GUIDING PRINCIPLES

2.7 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 as far as possible. 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.

Organisations should ensure that AI decision-making processes are explainable, transparent and fair, while AI solutions should be human-centric.

2.8 Like other technologies, AI aims to increase human productivity. However, unlike earlier technologies, some aspects of autonomous predictions or decisions made by AI may not be fully explainable. As AI technologies can make decisions that affect individuals, or have a significant impact on society, markets or economies, organisations should consider using this Model Framework to guide their deployment of AI.

2.9 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. Where necessary, organisations may wish to refer to the compilation of AI ethical principles in **Annex A**. 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 (e.g. safety) 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 and pre-existing corporate governance structures, and thereby aid in guiding the adoption of AI in an organisation.

ASSUMPTIONS

2.10 The Model Framework aims to discuss good data management practices in general. The Model Framework is mainly applicable to machine learning models (as compared to pure decision tree-driven AI models).

2.11 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.

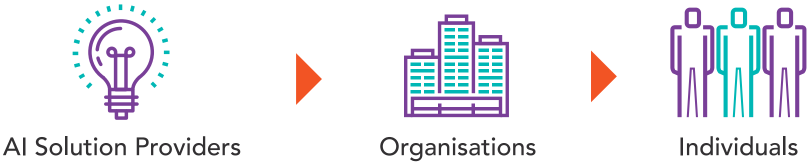
2.12 Adopting this voluntary Model Framework will not absolve organisations from compliance with current laws and regulations. However, as this is an accountability-based framework, adopting it will assist organisations in demonstrating that they had implemented accountability-based practices in data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

2.13 Further, it should be noted that certain industry sectors (such as in the finance, healthcare, and legal sectors) may be regulated by existing sector-specific laws, regulations or guidelines relevant to the sector. For example, the Monetary Authority of Singapore published the Principles to Promote Fairness, Ethics, Accountability and Transparency in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector (the "**FEAT Principles**") to provide guidance to firms that use AI and data analytics to offer financial products and services.[2] Organisations are advised to remain mindful of such laws, regulations and guidelines, as adopting the Model Framework does not mean that organisations are in compliance with such sector-specific laws, regulations or guidelines.

[2] Monetary Authority of Singapore, "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector" (12 November 2018) <https://www.mas.gov.sg/publications/monographs or-information-paper/2018/FEAT>.

DEFINITIONS

2.14 The following simplified diagram depicts the key stakeholders in an AI adoption process discussed in the Model Framework. The adoption process does not distinguish between business-to-consumer ("**B2C**"), business-to-business ("**B2B**"), and business-to-business-to-consumer ("**B2B2C**") relationships.


AI Solution Providers

Organisations

Individuals

2.15 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:

"AI" refers to a set of technologies that seek to simulate human traits such as knowledge, reasoning, problem solving, perception, learning and planning, and, depending on the AI model, produce an output or decision (such as a prediction, recommendation, and/or classification). AI technologies rely on AI algorithms to generate models. The most appropriate model(s) is/are selected and deployed in a production system.[3]

"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 B2B2C 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" can, depending on the context, 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] This definition of AI was adapted from various sources, and contextualised accordingly for the purposes of this Model Framework. It should not be taken to be an authoritative or exhaustive definition.

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 the level of human involvement in AI-augmented decision-making**
A methodology to aid organisations in setting its risk appetite for use of AI, i.e. determining acceptable risks and identifying an appropriate level of human involvement in AI-augmented decision-making.

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

d. **Stakeholder interaction and communication**
Strategies for communicating with an organisation's stakeholders, and the management of relationships with them.

3.2 Organisations adopting this Model Framework may find that not all elements are relevant. This Model Framework is meant to be flexible, and organisations can adapt the Model Framework to suit their needs and adopting those elements that are relevant.

3.3 To help organisations better understand the Model Framework, we have included (in each section) illustrations demonstrating how real-world companies have implemented certain practices described in that specific section. In addition, the PDPC has also released a Compendium of Use Cases that illustrates how various local and international organisations have put in place AI governance practices that are aligned to all sections of the Model Framework.

INTERNAL GOVERNANCE STRUCTURES AND MEASURES

3.4 This section is intended to guide organisations in developing appropriate internal governance structures that allow organisations to have appropriate oversight over how AI technologies are brought into their operations and/or products and services.

3.5 Internal governance structures and measures help to ensure robust oversight over an 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, while ethical considerations can be introduced as corporate values and managed through ethics review boards or similar structures.

Ethical considerations can be introduced as corporate values and managed through ethics review boards or similar structures.

3.6 Organisations may also consider determining 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 the operational level, if necessary. The sponsorship, support and participation of the organisation's top management and its board of directors in the organisation's AI governance are crucial.

3.7 Organisations may wish to consider including features that are relevant to the development of their internal governance structure, such as:

1. Clear roles and responsibilities for the ethical deployment of AI

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 can be allocated include:

   i. Using any existing risk management framework and applying risk control measures (see "Risk management and internal controls" below) to:

      o 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.).

      o Decide on the appropriate level of human involvement in AI-augmented decision-making.

      o Manage the AI model training and selection process.

ii. Maintenance, monitoring, documentation 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 stakeholders to provide disclosure and effective feedback channels.

iv. Ensuring relevant staff dealing with AI systems are properly trained. Where applicable and necessary, staff who are working and interacting directly with AI models may need to be trained to interpret AI model output and decisions and to detect and manage bias in data. Other staff whose work deals with the AI system (e.g. a customer relationship officer answering customer queries about the AI system, or a salesperson using an AI-enabled product to make a recommendation) should be trained to be at least aware of and sensitive to the benefits, risks and limitations when using AI, so that they know when to alert subject-matter experts within their organisations.

2. Risk management and internal controls

a. Organisations can consider implementing a sound system of risk management and internal controls that specifically addresses the risks involved in the deployment of the selected AI model.

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 could focus on why the AI model had a certain level of confidence.

   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.

CUJO AI:

ILLUSTRATION ON INTERNAL GOVERNANCE STRUCTURES AND MEASURES

CUJO AI is a network intelligence software company in the telecommunications operators' market. Headquartered in the US, it seeks to develop and deploy AI to improve security, control, privacy of connected devices in homes and businesses.

CUJO AI has implemented **clear internal governance structures and measures** to ensure robust oversight of its use of AI. Its multi-stakeholder governance structures facilitate decisions at appropriate levels:

A Research Board consisting of the Chief Technology Officer, the Head of Labs and the Chief Data Scientist, approves the AI development and deployment. In particular, the Chief Technology Officer oversees four technical teams which consists of more than 100 employees.

Their roles and responsibilities are clearly defined:

a. Research team performs data analysis, research and develop Machine Learning ("**ML**") models and AI algorithms;
b. Engineering team builds software, cloud services and applications;
c. Operation team deploys the AI model and upgrade platform; and
d. Delivery team engages with operators and integrate services.

An Architecture Steering Group ("ASG") consisting of the Chief Technology Officer, Chief Architect Officer and lead engineers, ensures the robustness of the AI/ML models before deployment. The ASG has bi-weekly meetings where the research team shares its findings on the ML models and AI algorithms (e.g. data, approach and assumptions).

PhD-level employees oversee the AI development and deployment process, and strive to implement academic review standards for each new feature development.

