roles, to identify and report potential ethical concerns relating to AI development and deployment

Risk management and internal controls

2.8	Does your organization have an existing risk management system that can be expanded to include AI-related risks?	– Consider implementing an internal policy explanation process to retain details of how decision-making on the deployment of AI was undertaken – Consider implementing a knowledge management registry to archive relevant documents to ensure proper knowledge transfer
2.9	Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g. personnel risk or changes to commercial objectives)?	

Section 3: Determining the level of human involvement in AI-augmented decision-making

To help organizations determine the appropriate extent of human oversight in their AI-augmented decision-making process

Guiding questions	Useful industry examples, practices and guides for consideration
3.1 Did your organization conduct an impact assessment (e.g. probability and/or severity of harm) on individuals and organizations who are affected by the AI solution?	– Consider whether it is necessary to list all internal and external stakeholders, and the impact on them accordingly – Consider whether it is necessary to assess risks from a technical perspective (e.g. system integrity tests) and from a personal data protection perspective (e.g. the PDPC's Guide to Data Protection Impact Assessments[3]) – Consider assessing risk at a societal/end-user level by conducting customer/society group testing – Consider whether it is necessary for AI-augmented decision-making to reflect prevailing societal norms and values

3.2 Based on the assessment, did your organization implement the appropriate level of human involvement in AI-augmented decision-making?	– Consider a **human-in-the-loop** approach when human judgement is able to significantly improve the quality of the decision made (e.g. pricing recommendation of million-dollar commodity bids) or when a human subjective judgment is required (e.g. market share forecasting for long-term decisions)
	– Consider a **human-out-of-the-loop** approach when it is not practical to subject every algorithmic recommendation to a human review. For example, when an AI model makes thousands or millions of micro-decisions (e.g. spare parts forecast for an airline company and daily replenishment recommendations in a retail environment). For such an approach, it would be important to ensure that the AI system is being developed and deployed in a manner that could provide simple and understandable explanations to individuals on the AI-augmented decision-making
	– Consider a **human-over-the-loop** approach to allow humans to intervene when the situation calls for it. To achieve this, organizations could consider using statistical confidence levels to determine when human is required to intervene (e.g. below a certain threshold, staff could be required to review a particular result generated by the AI model)
	– Organizations could also consider the following factors in determining the level of human involvement:
	  – Risk appetite. For example, organizations could have varying risk appetite in interrupting a transaction made by a retail customer as compared to a transaction made by a corporate customer that could result in more serious consequences (e.g. stopping a payroll)
	  – User experience of its clients' customers. For example, organizations might consider favouring a better user experience journey and reduce the level of human intervention
	  – Operational cost. For example, it might be costly for organizations to have a human to manually review all transactions, especially if there is a high volume of transactions and real-time decision-making is required

3.3 After deployment, did your organization continually identify, review and mitigate risks of using the identified AI solution?

- Consider whether it is useful to determine and implement an appropriate regular review period for retraining the AI model. For example, where image patterns are likely to change slowly (e.g. recognizing cats), to review and retrain the AI model less frequently. For patterns that are likely to change faster (e.g. phishing detection), consider a higher frequency of review and retraining
- Consider whether it is necessary to regularly review the AI model to assess the severity of harm to take into account evolving societal norms and values
- Consider defining key performance indicators for AI model's performance and alerting relevant staff when AI performance deteriorates
- Consider tracking the characteristics of the data that the AI is using, versus the data the AI was trained on, and alerting relevant staff when the data drifts too much (e.g. new categories appear, new values outside historical values appear, or the distribution of the values changes)
- Consider developing scenario-based response plans in the event that the risk management efforts fail

Relevant only in limited scenarios:

3.4 For safety-critical systems, did your organization ensure that:

A. The relevant personnel will be able to assume control where necessary?

B. The AI solution provides sufficient information to assist the personnel to make an informed decision and take actions accordingly?

- Consider whether it is necessary and feasible to put in place controls to allow the graceful shutdown of an AI system and/or bring it back to a safe state, in the event of a system failure
- When an AI model is making a decision for which it is significantly unsure of the answer/prediction, considering designing the AI model to be able to flag these cases and triage them for a human to review. This may occur when the data contains values that are outside the range of the training data, or for data regions where there were insufficient training examples to make a robust estimate

Section 4: Operations management

To help organizations adopt responsible measures in the operations aspect of their AI adoption process

Guiding questions	Useful industry examples, practices and guides for consideration
Data for Model Development – Ensuring personal data protection	
4.1 Did your organization implement accountability-based practices in data management and protection (e.g. the PDPA and OECD Privacy Principles)?	– Consider adopting industry best practices and engineering standards to ensure compliance with relevant data protection laws, such as the PDPA. It is important for organizations to implement proper personal data-handling practices, such as having policies for data storage, deletion and processing, particularly when the data deals with personal identifiable information – Consider whether model can be trained on pseudonymized or de-identified data[4] – Consider which data an AI system should have access to, and which sensitive data it should not have access to – Consider referring to (1) the PDPC's Advisory Guidelines on Key Concepts in the PDPA; (2) Guide to Accountability; and (3) Guide to Data Protection Impact Assessments – Consider applying applying for the PDPC's Data Protection Trustmark and Asia Pacific Economic Cooperation Cross Border Privacy Rules and Privacy Recognition for Processors (APEC CBPR & PRP) Systems certifications – Consider whether it is useful to implement a data governance panel/ dashboard to help with GRC on data protection

