DISCUSSION PAPER ON ARTIFICIAL INTELLIGENCE (AI) AND PERSONAL DATA – FOSTERING RESPONSIBLE DEVELOPMENT AND ADOPTION OF AI

Published 5 June 2018

1. INTRODUCTION

Data is the basic building block of the digital economy. The exponential growth in data volume and increasing computational power at decreasing cost work in tandem to promote data-driven technologies such as Artificial Intelligence (AI). The benefits and risks of AI have been the subject of great public debate. On the one hand, AI has the ability to boost productivity, transform businesses, grow the economy and enhance people's lives. On the other hand, AI may displace jobs and pose ethical challenges such as social profiling.

This paper presents the Singapore Personal Data Protection Commission (PDPC)'s preliminary analysis of some of the issues pertinent to the commercial development and adoption of AI solutions. The objective is to propose an accountability-based framework for discussing ethical, governance and consumer protection issues related to the commercial deployment of AI in a systematic and structured manner. In a services-driven economy like Singapore, AI will likely be deployed in intelligent systems that process personal data. Hence, this framework is also relevant to personal data protection. Using this framework in the design of systems or processes could also encourage "data protection by design."[1]

The proposed framework aims to encourage informed and constructive debate around this complex issue. Ultimately, we hope it sow the seeds for the private sector to develop voluntary governance frameworks, including voluntary codes of practice that can applied to organisations, sectors, or more generally across the digital economy.

(a) Striking the Right Balance in AI Governance

The nascence of AI development today presents a timely opportunity for key stakeholders such as regulators, AI developers, AI user companies and consumers to discuss the need for AI governance and regulations, as well as the form they might take. Key preliminary views include:

- **Governance frameworks around AI should be technology-neutral and "light-touch"** so that AI technology can develop in a direction that is not hindered or distorted by prescriptive rules that are laid down prematurely.

- **AI developers and user companies should be provided with regulatory clarity when developing AI technologies and translating them into AI solutions**. Consumers benefit from choice and product differentiation arising directly from

[1] Data Protection by Design refers to the approach by which organisations consider the protection of personal data from the earliest possible design stage, and throughout the operational lifecycle, of a new system, product or service. This way, the appropriate safeguards to protect personal data would have been embedded within. Extracted from the *Guide to Data Protection Impact Assessments* by the Singapore Personal Data Protection Commission.

the diffusion of AI technology into the marketplace. Regulatory clarity also contributes to healthy market competition.

- **Policies and regulations that promote explainability, transparency and fairness, as well as human-centricity, as clear baseline requirements can build consumer trust in AI deployments.** For example, explaining how AI-enabled decision-making can lead to more consistent decisions while providing more transparency in the decision-making process can increase consumer confidence. In addition to the Personal Data Protection Act (PDPA), sector-specific codes of practice can provide assurance to consumers about the use of AI, especially when applied in decision-making processes.

(b) The AI Value Chain and Deployment Process

This paper adopts the following model to describe the different stakeholders in the AI value chain:

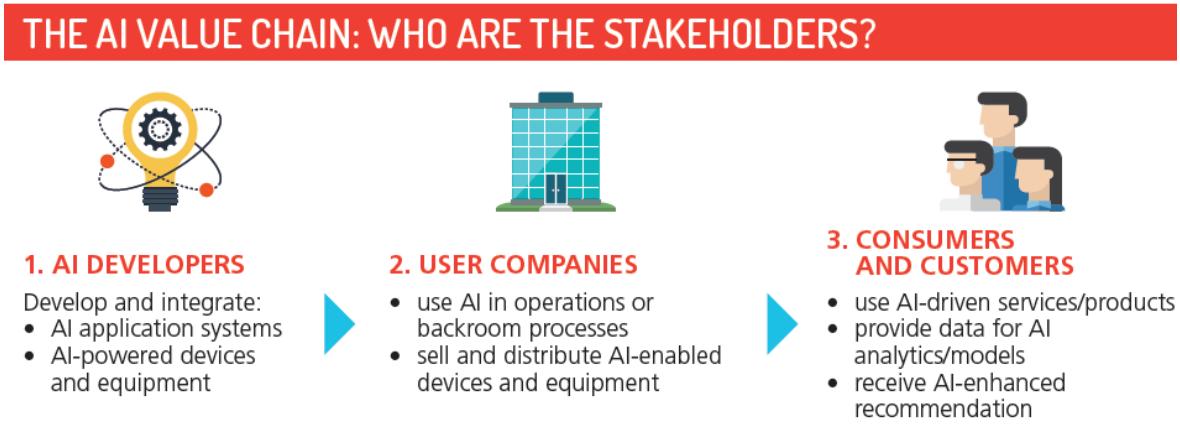


Fig. 1

The term "AI Developers" includes developers of application systems that make use of AI technology. These may be commercial off-the-shelf products, online services, mobile apps and other software that consumers can use directly. "AI Developers" also include device and equipment manufacturers that integrate AI-powered features into their products, as well as AI solution providers whose solutions are not stand-alone products but are meant to be integrated into a final product.

Meanwhile, the term "User Companies" refers to companies that make use of AI solutions in their operations. This could be a backroom operation (e.g. processing applications for loans) or a front-of-house service (e.g. e-commerce portal or ride-hailing app). Equally, it can refer to companies that sell or distribute devices or equipment that provide AI-powered features (e.g. smart home appliances).

This paper adopts the following process model to describe the different phases in an AI deployment:

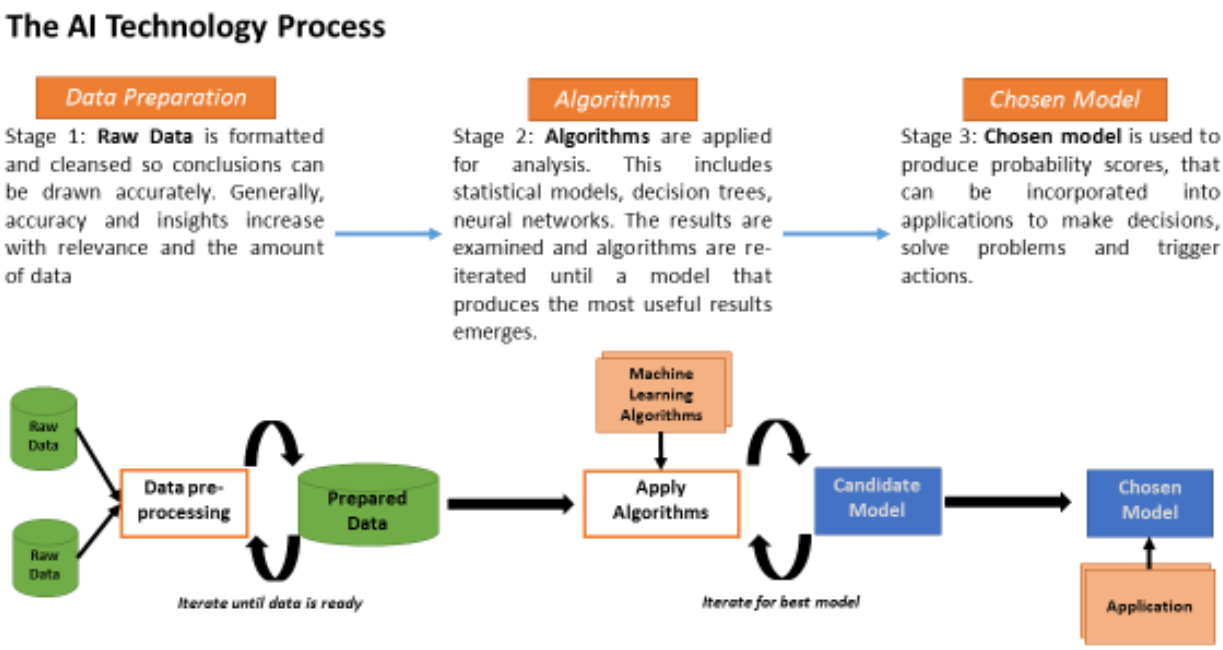


Fig. 2

As different stakeholders have different roles in an AI value chain, the relevance and applicability of the AI governance framework to each stakeholder may also be different. For example, the deployment process is potentially applicable to AI Developers that integrate a machine-learning AI in their products. Equally, a User Company that makes use of commercial off-the-shelf systems that employ supervised machine learning AI can also rely on this model to better understand the risks (e.g. of unintended bias hidden in the training dataset) and the need for ongoing maintenance (e.g. model tuning). Hence, it is necessary to consider both the AI value chain and the technology deployment process in discussing the development of the AI governance framework.