In addition, CUJO AI has developed a general **Code of Ethics** ("**Code**") for its employees. All new employees are introduced to the CUJO AI local country document and process repository. For example, CUJO AI's office in Finland provides its employees with an electronic "CUJO employee handbook". The handbook describes in detail the Code, while covering other topics such as business ethics and conduct. Employees carry out their tasks and responsibilities on the basis of the following **ethical principles**:

a. To conduct business in an honest and ethical manner across its various offices around the world;

b. To base decisions on honesty, fairness, respect, responsibility, integrity, trust, and sound business judgment;

c. That no illegal or unethical conduct on the part of officers, directors, employees, or affiliates is in the company's best interest; and

d. Not to compromise the company's principles for short-term advantage.

MASTERCARD:

ILLUSTRATION ON INTERNAL GOVERNANCE STRUCTURES AND MEASURES

Mastercard is a technology company in the global payments industry. Its global payments processing network connects consumers, financial institutions, merchants, governments and businesses in more than 210 countries and territories. To achieve its vision, Mastercard leveraged AI in many applications such as fraud prevention, forecasting future spending trends and improving user retail experience.

To ensure robust oversight of Mastercard's use of AI, Mastercard established a Governance Council to review and approve the implementation of AI applications that are determined to be high risk. The Governance Council is chaired by its Executive Vice President of the Artificial Intelligence Center of Excellence, and whose members include the Chief Data Officer, Chief Privacy Officer, Chief Information Security Officer, data scientists and representatives from business teams.

Mastercard has **defined clear roles and responsibilities** for the Governance Council. Each representative on the Council brings their expertise to the decision-making process:

a. **Chief Data Officer and Chief Privacy Officer** will review the proposal for implementation of AI to ensure that the:

- Data is fit for purpose for AI;
- AI is used for an ethical purpose; and
- Impact to an individual is appropriate and potential harms (including risks to privacy and data protection) are sufficiently mitigated.

b. **Chief Information Security Officer** will ensure that security by design is implemented.

c. **Data Science teams** that build and implement AI are in continued dialogue with the Data Office and the Privacy Office, so that there is continued information sharing regarding the required governance and the lifecycle of a particular implementation of an AI application.

Mastercard has also implemented risk management and internal controls to address the risk involved in the AI deployment. For example, Mastercard conducts initial risk scoring to determine the risk of the proposed AI activity, which includes an evaluation of multiple factors including alignment with corporate initiatives, the data types and sources utilised, and the impact on individuals from AI decisions.

In addition, Mastercard will identify potential mitigants as part of the process to reduce the level of risk posed by the data being collected or potential biases in the activity. If an AI project has been identified as high risk, it will be referred to the Governance Council for review. Low risk projects will not be subjected to a review and can proceed to the model development stage.

DETERMINING THE LEVEL OF HUMAN INVOLVEMENT IN AI-AUGMENTED DECISION-MAKING

3.8 This section is intended to help organisations determine the appropriate extent of human oversight in AI-augmented decision-making.

3.9 Having clarity on the objective of using AI is a key first step in determining the extent of human oversight. Organisations can start by deciding 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). These commercial objectives can then be weighed 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 or expectations of the territories in which the organisations operate.

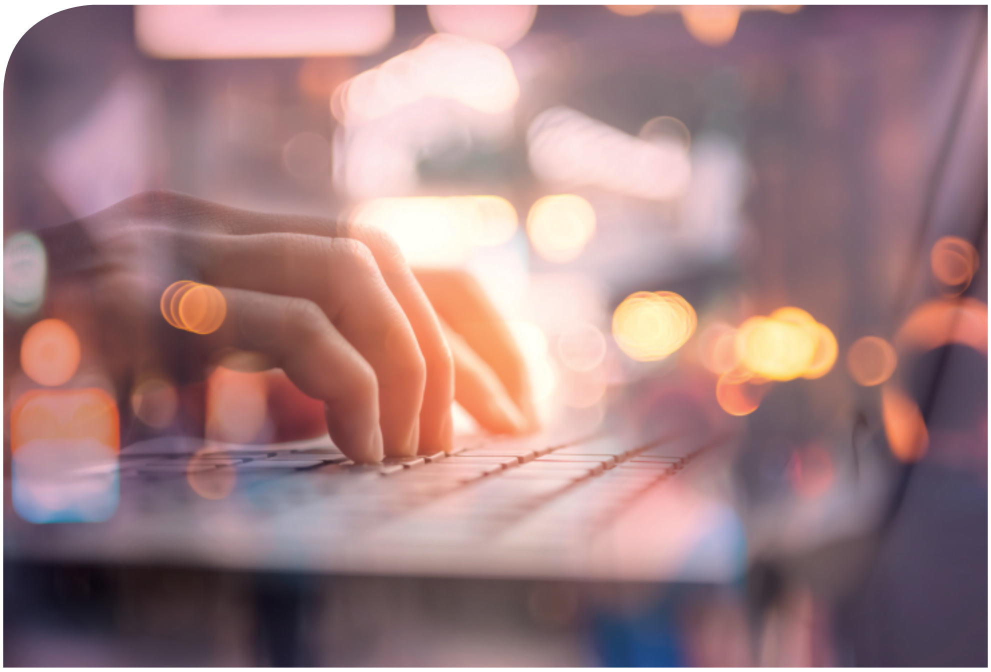
> *Before deploying AI solutions, organisations should decide on their commercial objectives of using AI, and then weigh them against the risks of using AI in the organisation's decision-making.*

3.10 It is also desirable for organisations operating in multiple countries to consider the differences in societal norms, values and/or expectations, where possible. For example, gaming advertisements may be acceptable in one country but not in another. 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 the risks associated with algorithmic trading facilities offered to sophisticated investors.

3.11 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).

3.12 Organisations' weighing of their commercial objectives against the risks of using AI should ideally 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 organisations with a clearly defined and documented rationale.

3.13 As identifying commercial objectives, risks and determining the appropriate level of human involvement in AI-augmented decision-making is an iterative and ongoing process, it is desirable for organisations to 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.



WHAT ARE THE THREE BROAD APPROACHES OF HUMAN INVOLVEMENT IN AI-AUGMENTED DECISION-MAKING?

3.14 Based on the risk management approach described above, the Model Framework identifies three broad approaches to classify the various degrees of human oversight in the decision-making process:

a. **Human-in-the-loop** 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** suggests that there is no human oversight over the execution of decisions. The AI system has full control without the option of human override.

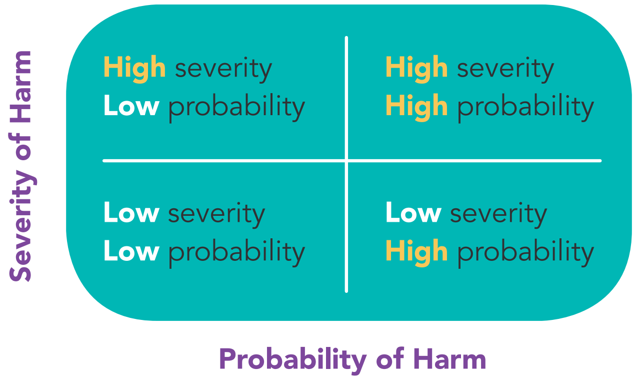
   For example, a product recommendation solution may automatically suggest products and services to individuals based on pre-determined demographic and behavioural profiles. AI can also dynamically create new profiles, then make product and service suggestions rather than relying on predetermined categories.