Data for Model Development – Understanding the lineage of data

4.2	Did your organization implement measures to trace the lineage of data (i.e. backward data lineage, forward data lineage and end-to-end data lineage)?	– Consider developing and maintaining a data provenance record – Consider whether it is useful to create a data inventory, data dictionaries, data change processes and document control mechanisms – Consider whether data can be traced back to the source at each stage – Consider whether it is useful to track data lineage by putting in place "feature repositories" with application programming interfaces (APIs), databases and files – Consider whether it is necessary to mandate developers to document data narratives/data diaries for accountability, as well as provide clear explanations of what data is used, how it is collected and why – Consider whether it is useful to establish a data policy team to manage tracking of data lineage with proper controls

Relevant only in limited scenarios:

4.3	If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?	– Consider obtaining datasets only from trusted third-party sources that are certified with proper data protection practices – Consider adopting the practices within IMDA's Trusted Data Sharing Framework[5] when establishing data partnerships (e.g. create a common "data-sharing language")

Data for Model Development – Ensuring data quality

4.4	Is your organization able to verify the accuracy of the dataset in terms of how well the values in the dataset match the true characteristics of the entity described by the dataset?	– Consider reviewing data in detail against its metadata – Consider whether it is useful to develop a taxonomy of data annotation to standardize the process of data labelling
4.5	Is the dataset used complete in terms of attributes and items?	– Consider whether it is useful to conduct validation schema checks (i.e. testing whether the data schema accurately represents the data from the source to ensure there are no errors in formatting and content)
4.6	Is the dataset used credible and from a reliable source?	– Consider whether it is necessary to put in place processes to identify possible errors and inconsistencies at the exploratory data analysis stage, before training the dataset
4.7	Is the dataset used up-to-date?	– Consider whether it is necessary and/or operationally feasible to implement data monitoring and reporting processes to remove and record all compromising data
4.8	Is the dataset used relevant?	
4.9	Where personal data is involved, is it collected for the intended purposes?	– Consider whether it is relevant to create internal data classification principles developed based on legal and data governance frameworks and standards (e.g. the International Organization for Standardization (ISO) guidelines)

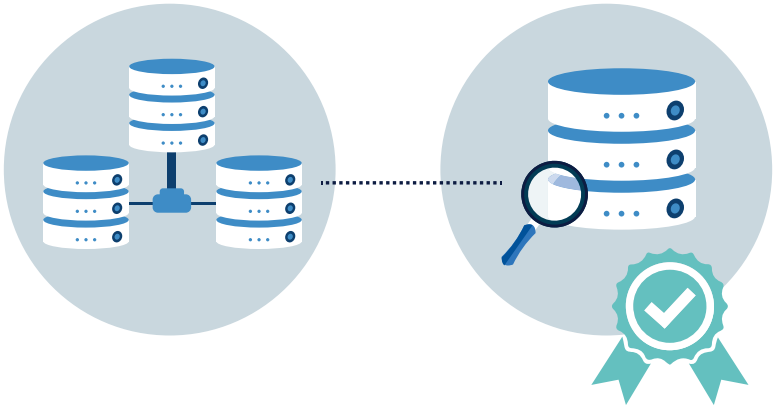
4.10 Is the dataset used well-structured and in a machine-understandable form?

- Consider setting up an extraction, transformation and loading (ETL) process. Prior to the ETL process, it might be useful for the data engineering team to be briefed on the objective of the AI solution and the business needs. After the ETL process, relevant teams (e.g. data engineering team and the business team) to check that the extraction and transformation of the datasets are performed correctly, and aligned to the business needs and intended purpose of the AI solution

Relevant only in limited scenarios:

4.11 If the dataset used has been joined from multiple datasets, were the extraction, transformation and other relevant operations performed correctly?

- Consider whether it is necessary to denormalise or transform the datasets to improve performance or to aid feature engineering
- Consider implementing unit tests to validate that each data operation is performed correctly prior to deployment
- Consider implementing monitoring mechanisms to ensure that changes to upstream data sources do not impact the model adversely, such as the removal of certain populations of data



4.12 If any human has filtered, applied labels, or edited the data, did your organization implement measures to ensure the quality of dataset used?

- Consider whether it is necessary to assign roles to the entire data pipeline to enforce accountability. This would allow an organization to trace who manipulated data and by which rule

Data for Model Development – Minimizing inherent bias

4.13 Did your organization take steps to mitigate unintended biases in the dataset used for the AI model, especially omission bias and stereotype bias?

4.14 Did your organization use a complete dataset by not removing data attributes prematurely to minimize risk of inherent bias?

