Response to the PDPC's "A Proposed Model AI Governance Framework"

Non-Profit Working Group on AI
Singapore, 26 June 2019

Dear Personal Data Protection Commission (PDPC),

We are the Non-Profit Working Group on AI, a group comprising members of DataKind SG, Effective Altruism SG, AI researchers, data scientists, academics and more.[1] We welcomed the PDPC's release of "A Proposed Model AI Governance Framework" on 23 January 2019, and appreciate the PDPC's commitment to improving the Model Framework by consulting with the wider public. We are responding to PDPC's call for feedback on the Model Framework.

Having studied the Model Framework very closely, we agree on the necessity of providing baseline guidance on how companies should internally govern their development and use of AI. This is important to help companies institute proper safeguards in their AI systems that forestall the myriad possible harms to society, such as discrimination or the loss of autonomy. We would like to help strengthen the Model Framework and help it to achieve its full potential by making some recommendations. These recommendations arose through careful analysis of the Model Framework and regular meetings and discussions within our group, over a period of three months. We drew on our experience working with AI and the societal complexities at the interface of software and society, knowledge of industry practices surrounding software and AI, as well as our knowledge of AI and software policy frameworks from around the world. We have also fleshed out our suggestions using examples where the situations discussed have been borne out in real life.

We were also informed by the burgeoning literature on AI policy, ethics, safety, and related issues. In particular, we have drawn inspiration from two other guiding documents: a paper from the Monetary Authority of Singapore (MAS) titled "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector"[2] (henceforth the *"MAS FEAT Principles"*), as well as the recent OECD Recommendation of the Council on Artificial Intelligence[3] (henceforth the *"OECD AI Recommendations"*).

The main themes of our recommendations and areas for future research are presented below, followed by an annotated copy of the Model Framework with specific recommendations under each paragraph that we have addressed. We have also written comments after the "use case" of UCARE.AI. We hope the PDPC derives value from our suggestions and examples, and incorporates our feedback into future iterations of the Model Framework. In the interest of public accountability, we are also making this response available at https://npwg-ai-sg.github.io/.

26 June 2019
Non-Profit Working Group on AI

[1] A list of contributors to this document is available on the "Contributors" section on the final page.
[2] MAS (2019), "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector". Retrieved on 16 June 2019 from http://www.mas.gov.sg/News-and-Publications/Monographs-and-Information-Papers/2018/FEAT.aspx
[3] OECD (2019), *Recommendation of the Council on Artificial Intelligence*. Retrieved on 16 June 2019 from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Major Themes from our Recommendations

Clarifying and broadening the scope of AI governance

The Model Framework currently proposes a broad definition of Artificial Intelligence (AI), and sets out to develop guidance that is both algorithm-agnostic and technology-agnostic. We commend and agree with this vision. However, there are a few areas where we believe the Model Framework could better live up to this intent. Firstly, the framework's current definition of AI, while appropriately broad, does not emphasize the policy-relevant features of AI. In combination with the goal of algorithm-agnosticity, this makes it difficult to determine which technologies fall under the framework. To address this, we have proposed a revised definition, consistent with the criteria suggested by Krafft *et al* that a policy definition of AI "should include both current and future applications of AI, be accessible to laypersons, and be implementable in policy through reporting and oversight."[4] By clarifying the scope in this way, we hope to provide better conceptual grounding for when and whether AI governance is relevant to a particular technology.

Secondly, the Model Framework tends to assume several features of AI that are not shared by all AI technologies. In particular, the "Determining AI Decision-Making Model" section focuses on AI that "makes decisions", while the "Operations Management" section is geared toward Machine Learning (ML) and other data-driven AI. However, many common uses of AI in industry fall outside one or both of these categories, such as expert systems (not data-driven) or content generation (not decision-making). Our response contains recommendations for broadening the scope of these sections, and adding suitable guidelines for non-data-driven AI, to help the entire Model Framework live up to its broad vision.