2. PRINCIPLES FOR RESPONSIBLE AI

In order for AI to benefit businesses and society at large, a set of principles needs to be incorporated into the AI governance framework. These principles aim to promote trust and understanding in the use of AI technologies. During PDPC's consultation, two main sets of principles surfaced:

(i) **Decisions made by or with the assistance of AI should be explainable, transparent and fair** so that affected individuals will have trust and confidence in these decisions**.**

- **Explainable**: How can automated algorithmic decisions and the data that drives such decisions be explained to end-users and other stakeholders in non-technical terms? The features that support explainability could possibly be designed into either the intelligent systems that deploy AI engines or in the AI engines proper. If how the AI engine works cannot be easily explained, can the ability to verify the automated algorithmic decision be equally effective?

- **Transparent:** AI developers, data scientists, application builders and user companies should be accountable for the AI algorithms, systems, applications and resultant decisions respectively in order to build trust in the entire AI ecosystem. What are the measures and processes that stakeholders in the different parts of the value chain can incorporate in order to be able to inform consumers or customers about how and when AI technology is applied in decisions affecting them?

- **Fair**: AI algorithms and models embedded in decision-making systems should incorporate fairness at their core. This could include the training dataset, AI engine and selection of model(s) for deployment in the intelligent system. What practices will avoid unintentional discrimination in automated algorithmic decisions? Examples include monitoring decisions to detect unintentional discrimination and accounting for how they were made.

(ii) **AI systems, robots and decisions made using AI should be human-centric.** Human-centric design refers to the design approach that puts the individual customer or consumer front and centre of the design of the AI deployment. Organisations that are perceived to have caused harm to consumers as a result of their AI deployments do not inspire consumer trust and confidence. **Beneficence or "Do no harm"** is a principle that can incorporate the following:

- Decisions should strive to confer a benefit on or provide individuals with assistance in the performance of a task;
- Decisions should not cause foreseeable harm[2] to an individual, or should at least minimise harm (in necessary circumstances, when weighed against the greater good);[3]
- Tangible benefits to individuals should be identified and communicated in order to build consumer understanding and confidence; and
- AI systems and robots should be designed to avoid causing bodily harm or affecting the safety of individuals.

[2] "Harm" in this case includes physical, psychological, emotional and economic harm.
[3] Adapted from UNICEF's Humanitarian Principles.

Illustrations

- Automating the decision to approve an application for travel insurance to promote consistency while catering for genuine differences in individual circumstances.
- Programming robot-assisted manufacturing such that the robotic arm will not swing beyond a specific safety parameter or comes to a halt if someone steps into its operational zone.
- Including safety limits by design in the operation of self-driving vehicles (e.g. speed limits, vehicle-to-vehicle safety limits for collision avoidance). Where collisions are unavoidable, parameters should also be included to avoid or minimise harm to humans.

3. EXPLORING A PROPOSED GOVERNANCE FRAMEWORK FOR AI

Building on the AI Value Chain (*Fig. 1*) and the aforementioned principles, this paper proposes a governance framework for AI to encourage discussion around how the two sets of principles could be adopted by different stakeholders. The proposed framework is generally applicable to all sectors and provides options that may be adopted (fully or in part) by organisations, depending on their needs and objectives.

Assumptions and Limits of the Proposed Framework

In exploring the proposed framework, the following assumptions and limits were identified:

Assumptions	**Limits**
• **Technology Neutral**: The proposed framework focuses on the design, application and use of technology in contexts affecting individuals without being specific to the AI technology. • **Sector-agnostic**: The proposed framework should be applicable to all sectors as a baseline standard. This does not preclude specific sectors and organisations to incorporate additional standards above the baseline set out in the proposed framework.	• **Legal Liability**: The proposed framework does not set out to address or resolve specific questions of legal liability or apportionment of damages or restitution. However, some of the practices advocated in this framework are likely to assist in the management of disputes and ensure the availability of evidence that may be required to resolve such questions.

4. PROPOSED FOUR-STAGE GOVERNANCE FRAMEWORK

(A) IDENTIFYING THE OBJECTIVES OF AN AI GOVERNANCE FRAMEWORK

This section sets out several objectives for the proposed framework.