- Consider taking steps to mitigate inherent bias in datasets, especially where social or demographic data is being processed for an AI system whose output directly impacts individuals
- Consider defining which data fields contain sensitive or protected attributes. In addition, consider checking for indirect bias by measuring which data fields are predictive of protected and sensitive attributes, and which of those data fields are causative of the target outcomes versus mere proxies for protected and sensitive attributes
- Consider whether it is useful to auto-mosaic any consumer physical features (e.g. face) and other personally identifiable information to prevent this information from being collected if it is not necessary. This could minimize potential risk for bias based on personal data instead of transactional behaviour

Relevant only in limited scenarios:

4.15 Did your organization take steps to mitigate biases that may result from data collection devices (e.g. cameras and sensors)?

- Consider whether it is necessary to identify potential biases of data annotation
- Consider whether not to remove data attributes and data items from the datasets prematurely
- Consider whether it is relevant to use statistical tools to evaluate bias (e.g. use "leave one out" to determine over-reliance on variables) – and implement continual monitoring to ensure the AI model stays within pre-defined parameters
- Consider whether it is useful to create an AI library containing datasets to test for potential unintended bias
- Consider defining which measure of bias the organization is trying to detect and remove (e.g. disparate treatment versus disparate impact)

4.16 Is the dataset used to produce the AI model fully representative of the actual data or environment the AI model may receive or function in?

To mitigate selection bias, consider:

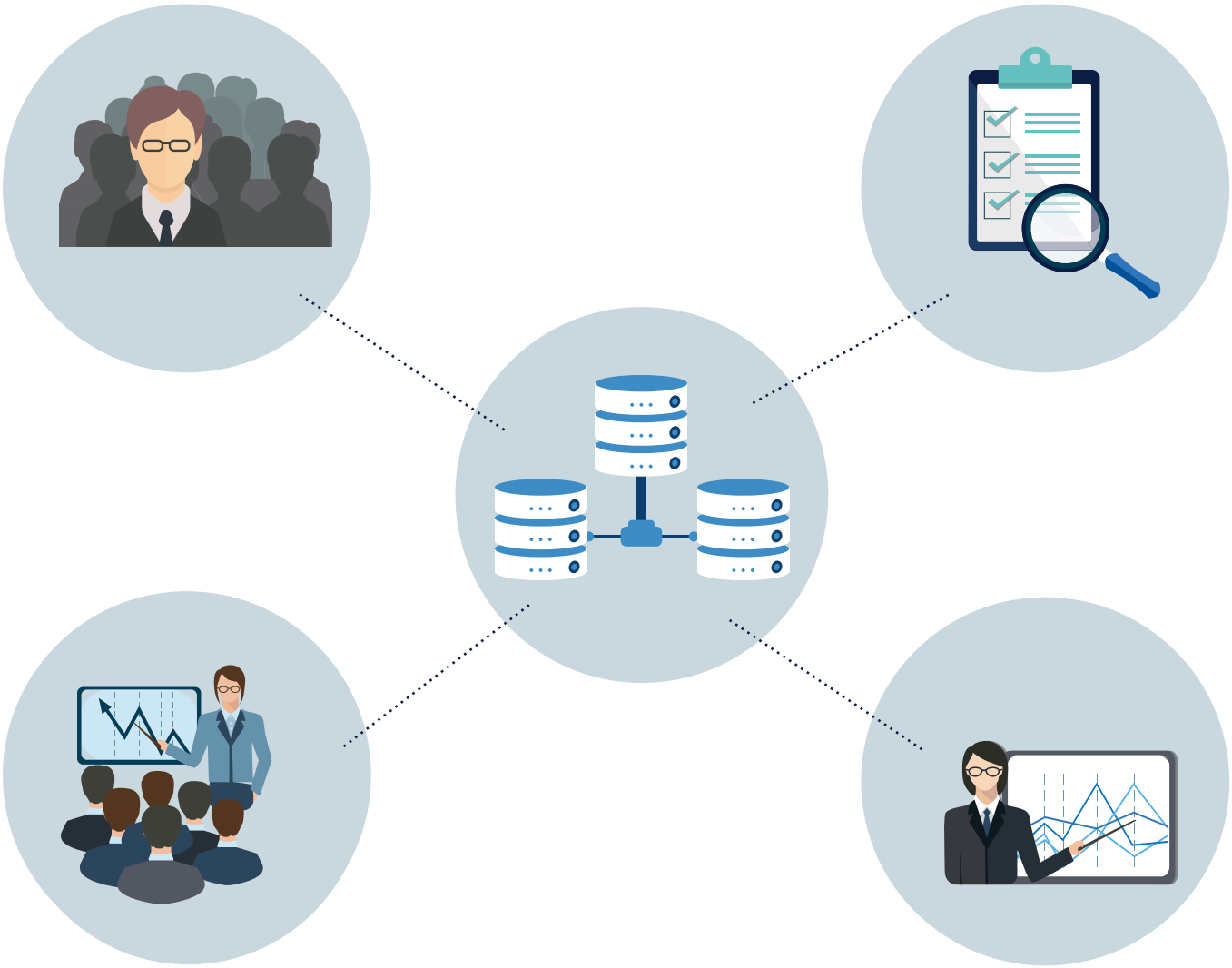
- Benchmarking data distributions against population statistics to identify and quantify how representative the data is
- Whether it is useful to adopt a random assignment approach for the sample data
- Whether it is useful to use quality metrics (e.g. completeness, freshness and context) to evaluate whether the dataset used for the AI model is fit for purpose and matches the population it is intended to represent
- Whether it is necessary to use a heterogeneous dataset (i.e. data collected from different demographic groups or from a variety of reliable sources)
- Whether it is necessary to use training data across different communities, events and attributes