   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** (or human-on-the-loop) suggests that human oversight is involved to the extent that the human is in a monitoring or supervisory role, with the ability to take over control when the AI model encounters unexpected or undesirable events (such as model failure). This approach allows humans to adjust parameters during the operation 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.

3.15 The Model Framework also proposes a design framework (structured as a matrix) to help organisations determine the level of human involvement required in AI-augmented decision-making. This design framework is structured along two axes: the (a) probability; and (b) severity of harm to an individual (or organisation) as a result of the decision made by an organisation about that individual (or organisation).

3.16 The definition of "harm" and the computation of probability and severity will depend on the context and vary from sector to sector. For example, the considerations of a hospital regarding the harm associated with a wrong diagnosis of a patient's medical condition will differ from the considerations of a clothing store's regarding the harm associated with a wrong product recommendation for apparels.



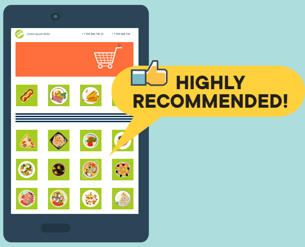
3.17 The matrix, however, should not be taken to imply that the probability of harm and severity of harm are the only factors to be considered in determining the level of human oversight in an organisation's decision-making process involving AI (although they are generally two of the more important factors).[4]

3.18 For safety-critical systems, it would be prudent for organisations to ensure that a person be allowed to assume control, with the AI system providing sufficient information for that person to make meaningful decisions or to safely shut down the system where human control is not possible.

[4] Other factors that organisations in various contexts may consider relevant, could also include: (a) the nature of harm (i.e. whether the harm is physical or intangible in nature); (b) the reversibility of harm, and as a corollary to this, the ability for individuals to obtain recourse; and (c) whether it is operationally feasible or meaningful for a human to be involved in a decision-making process (e.g. having a human-in-the-loop would be unfeasible in high-speed financial trading, and be impractical in the case of driverless vehicles).

USING THE PROBABILITY-SEVERITY OF HARM MATRIX

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 histories. The automation will meet the organisation's commercial objective of operational efficiency.

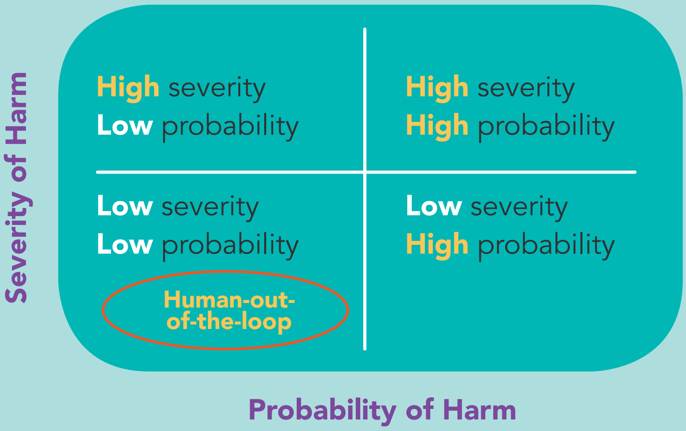


Probability-severity assessment

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 harm*** in 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 (i.e. human-out-of-the-loop).



Severity of Harm

High severity **Low** probability	**High** severity **High** probability
Low severity **Low** probability Human-out-of-the-loop	**Low** severity **High** probability

Probability of Harm

Regular review

The organisation regularly reviews its approach (i.e. human-out-of-the-loop) to re-assess the ***severity*** and ***probability of harm***, and as societal norms and values evolve.

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

SUADE LABS:

ILLUSTRATION ON DETERMINING THE LEVEL OF HUMAN INVOLVEMENT IN AI-AUGMENTED DECISION-MAKING

Suade Labs ("**Suade**") is a RegTech firm that operates globally and is a World Economic Forum Technology Pioneer. Suade provides an AI-enabled solution that allows financial institutions to process large volumes of granular data and generate the required regulatory data, calculations, and reports with the necessary controls and governance. Suade's solution also allows users to analyse the impact of the existing stock of regulation, including the impact of individual pieces of legislation.

In determining the level of human involvement in decision-making using AI, Suade considered the following key factors:

a. Degree of domain knowledge (e.g. legal or policy-making knowledge) required to accurately interpret the results of the algorithm.

b. Cost of non-compliance to regulation if the AI tool does not accurately analyse the impact of regulation and provide correct suggestions for regulatory compliance.

As Suade's solution requires a certain degree of domain knowledge from human experts, and given that the cost of regulatory non-compliance as a result of incorrect recommendations made by the AI solution will be significant to users, Suade has thus adopted a **human-in-the-loop** approach for its AI solution.

On the other hand, when it comes to tuning the AI model, Suade adopts a **human-over-the-loop** approach. In general, Suade tunes the AI model to automatically favour the identification of false positives over false negatives. However, Suade conducted user research, which informed them that some users prefer the model to favour false negatives over false positives. Therefore, Suade adopts a human-over-the-loop approach so that the AI model can be tuned to account for the differing preferences of its users with respect to whether the algorithm produces results that favours false positives or false negatives.

GRAB:

ILLUSTRATION ON DETERMINING THE LEVEL OF HUMAN INVOLVEMENT IN AI-AUGMENTED DECISION-MAKING

Grab is a Singapore-based company that offers ride-hailing transport services, food delivery and e-payment solutions. It uses AI across its platform, from ride allocation, detecting safety incidents, to identifying fraudulent transactions. In particular, Grab uses AI to improve the overall quality of trip allocations and minimise trip cancellations.

To allocate trips successfully, Grab's AI model considers drivers' preferences based on the following key factors:

a. Driver's preferences for certain trip types;

b. Preferred locations where a driver start and end their day; and

c. Other selective driving behaviours.

In determining the level of human involvement in its AI's decision-making for trip allocation, Grab considered the following key factors:

a. The scale of real-time decision-making required. As Grab has to make over 5,000 trip allocations every minute, this would mean an impact to customers in terms of efficiency and cost if a human had to review each trip allocation; and

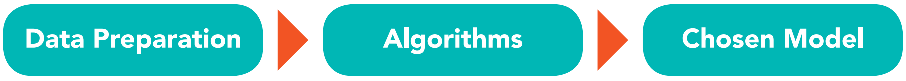
b. The severity and probability to users should the AI model work in a sub-optimal manner.

Among other factors, Grab considered that: (1) it is not technically feasible for a human to make such high volume of trip allocations in a short amount of time; and (2) there is often little or no harm to life should there be less than optimal trip allocations. Hence, Grab decided to adopt a human-out-of-the-loop approach for its AI model deployed for trip allocation, while continuously reviewing the AI model to ensure optimal performance.

OPERATIONS MANAGEMENT

3.19 This section is intended to help organisations adopt responsible measures in the operations aspect of their AI adoption process. A reference AI adoption process is set out in order to provide a context for the recommendations for good governance in respect of the organisation's data, algorithm and AI model.