Improving the explanation of harms and risks

We applaud the recognition by the PDPC of the risk of harm that the widespread use of AI could pose to society, if not governed carefully. The avoidance of such harms, and mitigation of associated risks, are fitting as foundational motivations for the Model Framework. However, many of the harms and risks can be counterintuitive and unexpected, as they can arise subtly and insidiously through the simplest of oversights in well-intentioned AI implementations. Even the most experienced companies have found it difficult to anticipate or avoid these harms. Hence we advocate for a more extended and comprehensive discussion specially dedicated for harms and risks, informed by the vast amount of relevant literature. Moreover, the Model Framework has left out an important class of harms, relating to discriminatory representations of human identity.[5]

---

[4] Peter Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo (2019), "Policy versus Practice: Conceptions of Artificial Intelligence". Preprint under review, retrieved on 16 June 2019 from http://people.csail.mit.edu/pkrafft/papers/critplat-policy-vs-practice.pdf

[5] See the response to paragraph 2.1 for further discussion of such *harms of representation*.

Improving the explanation of guiding principles

The guiding principles of *explainability*, *fairness*, *transparency*, and *human-centricity* are defined very briefly in paragraphs 2.5a-b. Although there are further elaborations deep in the body of the Model Framework, we advocate for dedicated paragraphs that elaborate on each guiding principle to make them concrete, meaningful and measurable, with examples where appropriate. We view this as necessary to properly ground internal governance, and to allow the PDPC to monitor the compliance performance of organizations that deploy AI, for example, by collecting sample data and decisions from organizations for compliance analysis. Elaborating the guiding principles would also bring the Model Framework better in line with emerging international standards for AI ethics and governance, including the *OECD AI Recommendations*, the *EU Guidelines for Trustworthy AI*, the *Beijing AI Principles*, and China's *AIIA Joint Pledge on AI Industry Self-Discipline*.

The Glossary includes some elaboration of each guiding principle, but at the same time makes it clear that those elaborations are for consideration by companies, and may not be addressed by the Model Framework. The dedicated explanations of guiding principles could incorporate relevant points from the Glossary, thus clarifying which principles in the Glossary constitute the "consistent core set of ethical principles" that the Model Framework upholds.

Safeguards for both software practice and human decisions

We advocate for a maxim which we have generalized and augmented from the *MAS FEAT Principles*. Since AI is **software** that assists or replaces certain **human capabilities**, both of these aspects should be subject to governance. The development, deployment and monitoring of AI must be subject to at least the same level of safeguarding scrutiny as software, involving testing, accountability, and so on. AI outputs — including decisions, generated content, and predictions — should also be held to at least the same ethical standards as human outputs for similar tasks.[6] It also naturally gives rise to the guiding principle of human-centricity, and extends existing internal governance of software to AI.

The social responsibility of companies

The Model Framework should emphasise the fundamental responsibility that companies have to proactively avoid causing harm to society through their operations. We advocate for the explicit mention of the ***social responsibility of companies beyond simply building consumer confidence and following procedures or best practices.***

AI ethics requires consultation with diverse voices

Voices from a diverse range of communities must be consulted throughout the development, deployment and monitoring of AI, to minimize the risk of harm and discrimination to those communities and beyond. This would be very difficult to realize without ensuring diverse

[6] Compare with point 6 of the Summary of Principles in MAS (2019), "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector". Retrieved on 16 June 2019 from http://www.mas.gov.sg/News-and-Publications/Monographs-and-Information-Papers/2018/FEAT.aspx

representation within the company workforce and leadership. Hence, companies should promote diversity of gender, race, age, dis/ability, and other social categories in hiring and promotion. As the AI Now Institute puts it in their white paper, *Discriminating Machines*, organizations needs to ask not just "Are humans in the loop?", but also "*Which* humans are in the loop?"[7]