Data for Model Development – Different datasets for training, testing and validation

4.17	Did your organization use different datasets for training, testing and validation of the AI model?	– After training of the AI model, consider validating the AI model using a separate validation dataset – Consider conducting statistical tests (e.g. Area under the Receiver Operating Characteristic Curve (ROC) and stationarity, multi-collinearity tests) to evaluate and validate the AI model's ability to predict results
4.18	Did your organization test the AI model used on different demographic groups to mitigate systematic bias?	– Consider whether it is necessary to check for data drift between the different datasets and making the AI robust to any differences – Consider whether it is necessary to test the results of different AI models to identify potential biases produced by a certain model

Relevant only in limited scenarios:

4.19	Did your organization split a large dataset into subsets to mitigate risks of systematic bias when validating the AI model?	



Data for Model Development – Periodic review and updating of datasets

4.20 Did your organization periodically review and update datasets to ensure its accuracy, quality, currency, relevance and reliability?

4.21 Did your organization implement measures to minimize reinforcement bias?

To ensure data accuracy, quality, currency, relevance and reliability, consider:

- Whether it would be useful to schedule regular review of datasets
- Whether it would be necessary to update the dataset periodically with new data that was obtained from the actual use of the AI model deployed in production or from external sources
- Allocating the responsibility to a relevant personnel to monitor on a regular basis whether new data is available
- Exploring if there are tools available that can automatically notify your organization when new data becomes available
- Deploying a new challenger model that shadows all of the predictions and decisions made by the main AI model, and train the challenger model on newer data than the main AI model. Flag when the challenger model is consistently outperforming the main deployed AI model as this indicates that the patterns in the data have changed and that the old data is no longer valid. This would be a trigger for a review of the data, and your organization would need to consider if the challenger model should become the new main deployed model
- Regularly retrain and build a new adversarial machine learning model that predicts whether a data row is from the current period or from the AI training period. If the adversarial model cannot predict significantly different probabilities for the data source-time period, you know that the data has not changed. However, if it has any success in predicting the source of any rows, that indicates that your data is changing and highlights how it is changing. This should trigger a review of the data and possibly retraining of the AI
- Mitigating bias by post-processing the model if the model bias is explainable and is in line with the bias in the data. For example, if the result from the AI model does not give a desired feature (e.g. gender mix) but the training model exhibits similar bias, consider running two AI models – one for each gender – and get the desired gender mix as a post-processing step. If the AI model bias is not understood, your organization has to evaluate whether the model is still applicable to the case it is used for

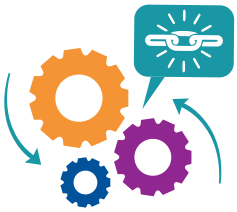
Algorithm and Model

4.22 Did your organization identify relevant features or functionalities that have the greatest impact on your organization's stakeholders?

4.23 Did your organization identify which measures will be the most effective in building trust with your organization's stakeholders?

– Consider prioritizing:

Explainability



Robustness

Regular tuning

– Measures pertaining to traceability, reproducibility and auditability could be more resource-intensive and may only be relevant for specific purposes such as accreditation and certification

Algorithm and Model – Explainability

4.24 Can your organization explain how the deployed AI model functions and arrives at a particular prediction?

- To enhance explainability, consider:
  - Implementing supplementary explanation strategies to explain AI models, especially for models that are less interpretable. Examples of these strategies include the use of surrogate models, partial dependence plots, global variable importance/interaction, sensitivity analysis, counterfactual explanations, or self-explaining and attention-based systems. These strategies help make the underlying rationale of an AI system's output more interpretable and intelligible to those who use the system. It is possible to use a combination of these strategies to improve the explainability of an AI model's decision
  - Generating model reports that contain the level of explainability of each feature
  - Putting in place a factsheet outlining the details on how the AI model operates, including how the model was trained and tested (with what types of data), its performance metrics, fairness and robustness checks, intended uses and maintenance
  - Developing a forecasting model that mimics the dynamics of the real-world business situation that is in line with the user's expectation of the business dynamics
  - Training a simpler version of the model to provide better explanation about the inner workings of the complex model
  - Having assessed trade-offs, use simpler models such as linear regression instead of more complex ones like neural networks
  - Identifying and explaining model limitations to minimize potential for misuse
- Consider whether it is relevant to request assistance from the AI solution provider to explain how the identified AI solution functions
- Consider whether it is useful to use visualizations (e.g. graphs) to explain technical predictions at the model and the individual level
- Consider whether it is useful to explain decisions in narrative terms (e.g. correlation between factors) and use simple indicators to measure output/outcomes (e.g. use "high/medium/low" instead of percentages to measure risk aversion)
- Consider documenting information/guiding descriptors (e.g. database description, model description, evaluation parameters) for AI modelling outputs to provide insights on major contributing factors of each model
- Consider using the Local Interpretable Model-Agnostic Explanations (LIME) technique to explain contributing factors that drive the output of the AI model and SHapley Additive exPlanation (SHAP) to explain how much a particular feature contributed to the decision of the AI model, and related techniques (e.g. Leave One Covariate Out, or LOCO, counterfactual, partial dependence and Individual Conditional Expectation, or ICE to explain the importance of a feature and how the values of that feature affect the outcome