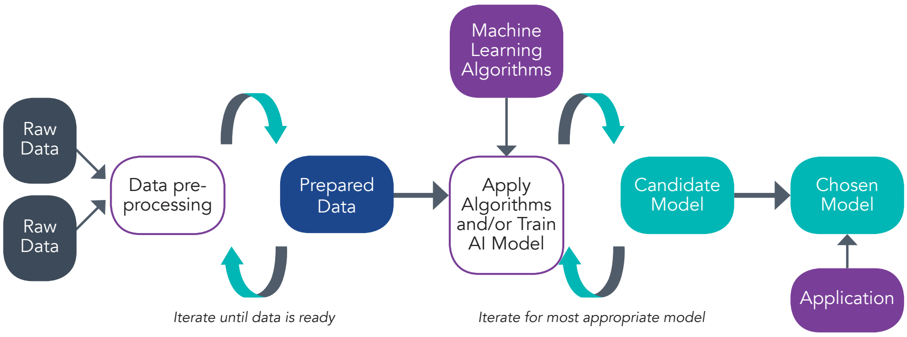
3.20 The Model Framework uses the following generalised AI model development and deployment process to describe phases in implementing an AI solution by an organisation.[5] It should be noted that this process is not always uni-directional – it can, and usually is, a continuous process of learning.



Stage 1:
Raw data is formatted and cleansed so conclusions can be drawn accurately. Generally, accuracy and insights increase with relevance and the amount of data.

Stage 2:
Models are trained on the dataset and **algorithms** may be applied. This includes statistical or machine learning models including decision trees and neural networks. The results are examined and models are iterated until the most appropriate model emerges.

Stage 3:
The **chosen model** is used to produce probability scores that can be incorporated into applications to offer predictions, make decisions, solve problems and trigger actions.


Machine Learning Algorithms
Raw Data
Raw Data
Data pre-processing
Prepared Data
Apply Algorithms and/or Train AI Model
Candidate Model
Chosen Model
Application
Iterate until data is ready
Iterate for most appropriate model

[5] Adapted from "Machine learning at scale" *Microsoft Azure* (2 December 2018) <https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/machine-learning-at-scale> (accessed December 2019).

3.21 During deployment, algorithms such as linear regression algorithms, decision trees, or neural networks are applied for analysis on training datasets. The resulting algorithmic models are examined and algorithms are iterated until a model that produces the most appropriate results for the use case emerges. This model and its results are then incorporated into applications to offer predictions, make decisions, solve problems 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

3.22 Datasets used for building models may come from multiple sources, and could include both personal and non-personal data. The quality and selection of data from each of these sources 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.

> *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.*

3.23 The persons who are involved in training and selecting models for deployment may be internal staff or external service providers. It is ideal for the models deployed in an intelligent system to 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, it would be helpful for relevant departments within the organisation with responsibilities over quality of data, model training and model selection to work together to put in place **good data accountability practices**. These 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.

   In some instances, the origin of data could be difficult to establish. One example could be datasets obtained from a trusted third-party which may have commingled data from multiple sources. It would be prudent for organisations to assess the risks of using such data and manage them accordingly.

b. **Ensuring data quality:** Organisations are encouraged to understand and address 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; and

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

c. **Minimising inherent bias**: 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 contain inherent biases and are encouraged to 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. For example, a dataset consisting only of Asian faces 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.

While identifying and addressing inherent bias in datasets may not be easy, organisations can mitigate the risk of inherent bias by having 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, if it does not lead to a significant reduction in the quality of data in terms of accuracy and representation. However, where this is not possible (e.g. if the organisation is not working with large datasets or are using pre-trained models as in the case of transfer learning), organisations are encouraged to be cognisant of the risks of systematic bias and put in place appropriate safeguards.

e. **Periodic reviewing and updating of datasets:** It would be prudent for datasets (including training, testing, and validation datasets) to be reviewed periodically to ensure accuracy, quality, currency, relevance and reliability. Where necessary, the datasets can be updated with new input data 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.

3.24 Even if only non-personal data are used for the training of AI models (including personal data that has been anonymised), the good data accountability practices above remain relevant.

SUADE LABS:

ILLUSTRATION ON MANAGING DATA FOR MODEL DEVELOPMENT

Suade (introduced above) has developed an AI-enabled solution that helps financial institutions generate the required data and reports to comply with regulatory requirements in the jurisdictions where they operate.

As the data used for Suade's AI model development directly affects its quality and performance, Suade has adopted several good data accountability practices. For example, to ensure that regulatory data comes from a credible and reliable source, Suade obtains and updates regulatory data only from the relevant regulators. In addition, Suade tags the datasets used with additional metadata. This allows Suade to trace datasets back to their original source when needed, such as where inconsistencies are found. Further, in order to trace which particular datasets were used in an AI model, Suade also documents and stores such information pertaining to model development on its database.

Suade also **minimises the inherent risks of AI models** through **responsible data tagging**. By using a larger number of taggers (i.e. people who tag data), Suade aims to make the output of its AI models as neutral as possible, and reduce the risk of its taggers being influenced by the context of the data (which often comprise of text) they are annotating. In other words, Suade uses as many individuals as practicable to tag data to reduce the risk of tagger bias.

In addition, Suade developed a tagging system to facilitate the annotation of data. This system is used to generate training data used by the algorithm. Suade will further develop this tagging system to enhance its ability to manage multiple annotators and to better select datasets used for model training. Suade also periodically updates the tagging system with new data. New training data is subsequently fed repeatedly back into the AI model. This way, the AI model is able to continuously learn from new sets of data.

Another data accountability practice that Suade adopts is the use of validation schema checks at various stages of data transformation. This is a process in which Suade verifies that the data schema accurately represents the data from the source, to ensure that there are no errors in factors such as the data's formatting and content.

PYMETRICS:

ILLUSTRATION ON MANAGING BIASES IN DATASETS FOR MODEL DEVELOPMENT

pymetrics is a technology provider that uses neuroscience insights and audited AI models to help evaluate applicants in a more predictive and less biased manner. To develop an AI model, pymetrics:

- Gets its clients' top-performing employees to go through pymetrics' assessments, and builds a trait profile of an employee that best fit the specific job role;
- Validates the trait profile with the client's HR team; and
- Collects behavioural data of applicants through pymetrics' gamified assessments, and assesses the suitability of the applicants based on the trait profile.

To deal with socially sensitive features and **mitigate the risk of inherent or unintentional bias in the datasets used by the AI model**, pymetrics:

- Uses objective data based on established neuroscience research (e.g. attention to detail, attention span and ability to recall), which are generally stable across gender, racial and age groups.
- Proactively de-biases all AI models to ensure that they are fully representative of the environment that they may function in, so that the AI models do not disadvantage people on the basis of their demographic features:
  - » The standards for fairness are informed by legal requirements. As a pre-hire assessment, pymetrics models must pass a test known as "the four-fifths rule", which is commonly cited in employment law. According to US' Equal Employment Opportunity Commission ("**EEOC**"), the selection rate for any legally protected group must be at least 80% of the selection rate for the majority group. For example, if an employer screens 200 qualified applicants (100 men and 100 women), a model that selects 50 men must also select at least 40 women.
  - » pymetrics will test the AI model against a dataset of users from diverse demographics to ensure that random patterns in the data are not learned by the model and to address any potential for bias.

  - pymetrics would conduct further de-biasing on additional demographic based on geographical relevance or legal requirements. pymetrics uses a bias ratio to compare the proportional pass rates of the highest-passing demographic group with the lowest-passing group for each demographic category (e.g. gender and ethnicity).
  - The AI model would be deployed only if they meet the EEOC standards.

- After AI deployment, pymetrics:
  - Will test the AI model's decisions on real applicants for adverse impact; and
  - Revisits the long-term impact of its system's predictions on retention for the role.