Proposed Objectives:

Explainability and Verifiability. An organisation that employs AI in its decision-making process should be able to explain how its AI engine functions. However, where this is not possible for certain types of AI engines (e.g. neural networks), the organisation should minimally be able to verify that the AI engine is performing to expectations and within the technical and ethical parameters set. This would provide reassurance that the decision-making process is supervised and not overly reliant on AI to suggest decisions or even simply delegated to a set of software codes.

To achieve explainable AI, it is necessary to consider the roles of different stakeholders in the AI value chain, for example:

- The requirement of explainability may be catered for by AI Developers in their design of AI engines or solutions. This will enable them to be in a better position to explain to User Companies how their AI solutions function.
- User Companies who are unable to explain how the AI engine functions can design for verifiability[4] of the decision-making process from the planning stages. This will enable them to ensure that the necessary data points for monitoring are catered for.

Good Data Accountability Practices. Organisations should put in place good data accountability practices. These include the following:

- **Understanding the lineage of data.** This means knowing where the data originally came from, how it was collected, curated and moved within an organisation, and how its accuracy is maintained over time.
- **Minimising the risk of bias.** Veracity or data quality refers to the risk of bias that may be inherent or latent within a dataset. Organisations should adopt practices that enable them to detect biases that may exist in their data so that they can take steps to address them.
- **Maintaining data provenance records.** This practice is important for establishing data lineage in general, but separate provenance records can also

[4] Verification methods for AI deployments can as a baseline reference traditional software verification and validation methods. These include testing, run-time monitoring, static analysis, model checking and theorem proving, and can be modified for the AI context. Human oversight is usually a core component of such verification processes. Extracted from Menzies, T. and Pecheur, C. (2005) "Verification and Validation and Artificial Intelligence", *Advances in Computers*, 65, 153 – 201.

be maintained to log the data that was used in the AI deployment process model and also the entire AI value chain[5].

Collectively, good data accountability practices provide reassurance to consumers that downstream decisions or suggestions provided by intelligent systems are not false (resulting in type I and/or type II errors[6]) and do not risk unintentional discrimination. It is also important for User Companies to distinguish between data accountability, which aims to ensure completeness and comprehensiveness in the data preparation and model-creation stages, and the responsibility of organisations to avoid making discriminatory, unfair or unlawful decisions.

Transparency. Open and transparent communication between stakeholders in the AI value chain will be conducive to building trust in the entire AI ecosystem. Examples of how this principle may be implemented include:

- Provision of information by different stakeholders should have a clear purpose, tailored to the intended recipient's interests and needs, and should not be inadequate; and
- Stakeholders are encouraged to explore and use a variety of communication channels in order to ensure effective and clear communication with consumers and customers.

(B) SELECTING APPROPRIATE ORGANISATIONAL GOVERNANCE MEASURES

This section identifies accountability-based practices that can help organisations render an account to a regulator, affected individual or interested stakeholder of how decisions are made. Not all elements of this section are relevant in all cases: options may be selected and customised according to the AI engine that is adopted (e.g. rule-based, machine learning), the sector that the organisation operates in, the type of decisions and degree of automation (e.g. man-in-the-loop or man-out-of-the-loop), etc.

Governance	A. **Internal governance.** When AI systems are used for decision-making, organisations should consider how their existing corporate governance or oversight mechanisms could be adapted or new ones created. For example: • Ensuring that the departments undertaking AI activities are **aware of their responsibilities**; • Introducing **oversight mechanisms** for actions or decisions within the responsible departments' sphere of responsibilities, e.g. to review exceptions identified by the automated decision-making process, ensure verifiability of automated decisions, or to review decisions in processes that have no human oversight elements;

[5] Maintaining provenance records from data used to build models to the AI end-user input data could provide a way to ascertain the quality of data used and trace back potential sources of errors.

[6] In statistical analysis, type I errors refer to false positive errors and type II errors refer to false negative errors.