Algorithm and Model – Repeatability

4.25 Where explainability cannot be practically achieved, did your organization consider lesser alternatives?

Where practical and/or relevant, consider:

- Conducting repeatability tests in a production environment
- Performing counterfactual fairness testing
- Identifying exceptions and implement measures to handle them
- Ensure that the AI model trained on time-sensitive data remains relevant
- Implementing measures to test repeatability to validate their models (e.g. observe model outcomes and out-of-time validations) and ensure AI models pass validation tests before deployment
- Testing for error rates of the AI model when applied to different subgroups of the target population
- Conducting simulations to collect and correlate data from different ecosystems for quality control and ensure real-world validation of the AI model before final deployment
- Implement version control so that it is possible to test an older version of the model
- Producing a candidate model. At the same time, produce different challenger models and select those that best represent the business issue. Compare it with the candidate model selected to demonstrate the process and rigour of evaluating AI models. Organizations may wish to consider documenting the justification of producing these models and how they have been used

Algorithm and Model – Robustness

4.26	Did your organization ensure that AI model deployed is sufficiently robust?	– Consider designing, verifying and validating the AI model to ensure that it is sufficiently robust – Consider whether it is relevant to conduct adversarial testing on the AI model to ensure that it is able to handle a broader range of unexpected input variables (e.g. unexpected changes or anomalies) – Consider whether it is necessary to put in place back-up systems, protocols or procedures in the event the AI model produces unacceptable/inaccurate results, or fails

Algorithm and Model – Active monitoring, review and tuning

4.27	Did your organization perform active monitoring, review and regular model tuning when appropriate (e.g. changes to customer behaviour, commercial objectives, risks and corporate values)?	Where practical and/or relevant: – Consider updating the AI model with new data points – set up an automated pipeline to update the model with newer data points via the extraction, transformation and loading (ETL) process, and retrain the model periodically when new data points are added. It might be useful to record when the AI model is being updated, how it is being updated and how this affects the outputs of the AI model – At each model update, consider including examples of output that were misclassified as true errors from the last model update into the training dataset. Before deploying the updated model to production, organization can apply two rounds of testing: compute certain cross-validation metrics for the model (e.g. accuracy, false positive/negative rate, ROC and confusion matrix), by excluding test example from the model's training datasets; if the cross-validation metrics are positive, apply a second, independent testing round on new examples that are not included in the datasets – For ad hoc changes (e.g. changes to market dynamics, commercial objectives and environment), consider whether it is useful to gather feedback from AI model users via multiple channels (e.g. email distribution lists, in-app feedback and periodic user discussion forums). Data scientists may use this feedback to update assumptions in the AI model – Consider conducting on-site observations to solicit feedback and assess performance of the AI model

4.28	Did the AI model testing reflect the actual production environment it is supposed to operate in?	Where practical and/or relevant, consider: – Whether the data used has similar characteristics and is in the same structure as the production environment – Using the same version of the AI model for testing and in products – Using consistent library and dataset versions – Using out-of-sample testing to ensure that the AI model balances accuracy versus over-fitting – Creating test cases and run several model scenarios (i.e. what-ifs) to test model efficacy. This might be relevant for applications where the AI model is solving a puzzle (e.g. assigning resources to create a plan or a schedule) – Running a proof-of-concept with customers and review its results to determine the real-life performance of the AI model and its impact
4.29	Did your organization assess the degree to which the identified AI solution generalized well and failed gracefully?	– To monitor the degradation of models, consider setting up an automated tool that will alert data scientists when the model performance is subpar or below an acceptable threshold To assess whether the AI solution failed gracefully, consider: – Using confidence levels and thresholds as a mechanism for accountability to consider perceived outcomes and aid communication to stakeholders – Whether the AI model produces an error log/message to explain why it failed – Whether a process owner has been identified to triage the problem – Whether there is adequate communication of AI system failure, especially to external stakeholders – Whether your organization has put in place a business continuity plan