If bias is found either before or after deployment, pymetrics will adjust the AI model to optimise for fairness towards applicants while ensuring the predictive performance of the AI model.

ALGORITHM AND MODEL

3.25 AI systems may have numerous features or functionalities enabled through algorithms in AI models. Measures such as explainability, repeatability, robustness, regular tuning, reproducibility, traceability, and auditability can enhance the transparency of algorithms found in AI models. It may not be feasible or cost-effective to implement even the most essential of these measures for all algorithms.

Organisations are encouraged to take a risk-based approach in making a two-fold assessment. First, identify the subset of features or functionalities that have the greatest impact on stakeholders for which such measures are relevant. Second, identify which of these measures will be most effective in building trust with their stakeholders. Some of these measures like explainability (or repeatability, when using models that are not easily explained), robustness and regular tuning are sufficiently essential that they could, to varying extents, be incorporated as part of the organisation's AI deployment process. Other measures, such as reproducibility, traceability and auditability are more resource-intensive and may be relevant for specific features or in specific scenarios.

Explainability

3.26 Explainability is achieved by explaining how deployed AI models' algorithms function and/or how the decision-making process incorporates model predictions. The purpose of being able to explain predictions made by AI is to build understanding and trust. 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. When an algorithm cannot be explained, understanding and trust can still be built by explaining how predictions play a role in the decision-making process.

3.27 Organisations deploying AI solutions are recommended to adopt the following practices:

a. Model training and selection are necessary for developing an intelligent system (i.e. a system that contains AI technologies). Documenting how the model training and selection processes are conducted, the reasons for which decisions are made, and measures taken to address identified risks will enable the organisation to provide an account of the decisions subsequently.

In this regard, the field of Automated Machine Learning aims to automate a significant portion of machine learning workflows, including feature engineering, feature selection, model selection and hyper-parameter tuning. Organisations using these types of tools can consider the transparency, explainability and traceability of the automated machine learning approach, as well as the models selected.

b. Incorporating descriptions of the solutions' design and expected behaviour into product or service descriptions and system technical specifications documentation demonstrates 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. These steps can help provide greater clarity on an AI model by giving understandable and digestible insights into how the model operates.

Where an organisation's AI system was obtained or procured from a third-party AI solution provider, the organisation can consider requesting assistance from the AI solution provider as they may be better placed to explain how the solution functions.

c. Supplementary explanation tools are helpful for explaining AI models,[6] especially models that are less interpretable (also known as "black box" systems). These tools 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 tools to improve the explainability of an AI model's decision.

3.28 Technical explainability may not always be enlightening, especially to the man on 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") and/or comparisons (such as "these are users with similar profiles to yours that received a similar decision") can be a powerful type of explanation that organisations could consider.

3.29 Nevertheless, there may be scenarios where it might not be practical or reasonable to provide information in relation to an algorithm. For example, disclosing algorithms deployed for anti-money laundering detection, information security, and fraud prevention may allow bad actors to avoid detection; likewise, providing detailed information about proprietary algorithms or the decisions made by the algorithms may expose confidential business information.

[6] These tools are known as "supplementary" as there is at present no single comprehensive technical solution for making AI models explainable. These tools thus play a supplementary role in providing some level of interpretability on an AI model's operation. Examples of these tools include the use of surrogate models, partial dependence plots, global variable importance/interaction, sensitivity analysis, counterfactual explanations, or Self-Explaining and Attention-Based Systems.

Repeatability

3.30 Where explainability cannot practicably be achieved given the state of technology, organisations can consider documenting the repeatability of results produced by the AI model. Repeatability refers to the ability to consistently perform an action or make a decision, given the same scenario. While repeatability (of results) is not equivalent to explainability (of algorithm), some degree of assurance of consistency in performance could provide AI users with a larger degree of confidence. Helpful practices include:

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

b. Performing **counterfactual fairness testing**. Counterfactual fairness testing ensures that a model's decisions are the same in both the real world and in a counterfactual world where attributes deemed sensitive (such as race or gender) are altered;[7]

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;

In this regard, it may be helpful to use AI models that are able to recognise when a given set of facts contains new variables not previously considered and are able to highlight these new variables to a human;

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

[7] James Manyika, Jake Sitberg, and Brittany Presten, "What Do We Do About the Biases in AI?" Harvard Business Review (25 October 2019) <https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai> (accessed 31 October 2019).

Robustness

3.31 Robustness refers to the ability of a computer system to cope with errors during execution and erroneous input, and is assessed by the degree to which a system or component can function correctly in the presence of invalid input or stressful environmental conditions. Ensuring that deployed models are sufficiently robust will contribute towards building trust in the AI system.

3.32 The concept of robustness arises because it is not possible for models to be able to enumerate and set out all preconditions and consequences for an action. This creates the possibility of models producing insensible or unexpected results even with minor modifications to input data (that may not even be perceptible to humans). Testing for robustness can be achieved through scenario-based testing for foreseeable erroneous input.[8] To ensure that models are more robust, organisations can consider working with AI developers to conduct adversarial testing on their models to ensure that their models are able to handle a broader range of unexpected input variables (especially for public-facing AI systems). As this is a resource-intensive exercise, organisations can take a risk-based approach towards identifying the subset of AI-powered features in their products or services that requires adversarial testing.

3.33 No model can be perfectly robust as it is not possible to detect all possible modifications to a set of input data. For this reason, organisations intending to use continual learning (i.e. where the learned parameters of a machine learning model are not fixed but the model continues to change its learned parameters after being deployed into production) are encouraged to be aware of the risks of doing so, should the continual learning model behave in an unpredictable manner.

[8] This is distinct from user acceptance testing ("**UAT**"), which is a process where actual software users test a piece of software to ensure that it can handle required tasks in real-world scenarios, based on specifications. UAT is often a critical step that is taken before a newly-developed software is released to the market.

Regular tuning

3.34 Establishing an internal policy and process to perform **regular model tuning** is effective for ensuring that deployed models cater for changes to customer behaviour over time. This allows organisations 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.

3.35 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 up packages) could 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 that do not reflect actual production environment 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 advised.

Traceability

3.36 An AI model is considered to be traceable if (a) its decisions, and (b) the datasets and processes that yield the AI model's decision (including those of data gathering, data labelling and the algorithms used), are documented in an easily understandable way. The former refers to traceability of AI-augmented decisions, while the latter refers to traceability in model training. Traceability facilitates transparency and explainability, and is also helpful for other reasons. First, the information might also be useful for troubleshooting, or for an investigation into how the model was functioning or why a particular prediction was made. Second, the traceability record (in the form of an audit log) can be a source of input data that can be used as a training dataset in the future.

3.37 Practices that organisations may consider to promote traceability include:

a. Building an **audit trail** to document the model training and AI-augmented decision.

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.[9]

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

3.38 As traceability measures may lead to an accumulation of a large volume of activity data, organisations can consider which of their product features require traceability and which traceability measures might be sufficient for their needs, bearing in mind the resources needed to document the AI model's decisions, datasets and processes. Organisations could assess this based on several factors, including:

a. Their assessment of the probability and/or severity of harm arising from the use of the AI system;

b. The extent to which the AI model had previously been trialled or used; and

c. The regulatory needs of their industry.

[9] It should be noted that a black box recorder does not refer to a "black box" in the AI model sense (i.e. where the decision-making process of an AI model is inherently difficult to interpret and explain).