	• Establishing **monitoring or reporting** systems to ensure that information flows to the right level within the corporate governance hierarchy; and • **Periodically reviewing** the corporate governance or departmental oversight mechanisms, especially when there are significant changes to the organisational structure or key personnel involved in the governance or oversight mechanisms, to ensure their relevance.
	B. **Risk and/or harm mitigation.** When deploying intelligent systems for real world uses, it is important to identify potential risk or harm that may foreseeably arise in anticipated use-cases. A **risk and impact assessment** is a tool that can assist with risk identification and harm mitigation, as follows: • When **selecting candidate models** for AI deployments, risk and impact assessments can assist in identifying and understanding the expected and worst case implications, and inform the crafting of mitigation processes; • **Ethical considerations** should also be incorporated into the overall risk and impact assessments; and • **Documenting** the risk and impact assessments can be helpful, e.g. to produce audit trails for internal (or external) accountability.
	C. ***Ad hoc* and periodic reviews of AI deployments and decisions.** After initial deployment, organisations should consider periodically reviewing their decisions and processes to be satisfied that reliance on AI systems for decision-making remains relevant and appropriate. In addition to periodic review, specific triggers for reviews should also be defined, e.g. when there are significant changes to the intelligent system or deployed models.
Operations Management & Systems Design	D. **Data Accountability.** AI models deployed in intelligent systems that make or assist the making of algorithmic decisions must operate on accurate information. Accuracy of information is affected by: • The **completeness** of the data required; • How **recently** the data was collected and updated; • Whether the data is structured in a **machine-understandable** form; and • The **source** of the data, as the context for initial collection may affect the interpretation and reliance of the data for a secondary purpose. Organisations should consider generating records or documenting processes to mitigate potential issues with data accountability, including: • Keeping a **data provenance record** or audit trail of the data that was used in the model creation or decision making, which may help uncover any inherent limits or errors;

- AI Developers can consider assigning a **veracity score**[7] to training datasets during model creation that may be helpful during model selection to minimise the risk of bias; and
- User Companies can consider assigning **secondary veracity scores** for operational (or input) data which may be helpful when it is used as feedback data during model tuning.

E. **Repeatability.** An intelligent system that is able to perform an action or make a decision consistently within the same scenario will promote consumer confidence. Practices that may be helpful include:
- **Repeatability assessments** for commercial deployments in live environments;
- Where a decision is not repeatable, one possible design consideration is how **exceptions** should be identified and handled; and
- Measures can also be put in place to identify and account for **changes over time**, especially if models are trained on time-sensitive data or are designed to evolve.

F. **Traceability** relates to how the AI module makes decisions or provides suggestions. Practices that promote traceability include:
- Building an **audit trail** to document the decision-making process;
- Implementing a **black box recorder** that captures all input data streams. Data relevant to traceability should be stored appropriately to ensure that there is no degradation or alteration, and for retention durations relevant to the industry; and
- **Access to and/or audits of the AI algorithm,** for organisational risk management. Regardless of whether algorithm audits should be universally practised, if executed well, algorithmic audits can foster consumer trust. In considering algorithm audits, the following matters are relevant:
  - **Expertise** required to effectively understand the algorithm, rules or models;
  - Considerations of **commercial confidentiality** of the AI technology provider, including mitigating measures; and
  - The **usefulness** of this information to assist in providing an account to regulators, affected individuals and/or interested stakeholders.

G. **Tuning of AI Models.** The selection of model(s) for eventual deployment into the intelligent system should be a considered decision and the process and reasons for the choice should be documented. Moreover, models need to be periodically updated. Relevant considerations include:

[7] In data science, data veracity refers to false or inaccurate data. A data veracity score refers to the assignment of a score to account for elements that could affect the accuracy of data, e.g. inconsistencies, contradictions, or "staleness" of data.

	• Adopting internal governance processes and **tuning AI models periodically** to cater for changes to data and/or models over time; • Carry out active monitoring and tuning where AI systems developed in a comparatively static environment[8] display model instability when deployed in dynamic environments.[9]

(C) CONSIDERING CONSUMER RELATIONSHIP MANAGEMENT PROCESSES

The third step in the AI governance framework is the management of communications with affected individuals and providing measures for recourse, which are important for building consumer trust and confidence. The following measures can be incorporated into existing consumer management processes. Depending on the nature of the AI deployment, organisations may select measures that would best suit their needs. These measures include:

Transparency	A. **Policy for disclosure.** Organisations should consider disclosing the use of AI in the decision-making process (whether fully automated or to assist in decision making), including how this disclosure can be made. **Increased transparency** would contribute to building consumer confidence and acceptance by increasing the level of knowledge and trust in the customer relationship. Organisations should also consider carrying out **ethical evaluations** and making meaningful summaries of these evaluations available to their customers.
	B. **Policy for explanation.** Organisations should consider explaining how AI is deployed in the decision-making process, and/or how a specific decision was made including the reasons underpinning the decision where appropriate or available. Explanations could take the form of *ex ante* information provided as part of general communication by the organisation, or of specific information provided by the organisation in respect of a decision affecting the individual making the request. In the context of decisions made using profile information, the affected individual may be given information about how his personal data is associated with user targeting profiles.