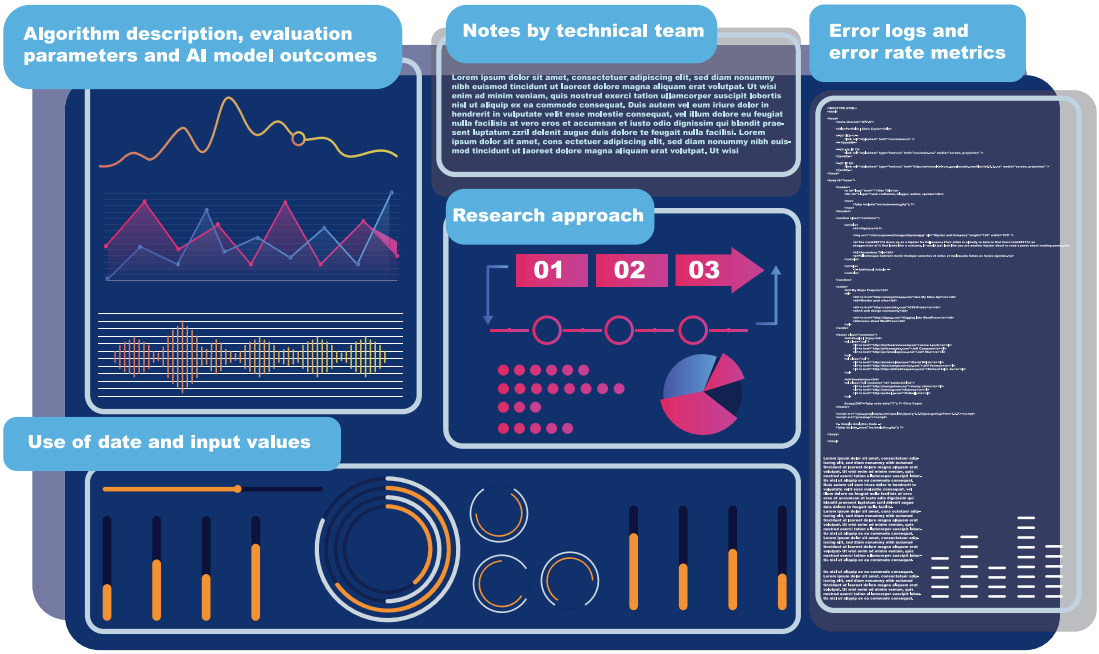
Algorithm and Model – Traceability

Relevant only in limited scenarios:

4.30 Did your organization document the relevant information such as datasets and processes that yield the AI models' decisions in an easily understandable manner?

Where practical and/or relevant, consider:

- Whether it is useful to track the AI model's decision-making process and performance using standard documentation (e.g. dashboard). Examples of information to track could include:
  - Project objectives
  - Use of data and input values
  - Research approach
  - Algorithm description, evaluation parameters and AI model outcomes
  - Error logs and error rate metrics (e.g. false acceptance rate and throughput metrics)
  - Notes by technical team
- Building an audit trail to document the decision-making process
- Keeping a copy of training data and documenting how the data was processed
- Whether it is necessary to implement a black box recorder
- Whether it is useful to ensure that all data relevant to traceability is stored appropriately and retained for durations relevant to the industry
- Whether it is necessary to document the version if the AI model used open-source codes



Algorithm and Model – Reproducibility

Relevant only in limited scenarios:

4.31 Did your organization engage an independent team to check if they can produce the same or very similar results using the same AI method based on the documentation relating to the model made by your organization?	Where practical and/or relevant, consider: – Whether it is relevant to take into account specific contexts or particular conditions that have an impact on the results produced by the AI method – Whether it is useful to make available replication files (i.e. files that replicate each step of the AI model's developmental process) to facilitate the process of testing and reproducing behaviour – Whether it is relevant to check with the original developer on whether the model's results are reproducible

Algorithm and Model – Auditability

Relevant only in limited scenarios:

4.32 Has your organization put in place relevant documentation, procedures and processes that facilitate internal and external assessments of the AI system?	Where practical and/or relevant, consider: – Whether the AI system can be evaluated by internal or external assessors – Whether it is useful to keep a comprehensive record of data provenance, procurement, pre-processing, how the data has been processed, lineage of the data, storage and security – Whether it is useful to centralize information digitally in a process log

Section 5: Stakeholder Interaction and Communication

To help organizations implement good communication practices to inspire trust and confidence among their stakeholders when deploying AI

Guiding questions	Useful industry examples, practices and guides for consideration
Operationalizing communication strategy based on purpose and audience	

5.1 Has your organization identified the various internal and external stakeholders that will be involved and/or impacted by the deployment of the AI solution?
 	
5.2 Did your organization consider the purpose and the context under which the explanation is needed?
 	
5.3 Did your organization tailor the communication strategy and/or explanation accordingly after considering the audience, purpose and context?
	Where practical and/or relevant, consider:
	
	– Customizing the communication message for the different stakeholders who are impacted by the AI solution
	
	– Providing different levels of explanation at:
	
	  – Data (e.g. types and range of data used in training the algorithm)
	
	  – Model (e.g. features and variables used and weights)
	
	  – Human element (e.g. nature of human involvement when deploying the AI system)
	
	  – Inferences (e.g. predictions made by the algorithm)
	
	  – Algorithmic presence (e.g. if and when an algorithm is used)
	
	  – Impact (e.g. how the AI solution affects users)
	
	After identifying the audience, purpose and context, organizations should consider prioritising what needs to be explained to the different stakeholders:
	