Reproducibility

3.39 While repeatability refers to the internal repetition of results within one's organisation, reproducibility refers to the ability of an independent verification team to produce the same results using the same AI method based on the documentation made by the organisation. Reproducibility can influence the trustworthiness of the AI product and the organisation deploying the AI model. As implementing reproducibility entails the involvement of external parties, organisations can take a risk-based approach towards identifying the subset of AI-powered features in their products or services that requires external reproducibility testing.

3.40 The following practices contribute towards reproducibility:

a. Testing whether specific contexts or particular conditions would need to be taken into account to ensure reproducibility;

b. Putting in place verification methods to ensure different aspects of the AI model's reliability and reproducibility;

c. Making 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 behaviours;

d. For companies that procure commercial off-the-shelf AI systems, checking with the original AI solution provider about whether the model's results are reproducible; and

e. Adopting points in paragraph 3.30 (c)-(e) under repeatability (namely, assessing how exceptions can be identified and handled, ensuring that exception-handling is in line with organisational policies, and identifying and accounting for changes over time).

Auditability

3.41 Auditability refers to the readiness of an AI system to undergo an assessment of its algorithms, data and design processes. The evaluation of the AI system by internal or external auditors (and the availability of evaluation reports) can contribute to the trustworthiness of the AI system as it demonstrates the responsibility of design and practices and the justifiability of outcomes. It should, however, be noted that auditability does not necessarily entail making information about business models or intellectual property related to the AI system publicly available.

3.42 Implementing auditability not only entails the involvement of external parties but requires disclosure of commercially sensitive information to the auditors, who may be external. Organisations can take a risk-based approach towards identifying the subset of AI-powered features in their products or services for which implementing auditability is necessary, or where implementing auditability is necessary for an organisation to align itself with regulatory requirements or industry practice.

3.43 To facilitate auditability, organisations can consider keeping a comprehensive record of data provenance, procurement, pre-processing, lineage, storage and security. The record could also include qualitative input about data representations, data sufficiency, source integrity, data timelines, data relevance, and unforeseen data issues encountered across the workflow.

3.44 Organisations may also wish to centralise such information digitally in a process log. This would enable the organisation to make available, in one place, information that may assist in demonstrating to concerned parties and affected decision subjects both the responsibility of design and practices and the justifiability of the outcomes of your system’s processing behaviour. Such a log would also enable better organisation of the accessibility and presentation of information yielded, assist in the curation and protection of data that should be kept unavailable from public view, and increase the organisation’s capacity to cater the presentation of results to different tiers of stakeholders with different interests and levels of expertise.

SYMPHONY AYASDIAI:

ILLUSTRATION ON DOCUMENTING MODEL DEVELOPMENT

Symphony AyasdiAI ("**Ayasdi**") offers a solution that helps its clients, mainly in the US banking and finance sector, to build AI models that can adequately forecast revenues and the capital reserve required to absorb losses under stressed economic conditions. Its clients need to prove to the US Federal Reserve that their AI models are accurate and defensible.

The solution – Ayasdi's Model Accelerator (the "**AMA**") – first identifies relevant variables to include in the model and then explains why they are selected. AMA does this by looking at possible relationships encoded within the enriched base-level data, and finding hidden patterns that hold predictive value. The AMA then uses the variables selected to build AI models. It presents a candidate model and several viable challenger models for the clients' selection. Business units within the clients' organisations will evaluate the candidate and challenger models and select those that best represent their business units.

The entire model creation process is documented automatically. The clients can use AMA to institutionalise both their variable selection and modelling methodology, systematically and deterministically, to produce a repeatable process with consistent supporting reports on model lineage, variable selection and cross-validation. This allows the clients to ensure that initial selections of features and models are recorded and documented. At the same time, the entire modelling and approval process is tracked and catalogued, thus facilitating subsequent processes from review to model re-use.

The ability to demonstrate the detailed process of model building and the rigour of evaluating challenger models allows Ayasdi's clients to explain to the US Federal Reserve how their final models are selected.

STAKEHOLDER INTERACTION AND COMMUNICATION

3.45 This section is intended to help organisations take appropriate steps to build trust in the stakeholder relationship strategies when deploying AI.

General disclosure

3.46 Organisations are encouraged to provide general information on whether AI is used in their products and/or services. Where appropriate, this could include information on what AI is, how AI is used in decision-making in relation to consumers, what are its benefits, why your organisation has decided to use AI, how your organisation has taken steps to mitigate risks, and the role and extent that AI plays in the decision-making process. For example, an online portal may inform its users that they are interacting with an AI-powered chatbot and not a human customer service agent.

3.47 Organisations can consider disclosing the manner in which an AI decision may affect an individual consumer, and whether the decision is reversible. For example, an organisation may inform the individuals that their credit ratings may lead to a loan refusal not only from this organisation but also from other similar organisations, while also informing them that such a decision is reversible if individuals can provide more evidence on their credit worthiness.

Policy for explanation

3.48 Organisations are encouraged to develop a policy on what explanations to provide to individuals and when to provide them. Such policies help ensure consistency in communication, and clearly sets out roles and responsibilities of different members of your organisation. These can include explanations on how AI works in an AI-augmented 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. In this regard, the principle of equivalence can provide some guidance such that the same standards of disclosure for human-driven decisions is applied to decisions that have been made or augmented by an AI system.

Bringing explainability and transparency together in a meaningful way

3.49 Appropriate interaction and communication inspire trust and confidence as they build and maintain open relationships between organisations and individuals (including employees). Stakeholder relationship strategies should also not remain static. Companies are encouraged to test, evaluate and review their strategies for effectiveness. Further, the extent and mode of implementation of these factors could vary from scenario to scenario.

3.50 As different stakeholders have different information needs, an organisation can start by first identifying its **audience** (i.e. its external and internal stakeholders). An organisation's external stakeholders may include consumers, regulators, other organisations it does business with, and society at large. Its internal stakeholders may include the organisation's board, management and employees. An organisation can also consider the **purpose** and the **context** of the interaction with its stakeholders. For the purposes of illustration, this Model Framework provides considerations for interacting with consumers and other organisations.

> *As different stakeholders have different information needs, an organisation can start by first identifying its **audience** and considering the **purpose** and the **context** of the interaction.*

Interacting with consumers

3.51 Organisations are encouraged to consider the information needs of consumers as they go through the journey of interacting with AI, from considering whether to use an AI solution, to understanding how the AI solution works as they use it, to requesting for reviews on the decisions made by the AI solution. A typical consumer journey may entail meeting the following information needs of consumers:

a. Making sure that consumers are aware that the products or services that they are considering are AI-enabled. Such information could be provided as part of a general product description.

b. Providing information so that consumers know how the AI-enabled features are expected to behave during normal use. The information could be provided in more detailed descriptions or specifications of product features.

   This, however, may not be necessary for every feature that is AI-enabled. Organisations are encouraged to identify those features where providing additional information in this manner will enhance consumer trust. Similarly, if AI is used in decision-making, information may be provided so that consumers understand how decisions made with the assistance of AI may affect them. This can likewise be provided through descriptions of how the service will be provided.

c. For AI-enabled features that consumers interact with regularly, providing information so that they understand why the AI-enabled feature is behaving in a certain way, and providing preference settings to allow consumers some influence over future behaviour where possible. As doing so requires more engineering effort (such as in the providing additional user interfaces to user history), and the level of information provided may be somewhat more detailed and personalised than feature descriptions, organisations will have to decide which of their product features will benefit from provision of this level of detail.

d. For AI-augmented decisions that affect consumers, consider providing additional information so that they understand why the decisions were made; and for certain categories of such decisions, providing an appropriate channel to contest such decisions. The level of information that is provided will necessarily be detailed but this may not be necessary except for those scenarios where a customer is affected by the decision.