[8] Stoica, I. et al. (2017) *A Berkeley View of Systems Challenges for AI.* and Mishra, N. et al. (2018) *Controlling AI Engines in Dynamic Environments*.

[9] These are environments that change rapidly, frequently and in non-reproducible ways.

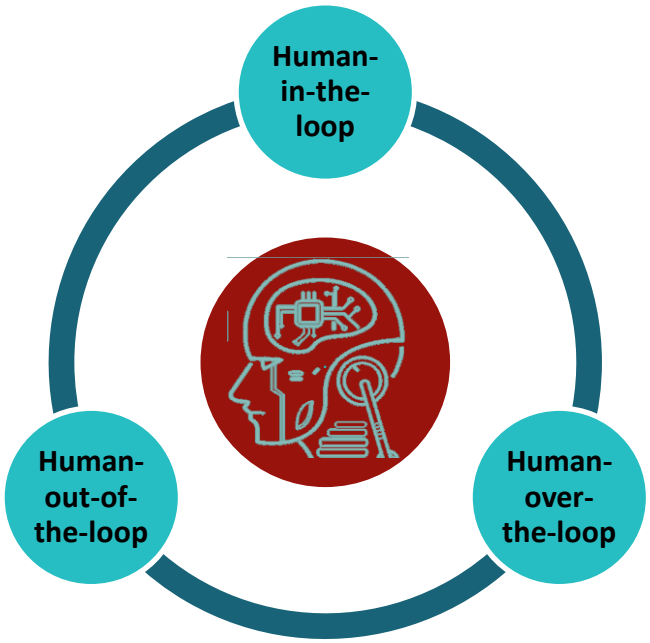
Interaction	C. **Heuristic evaluation.** This ensures that usability problems are addressed, and that user interfaces are tested, so that the user interface layer serves its intended purpose. Expectations of consumers will be a relevant area to address to preserve the consumer experience. For example, with the increasing use of chat bots, organisations should consider whether consumers should be informed when they interact with chat bots instead of a human agent.
	D. **Option to opt out.** In a fully automated deployment scenario, an organisation providing an option to opt out could lead to a decrease in operational efficiency, may not be technically feasible, or lead to a process being commercially uncompetitive. Nonetheless, providing an option to opt out could be beneficial, particularly if it builds consumer trust, for example in deployments that could have significant impact to the individual and risk to the organisation. Hence, organisations should consider providing an option for opt out.
Communication	E. **Feedback channel**. Organisations should consider providing a channel (e.g. email address) for affected individuals to provide any feedback or raise any queries they may have for the organisation to address. In particular, where customers find inaccuracies in the data, a channel that allows customers to access and correct their own data will be useful to maintain data veracity.
	F. **Review of decision.** Organisations should consider providing the affected individual with an avenue to request a review of a decision affecting him. Relevant considerations include the degree of automation and the extent to which AI is deployed in decision-making, as well as the impact to the individual. In cases where a fully automated decision is made, it could be reasonable to provide an avenue for the affected individual to request a review of the decision.

(D) BUILDING A DECISION MAKING AND RISK ASSESSMENT FRAMEWORK

The final stage is to incorporate decision-making and risk assessment considerations into the framework. The risk and severity of harm to the customer are factors that affect which decision-making approach should be adopted, and in turn how organisations calibrate governance and consumer management processes. The following decision-making approaches can assist businesses in determining the appropriate method of deployment of AI by maximising benefits while minimising the risks of harm:[10]

[10] Adapted from Citron, D. K. and Pasquale, F. A. (2014) "The Scored Society: Due Process for Automated Predictions", *Washington Law Review*, 89

Human-in-the-loop models involve a human decision maker who relies on the intelligent system to suggest one or more possible options, but who ultimately makes the final decision. For example, an operational system that provides an employee with one or more options customised for the case that he is handling.