	– Providing process-based explanation (e.g. considerations on the data used, model selection and steps to mitigate risk of the AI solution) and/or outcome-based explanation (i.e. the purpose and impact/consequences of the AI solution on users)
	
	– Both the language and complexity of concepts in communication, and use heuristics for stakeholders that are less technical
	
	– Consider charting the stakeholder journey and identifying the type of information, level of details and objective of informing the customer at each significant milestone. This could minimize information fatigue

General information on the AI used is provided to potential users for them to decide whether to use it

More specific information is provided to users so they can understand how the app works during use

More in-depth information is provided to users if they challenge a decision or wish to provide feedback

5.4 Did your organization inform relevant stakeholders that AI is used in your products and/or services?

Relevant only in limited scenarios:

5.5 In circumstances where technical explainability/explicit explanations may not be useful to the audience, did your organization provide implicit explanation (e.g. counter-factuals)?

In disclosing information to relevant stakeholders, consider:

- Disclosing to consumers which data fields were most important to the decision-making process and the values in those data fields
- Whether it is relevant to provide information at an appropriate juncture on what AI is and when, why and how AI has been used in decision-making about the users. Organizations could also document and explain the reason for using AI, how the AI model training and selection processes were conducted, the reasons for which decisions were made, as well as steps to mitigate risks of the AI solution on users. By having a clear understanding of the possible consequences of the AI-augmented decision-making, users could be better placed to decide whether to be involved in the process and anticipate how the outcomes of the decision may affect them
- Whether it is necessary to provide information on the role and extent that AI played in the decision-making process (e.g. statistical results and inferences) in plain language and in a way that is meaningful to the individuals impacted by the AI solution (e.g. infographics, summary tables and simple videos). Organizations could also use decision trees or simple proxy model representations to visualize complexity and justify decisions by the AI model to stakeholders
- Publishing a privacy policy on your organization's website to share information about AI governance practices (e.g. data practices, and decision-making processes). The general disclosure notice could include:
  - Disclosure of third-party engagement
  - Definition of data ownership and portability
  - Depiction of the data flow and identify any leakages
  - Identification of standards the company is compliant with as assurance to customers
- Informing users if an interaction involves AI, and how the AI-enabled features are expected to behave during normal use. For example, your organization could consider informing users on the website landing page that they are interacting with an AI-powered chatbot
- In the context of B2B, stating clearly in client agreements, contracts or licences when and how AI technology will be used

5.6 Did your organization disclose the manner in which an AI decision affects individuals and if the decision could be reversible?

In disclosing information to relevant stakeholders, consider:

- Using easy-to-understand language
- Whether it is useful to document the AI model workflow with a decision tree to help regulators visualize the complexity of decision-making and justify decisions made by the AI model
- Whether it is relevant to refer to PDPC's Guide to Notification, Fry readability graph, the Gunning Fog Index, the Flesch-Kincaid readability tests. Besides textual communications, organizations could use visualization tools, graphical representations, summary tables, or a combination of these to aid in communication with stakeholders
- Notifications to customers could include:
  - Explanation of the outcomes of automated decisions on users, and show data source and lineage where possible
  - Depiction of how the data is trained and labelled
  - Disclosure of statistics and information on outcomes and model performance

---

5.7 Did your organization evaluate whether your AI governance structure and processes are in line with changing standards?

- Consider whether it is relevant to keep abreast of local and international developments relating to AI governance
- Consider whether it is necessary to also provide an explanation on how/why an ethical evaluation was conducted

5.8 Did your organization make available the outcome of the evaluation to relevant stakeholders?

Policy for explanation

5.9	Did your organization develop a policy on explanations to be provided to individuals, and implement the policy accordingly?	– Consider whether it is applicable to publish an explanation of when AI is used – Consider identifying educational tools (e.g. leaflets, newsletters, user guides and white papers) and conducting briefing sessions or information campaigns that could help clients/customers understand the explanation

Testing the user interface

5.10	Did your organization address usability problems and test whether user interfaces served their intended purposes?	– Consider whether it is useful to conduct user testing
		– Consider placing clients/consumers at the centre, when designing the user interface and deploying the AI solution by:
		  – Consulting the community or end users at the earliest stages of development to ensure there is transparency on the technology used and how it is deployed
		  – Co-designing the identified AI solution with clients/users from the beginning to create a friendly user interface
		  – Conducting outreach and building a feedback loop to collect client feedback during the co-design process
		  – Sharing part of the operations dashboard with customers to build trust
		  – Assessing whether there is a need to have just-in-time consent over push notifications to customers
		– Consider whether it is necessary to embed contextual consent in the user experience and design process of AI-powered applications, and collect data from users only when they need to access a function in the application