FEATURE:
A:
B:
C:
D:
E:
FEATURE:
A:
B:
FEATURE:
A:

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

Option to opt-out

3.52 Organisations may wish to consider carefully when deciding whether to provide individuals with the option to opt out from the use of the AI product or service, and whether this option should be offered by default or only upon request. Relevant considerations include:

a. Degree of risk/harm to the individuals;

b. Reversibility of the decision made;

c. Availability of alternative decision-making mechanisms;

d. Cost or trade-offs of alternative mechanisms;

e. Complexity and inefficiency of maintaining parallel systems; and

f. Technical feasibility.

3.53 Where an organisation has weighed the factors above and decided not to provide an option to opt out, it is prudent for the organisation to consider providing modes of recourse to the consumer such as providing a channel for reviewing the decision. Where appropriate, organisations may also wish to keep a history of chatbot conversations when facing complaints or seeking recourse from consumers.

Communication channels

3.54 Organisations are encouraged to put in place the following communications channels for their customers:

a. **Feedback channels** This channel could be used for customers to raise feedback or raise queries. It could be managed by an organisation's Data Protection Officer ("**DPO**") if this is appropriate. Where customers 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 stakeholders wish to raise feedback and queries on material inferences made about them.

b. **Decision review channels** Apart from existing review obligations, organisations can consider providing an avenue for individuals (such as an aggrieved consumer) to request a review of material AI decisions that have affected them. Where the effect of a fully-autonomous decision on a consumer may be material, it would be reasonable to provide an opportunity for the decision to be reviewed by a human.

Testing the user interface

3.55 Organisations are encouraged to test user interfaces and address usability problems before deployment, so that the user interface serves its intended purposes. If applicable, organisations are also encouraged to inform individuals that their responses would be used to train the AI system (e.g. a chatbot). Organisations should be aware of the risks of using such responses as some individuals may intentionally use "bad language" or "random replies" which would affect the training of the AI system.

Easy-to-understand communications

3.56 Organisations are encouraged to **communicate in an easy-to-understand manner** to increase transparency. There are existing tools to measure readability, such as the Fry readability graph, the Gunning Fog Index, the Flesch-Kincaid readability tests, etc. It would be helpful for decisions with higher impact to be communicated in an easy-to-understand manner, with the need to be transparent about the technology being used. Besides textual communications, organisations can also consider using visualisation tools, graphical representations, summary tables, or a combination of these. The priority is to convey your information, such as an explanation or interpretation, in a way that is understandable by an organisation's consumers and other stakeholders.

Acceptable user policies

3.57 In certain cases, organisations may be implementing AI-powered solutions that are also trained on real-life input data (i.e. active learning). These organisations may wish to consider setting out certain acceptable user policies ("**AUPs**") to ensure that users do not maliciously introduce input data that unacceptably manipulates the performance and/or results of the solution's model. This is pertinent, given past examples of AI chatbot systems that have been unduly manipulated to issue publicly-unacceptable responses.

3.58 In this regard, AUPs serve to set broad boundaries for the interactions that individuals can perform with the AI system, such as restrictions with regard to intentional actions or attempts to reverse engineer, disable, interfere or disrupt the functionality, integrity or performance of the AI-powered service.

Interacting with other organisations

3.59 Some of the approaches and methodologies described in the preceding section are also relevant when organisations interact with AI solution providers (such as procuring AI solutions and obtaining regulatory approval), or other organisations (such as facilitating industry collaboration, enabling interoperability of systems). Organisations would thus need to obtain sufficient information from AI solution providers to help them meet their business objectives (for example, this could be a back-to-back arrangement for providing the information described in paragraph 3.51). This could be as straightforward as obtaining the AI solution providers' support to provide the information[10] and to build the features[11] necessary such that the deploying organisation can align itself to the Model Framework.

[10] For example, information related to lineage of the training dataset and documenting the key steps in the model training and selection process.

[11] For example, the user-facing interactions providing information about the expected behavior of an AI-powered feature and building the function to allow users to manage preference settings that influence how the AI-powered feature will perform for them in future.

3.60 Organisations may have to consider the level of support and detailed information that they may need to obtain from AI solution providers pertaining to:

a. Data (e.g. types and range of data used in training the algorithm, source and quality of external training data);

b. Model training and selection (e.g. features and variables used and weights of the commercial models supplied, documenting the key decisions made with respect to the model training and selection);

c. Human elements (e.g. nature of human involvement in developing the algorithm, or in the decision-making process);

d. Inferences (e.g. predictions made by the algorithm and how these are incorporated into product features or decision-making);

e. Algorithmic presence (e.g. where in the solution that an algorithm is used); and

f. Measures and safeguards in place to mitigate biases in data and algorithms.

3.61 Depending on the purpose and context, the type and level of detail of information required may be different. For example, a regulator may require a regulated entity to demonstrate that its model development and selection process is sufficiently rigorous, and the AI solution provider may be required to provide more information and be involved in the clarification process with the regulator. An industry collaborator, on the hand, may be more concerned with factors pertaining to compatibility and interoperability.

Ethical evaluation

3.62 Finally, as ethical standards governing the development and use of AI evolve, organisations are encouraged to evaluate whether their AI governance practices and processes are in line with evolving AI standards, and make available the outcome of such evaluations to relevant stakeholders.

FACEBOOK:

ILLUSTRATION FOR STAKEHOLDER INTERACTION AND COMMUNICATION

As a social media and technology company, Facebook is committed to being transparent with the public and users on its operations and services, which includes the use of AI.

In particular, Facebook strives to be **meaningfully transparent with its users** by:

a. **Providing a general disclosure** about Facebook's collection and use of data in an **easy-to-understand manner.** This is achieved through its Terms of Service and Data Policy, accompanied by explanatory and user-friendly videos;

b. Giving users easy-to-use and meaningful control over how their information will be used and shared;

c. Publishing a **policy for explanation** through a series of blogposts to discuss complex subjects, explain the rationale of Facebook's decisions and invite experts to share their opinions. For example, Facebook published a "Hard Questions" blogpost on Face Recognition, which discussed how Facebook used face recognition to help users to tag photos and the controls implemented;

d. Promoting a series of AI educational initiatives and campaigns to help users learn about the technology that underlies the various products and features. For example, Facebook's Artificial Intelligence Research Lab developed and published a series of AI Education videos to explain Machine Learning algorithms and how it is being used.

Facebook currently provides users with a customised News Feed that shows posts that are most relevant to them. The content of the News Feed is determined by the people and pages a user chooses to friend and follow.

As part of Facebook's consultation on its News Feed feature, Facebook took into consideration:

a. The need to be transparent and provide more information about the algorithms behind the feature;

b. The type of information that would be most valuable to users. For example, Facebook included examples of people's interactions that contributed to posts on News Feed; and

c. The need for users to take control or manage its News Feed.