Policy recommendations such as the Model Framework should also be crafted in consultation with diverse segments of society. The Acknowledgements of the Model Framework lists only corporations and trade organizations, but the interests of broader society need to be represented by other parties. In accordance with point 5.5d in the Glossary, which enshrines the need to "give weight to the considered judgments of people or communities affected by data practices", we recommend that the PDPC proactively seek out the voices of academics and members of civil society, who may be able to point out more areas for further scrutiny or tighter governance. For example, the PDPC could consider convening citizen juries on the impact of AI to gather input from a broad spectrum of society.[8]

The need for critical reflection beyond procedures and frameworks

While standard procedures and frameworks help companies to implement AI more carefully for the benefit of society, they are not wise to the entire interaction between AI systems and society at large, so they must never limit the companies' understanding or engagement with the risk to society. We recommend that the Model Framework stresses the importance of critical examination and reflection *beyond the framework* on the part of the company personnel responsible for AI development, deployment and monitoring. ***On their own, internal governance structures, standardized procedures, and technical safeguards cannot guarantee fairness or human-centricity.*** Not even the lack of statistical bias in data can completely prevent discrimination: reductive assumptions about a population—such as that members speak at least one out of a set of common languages like English, Malay, Mandarin, and Tamil—can result in discrimination against those who do not fit within that assumption. These openings for discrimination can only be eradicated through critical reflection about the entire process of building and using AI, and a willingness to adapt when such seeds of discrimination are found.

Eileen Oak, a researcher on risk assessment in social work, has argued that risk management frameworks can potentially neglect ethical and social realities that are too complex to fit into a standardized framework. This erodes the space for those using the framework to articulate problems or negotiate solutions that must be sensitive to humane contexts.[9] The Model Framework cannot be an end in itself; it has to be a tool for the overarching goal of ethical AI that

---

[7] Sarah Myers West, Meredith Whittaker, and Kate Crawford (2019), "Discriminating Machines: Gender, Race, and Power in AI", *AI Now Institute*. Retrieved on 21 June 2019 from https://ainowinstitute.org/discriminatingsystems.pdf

[8] Annie Pottorff (2019), "Citizens Juries on Artificial Intelligence", Retrieved on 24 June 2019 from https://participedia.net/case/5820

[9] Eileen Oak (2016), "A Minority Report for Social Work? The Predictive Risk Model (PRM) and the Tuituia Assessment Frameworkin addressing the needs of New Zealand's Vulnerable Children", *British Journal of Social Work* **46**, pages 1208–1223

benefits the population it serves—a tool that is complemented by the critical reflection of employees implementing AI that scours for possible sources of risk, discrimination, or harm.

Areas for Further Research and Governance

The competitive dynamics between AI companies are an area for further attention. Unregulated competition can be thought of as an underlying risk that causes AI companies to under-invest in explainability, fairness, transparency, and human-centricity. Investing in the measures suggested by the Model Frameworks takes time and money. Even if one company desires to slow down and adopt these measures, pressure to out-compete other companies erodes the option to do so. A company is less likely to adopt the Model Framework if it perceives that other firms are neglecting the Model Framework and rushing to market. As such, we suggest that follow-up documents lay out approaches for AI companies to build trust—not just with the public, but with each other. For example, the PDPC may convene meetings for companies to demonstrate their ability and intention to adhere to the Model Framework. The goal is to make it easier for AI companies to compete responsibly.

AI research undertaken by organizations also deserves scrutiny. AI research refers to both the research and development of a particular AI system for organizational use, and efforts to improve more general capabilities of AI, performed internally or with partners (e.g. in academia). Some of the potential concerns stem from AI research regardless of who is conducting it, such as an emphasis on AI capabilities research without corresponding research in AI safety, robustness, verification, fairness, and oversight, or the irresponsible pursuit of artificial general intelligence in a way that is misaligned with human values. As the *Asilomar AI Principles* put it, "The goal of AI research should be to create not undirected intelligence, but beneficial intelligence".[10]