5.11	Did your organization inform users that they are interacting with AI, and their responses would be used to train the AI model?	– Consider implementing an Acceptable User Policy or Code of Conduct to inform users of the terms and conditions of using the AI system (e.g. to prohibit hate speech and bullying)
Relevant only in limited scenarios:		
5.12	If users' responses are used to train the AI model, did your organization implement measures to filter out misleading and/or inaccurate responses?	– Consider designing the AI model to identify abnormal behaviour and prevent manipulation (e.g. for chatbots, identify users who appear to respond too fast, or trigger parts of the bot code that other users do not) – For bots that employ automatic or supervised learning techniques, consider whether it is necessary to ensure that the AI system is able to distinguish between maliciously-introduced data and data that is rare, yet valid and important
Option to opt-out		
Relevant only in limited scenarios:		
5.13	Did your organization offer the option to opt out of the identified AI solution by default or only on request?	– Consider informing users of the consequences of choosing to opt-out, if such an option is available

Communication channels for feedback, queries and decision review

5.14	Did your organization provide a feedback channel for feedback or queries?	– Consider providing an avenue for individuals to submit updated data about themselves – Consider whether it is necessary to set expectations as to whether the user will receive any response to feedback provided – Consider providing a hotline or email contact of relevant personnel such as a data protection officer or quality service manager on the organization's website – Consider whether it is necessary to put in place measures to ensure that public queries and feedback are addressed in a timely manner (e.g. a minimum response time)
5.15	Is the feedback channel managed by appropriate personnel?	
5.16	Did your organization provide an avenue for users to request for a review of material AI decisions that have affected them?	– Consider whether it is useful to describe the process for appealing a decision – Consider whether it is useful to keep a record of chatbot conversations with users

Annex

The International Organization for Standardization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE) are developing relevant AI standards. Organizations may consider referring to them, as and when they become available.

Some relevant ISO Standards include:

ISO/IEC 22989	Information technology — Artificial intelligence – Concepts and terminology
ISO/IEC 23053	Information technology — Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
ISO/IEC 20546	Information technology — Big data — Overview and vocabulary
Parts 1, 2, 3 and 5 of ISO/IEC 20547	Information technology – Big data reference architecture
ISO/IEC 24668	Information technology — Artificial intelligence —Process management framework for Big data analytics
ISO/IEC DTR 24027	Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making
ISO/IEC DTR 24028	Information technology — Artificial Intelligence — Overview of trustworthiness in Artificial Intelligence
ISO/IEC DTR 24029-1	Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
ISO/IEC DTR 24368	Information technology — Artificial intelligence — Overview of ethical and societal concerns
ISO/IEC 23894	Information Technology — Artificial Intelligence — Risk Management
ISO/IEC DTR 24030	Information technology — Artificial Intelligence — Use cases
ISO/IEC DTR 24372	Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems
ISO/IEC 38507	Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

Some relevant IEEE standards include:

IEEE P7000™	Model Process for Addressing Ethical Concerns During System Design
IEEE P7001™	Transparency of Autonomous Systems
IEEE P7002™	Data Privacy Process
IEEE P7003™	Algorithmic Bias Considerations
IEEE P7004™	Standard on Child and Student Data Governance
IEEE P7005™	Standard for Transparent Employer Data Governance
IEEE P7006™	Standard for Personal Data Artificial Intelligence (AI) Agent
IEEE P7007™	Ontological Standard for Ethically Driven Robotics and Automation Systems
IEEE P7008™	Standard for Ethically Driven Nudging for Robotic, Intelligent, and Automation Systems
IEEE P7009™	Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
IEEEP70010™	Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems

Acknowledgements

The Personal Data Protection Commission, Info-communications Media Development Authority and World Economic Forum's Centre for the Fourth Industrial Revolution express their sincere appreciation to the following for their valuable feedback to this Implementation and Self-Assessment Guide for Organizations:

- Accenture
- Adobe
- Alibaba Group
- Amazon Web Services
- AsiaDPO
- Basis.AI
- Best Practice AI Ltd
- Callsign
- ConnectedLife
- CUJO AI
- DataRobot Singapore
- Deepen AI
- DBS Bank
- Element AI
- Facebook
- Fullerton Health
- Google
- Great Eastern Life Assurance Co. Ltd
- IBM Singapore and ASEAN
- IKEA
- KPMG
- Llamasoft
- Malong Technologies
- Manulife
- Mastercard
- Microsoft
- MSD International GmBH (Singapore branch)
- Nanyang Technological University, School of Computer Science and Engineering
- National University of Singapore, Institute of Systems Science
- OCBC Bank
- Omada Health
- Primer
- PwC
- pymetrics
- Salesforce
- Singapore Computer Society
- Standard Chartered Bank
- Suade Labs
- Taiger
- Telenor Group
- Temasek International
- Unipol Group
- Untangle AI
- Visa WorldWide Pte. Limited

Endnotes

1. The PDPC's Second Edition of the Model AI Governance Framework can be downloaded at Go.gov.sg/ai-gov-mf-2
2. The PDPC's Compendium of Use Cases can be downloaded at Go.gov.sg/ai-gov-use-cases
3. The PDPC's Guide to Data Protection Impact Assessments can be downloaded at https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/guide-to-dpias---011117.pdf
4. These terms are used as defined in the PDPC Anonymization Advisory Guidelines and Technical Companion Guide
5. The IMDA's Trusted Data-Sharing Framework can be downloaded at www.imda.gov.sg/AI-and-Data

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744

contact@weforum.org
www.weforum.org