With this, Facebook put in place the following to **build trust with users**:

a. Implement a "Why am I seeing this post?" feature to explain how users' past interactions impacted the ranking of posts in the News Feed. Specifically, users were able to learn:

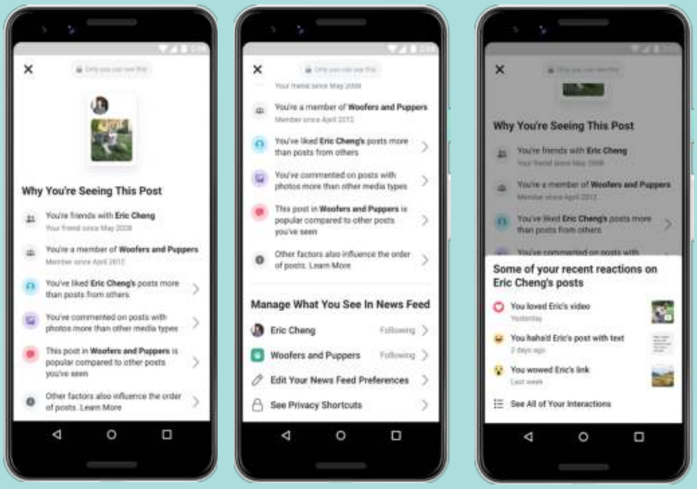
   i. The reason for viewing a certain post in the News Feed. For example, the post could be from a friend or Group or Page that the user has followed.

   ii. Information that had the largest influence over the order of posts, including: (a) the frequency in which the user interacts with posts from people, Pages or Groups; (b) the frequency in which the user interacts with a specific type of post (e.g. videos, photos or links); and (c) the popularity of the posts shared by the people, Pages and Groups that the user follows.

   iii. Shortcuts to controls that help users personalise its News Feed such as See First, Unfollow, News Feed Preferences and Privacy Shortcuts.

b. Publish a series of "News Feed FYI" blog posts that highlighted and explained the rationale for key updates to News Feed.

c. Launch a new "Inside Feed" website that provided greater detail on how Facebook's systems worked and the way Facebook evaluated the changes.



Only you can see this

Why You're Seeing This Post

You're friends with **Eric Cheng**
Your friend since May 2008

You're a member of **Woofers and Puppers**
Member since April 2012

You've liked **Eric Cheng's** posts more than posts from others

You've commented on posts with photos more than other media types

This post in **Woofers and Puppers** is popular compared to other posts you've seen

Other factors also influence the order of posts. Learn More

Your friend since May 2008

You're a member of **Woofers and Puppers**
Member since April 2012

You've liked **Eric Cheng's** posts more than posts from others

You've commented on posts with photos more than other media types

This post in **Woofers and Puppers** is popular compared to other posts you've seen

Other factors also influence the order of posts. Learn More

Manage What You See In News Feed

Eric Cheng Following

Woofers and Puppers Following

Edit Your News Feed Preferences

See Privacy Shortcuts

Why You're Seeing This Post

You're friends with **Eric Cheng**
Your friend since May 2008

You're a member of **Woofers and Puppers**
Member since April 2012

You've liked **Eric Cheng's** posts more than posts from others

You've commented on posts with

Some of your recent reactions on Eric Cheng's posts

You loved Eric's video
Yesterday

You haha'd Eric's post with text
2 days ago

You wowed Eric's link
Last week

See All of Your Interactions

MSD:

ILLUSTRATION FOR STAKEHOLDER INTERACTION AND COMMUNICATION

MSD is a multinational pharmaceutical company that deploys an in-house chatbot, Jennie, to answer queries on IT-related matters.

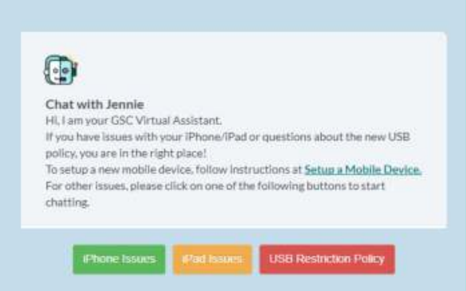
Prior to deployment, MSD's User Experience ("**UX**") team tested the **human-AI interface** and addressed usability problems to ensure optimal user interaction with Jennie. In particular, three tenets guided the development and deployment of Jennie:

a. **Understanding a user's mental model**: The UX team conducted user research with representative users to understand users' expectations when interacting with a chatbot. The research covered the scope of IT questions, expected answers and the kinds of answers provided (e.g. how technical the answers should be phrased). By building a user-friendly interface, users would be more comfortable to interact and use the chatbot.

b. **Taking a human-centric approach**: To understand patterns of human behaviour, the team analysed how its employees reacted when faced with challenges in interacting with the chatbot. Examples include how users formulate questions, what types of answers satisfy users and how many times a chatbot should be allowed to attempt an answer. After understanding the human interaction touchpoints, the team used these insights to create an information flow architecture to deliver a better experience for users.

c. **Managing the bot-human handover**: There could be instances where Jennie might not be able to provide a satisfactory answer. In such instances, the UX team determined that the chatbot would have a maximum of three attempts to provide a satisfactory reply, before forwarding the user's request and chat logs to a customer care executive for follow-up.

When employees engage Jennie, MSD will **disclose** on the landing page that Jennie is AI-powered and is a beta version which will improve over time (see figure below).



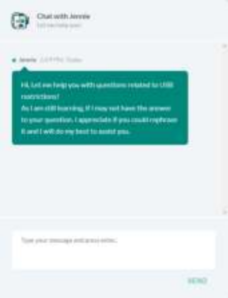
Chat with Jennie

Hi, I am your GSC Virtual Assistant.

If you have issues with your iPhone/iPad or questions about the new USB policy, you are in the right place!

To setup a new mobile device, follow instructions at Setup a Mobile Device.

For other issues, please click on one of the following buttons to start chatting.



Chat with Jennie

Jennie [illegible]

Hi, Let me help you with questions related to USB restrictions!

As I am still learning, if I may not have the answer to your question. I appreciate if you could rephrase it and I will do my best to assist you.

ANNEX A

FOR REFERENCE: A COMPILATION OF EXISTING AI ETHICAL PRINCIPLES

This annex comprises a collection of foundational AI ethical principles, distilled from various sources.[12] **Not all are included or addressed in the Model Framework.** Organisations may consider incorporating these principles into their own corporate principles, where relevant and desired.

1. **Accountability:** Ensure that AI actors are responsible and accountable for the proper functioning of AI systems and for the respect of AI ethics and principles, based on their roles, the context, and consistency with the state of art.

2. **Accuracy:** 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.

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

4. **Explainability:** 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. **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.

[12] These include the 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). There is also the European Commission's *Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions - Building Trust in Human-Centric Artificial Intelligence* (https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58496), and the OECD's Recommendation of the Council on Artificial Intelligence (https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449). They also include principles raised through consultation feedback from the industry.

6. **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.

   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 judgements 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.

   g. To ensure that the overall well-being of the user should be central to the AI system’s functionality.

7. **Human rights alignment:** Ensure that the design, development and implementation of technologies do not infringe internationally recognised human rights.

8. **Inclusivity:** Ensure that AI is accessible to all.

9. **Progressiveness:** Favour implementations where the value created is materially better than not engaging in that project.

10. **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 stakeholders to verify how and when AI technology is being applied.

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

11. **Robustness and Security**: AI systems should be safe and secure, not vulnerable to tampering or compromising the data they are trained on.

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