Human-out-of-the-loop models usually involve automated decision making by the intelligent system based on a pre-determined set of scenarios. The identification and handling of exceptions will be important. For example, an autonomous cleaning bot can be left to map out the best path for cleaning a location, excluding "no-go zones" which humans can pre-set.

Human-over-the-loop models involve a human decision maker who has made a choice but relies on the intelligent systems to suggest options of how to perform the action. For example, an individual specifies his destination in a navigation system which makes suggestions of one or more navigation routes.

Fig. 3

To determine the appropriate decision-making approach, organisations can consider relying on the proposed decision matrix (see *Fig. 4* below), which takes into account the probability and severity of harm to consumers.

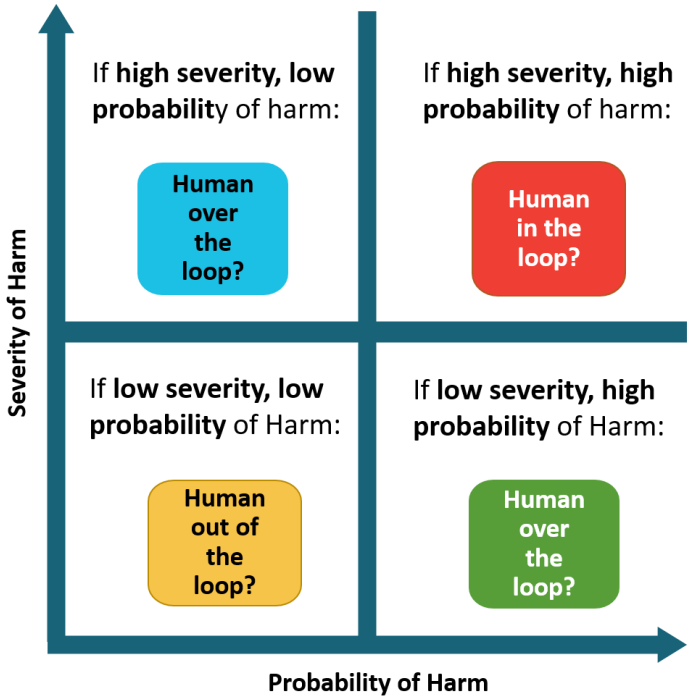


Fig. 4 Illustration of how to use the decision matrix

To illustrate how this proposed framework may be used, where an organisation assesses that both the probability and severity of harm to a customer is high, it may decide that the human-in-the-loop model may be the appropriate decision-making approach.[11] Accordingly, given the risk of harm, it is appropriate for the organisation to decide to also implement governance processes focused on repeatability and traceability. Additionally, the organisation may decide that since the algorithmic decision is not automated, there is no need to provide too much information about its internal processes to its customers.

5. NEXT STEPS

This discussion paper is intended to promote healthy discussion on promoting the responsible development and adoption of AI solution and mitigating potential risks and negative impact. PDPC invites organisations to use this document for internal discussion. Organisations are free to adapt it for internal use. Trade associations and chambers, professional bodies and societies, and interest groups are encouraged to adapt this proposed framework for their sectors in the form of voluntary codes of practice.

[11] However, where an organisation chooses to use human-in-the-loop model, it does not necessarily mean that the probability and severity of harm is high.

6. ACKNOWLEDGEMENTS

PDPC expresses its sincere appreciation to the following organisations and individuals for their valuable feedback to this discussion paper: *(in alphabetical order)*

AI Singapore
Attorney-General's Chambers
Centre for Strategic Futures
Competition and Consumer Commission of Singapore
Government Technology Agency
Infocomm Media Development Authority
Integrated Health Information Systems Pte Ltd
Inter-Agency Project Team on Ethics and Governance of Artificial Intelligence
Land Transport Authority
Ministry of Communications and Information
Ministry of Health
Monetary Authority of Singapore
Smart Nation and Digital Government Office

Adobe Systems, Inc.
Amazon Web Services
AsiaDPO
BSA | The Software Alliance
Cisco Systems, Inc.
Facebook
Microsoft
Symantec Corporation
Dr. Steven Tucker, Founder, Tucker Medical Pte Ltd and Chief Medical Officer, CXA Group

END OF DOCUMENT