Other concerns stem from companies being less-regulated sites of research. University research involving human subjects is subject to ethical review by Institutional Review Boards (IRBs). However, technology companies, which generally do not have comparable ethical review systems for human-subject research, regularly experiment on their customers with A/B testing, where such testing often involves AI technology such as personalized search and recommender systems. Moreover, in 2014, Facebook controversially tested for "emotional contagion" in its social network by surreptitiously manipulating newsfeed algorithms.[11] Facebook has since pointed out limitations to the IRB framework in how well it can guide research in companies, and presented its own internal ethical review mechanism adapted from IRBs.[12] These realities and initiatives could help to inform the proper governance of AI research in industry.

---

[10] Future of Life Institute (2017), "Asilomar AI Principles". Retrieved on 20 June 2019 from https://futureoflife.org/ai-principles/

[11] Kashmir Hill (2014), "Facebook Added 'Research' To User Agreement 4 Months After Emotion Manipulation Study", *Forbes*. Retrieved on 20 June 2019 from https://www.forbes.com/sites/kashmirhill/2014/06/30/facebook-only-got-permission-to-do-research-on-users-after-emotion-manipulation-study/#32223fbb7a62

[12] Molly Jackman and Lauri Kanerva (2016), "Evolving the IRB: Building Robust Review for Industry Research", *Washington and Lee Law Review Online* **72**(3), pages 442-457

The distribution of AI development, deployment, and monitoring across several companies may be another challenge in accountability that needs to be addressed. Companies deploying off-the-shelf AI developed by another company were specially mentioned in paragraph 2.2 of the Model Framework. However, it is likely that in many companies, the linear process of AI deployment in paragraph 3.13, or any iterative or non-linear versions of it, will be split up among several different companies.

This may not be as simple as purchasing off-the-shelf solutions: for example, company *A* may purchase an ML base model from company *B* which company *A* trains on data from a repository updated and maintained by company *C*. Company *A* then deploys this ML system. Company *A* needs to perform due diligence in each of these decisions, and consider which parts of the responsibility for the ML system it should retain and which should lie with Companies *B* and *C*. Parts of the "Operations Management" and "Customer Relations Management" sections may have to be adapted to a decentralized AI development and deployment similar to the following:

- The deployer of the AI should hold responsibility for ensuring that the AI it uses can satisfy the requirements in "Operations Management" and "Customer Relations Management", possibly by getting guarantees from the suppliers of each part of its AI system.
- The deployer should mitigate prejudice in data either by checking it themselves, or securing an undertaking from the data supplier.
- The responsibility for explainability may be split among several companies. For instance, when a base model has been purchased and trained on further data, both the companies which train the base model and train it afterwards may have to bear some responsibility for explanation.

The Model Framework may draw some inspiration from the way that liability is divided among the myriad manufacturers and suppliers that contribute the parts for a single product, as well as the web of accountability between them.

TABLE OF CONTENTS

FOREWORD

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

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

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

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

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

S Iswaran
Minister for Communication and Information
Singapore
January 2019

1. PREAMBLE

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

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

[1.1a] We support the algorithm-agnostic vision for the Model Framework, but the current Model Framework, especially the Operations Management section, appears to focus on data-driven machine learning. On the other hand, most of the Model Framework generalizes well to AI that is not data-driven or machine learning, such as expert systems, robotics, and voice synthesis. We believe that regardless of whether an AI system is data-driven, it has features that warrant governance under a unified framework. Moreover, AI that is not data-driven will continue to play significant roles in commerce.

To ensure that companies appreciate the full scope of application of the Model Framework, we recommend integrating discussions of wide array of AI types, data-driven or otherwise, into this document. Recommendations to this effect are made after the definition of AI (paragraph 2.12) and at the end of the Operations Management section.

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

[1.1b] We suggest replacing the term "Technology-agnostic" with "Implementation-agnostic", because "technology" can be interpreted to mean the combination of algorithms with specific software and hardware implementations, leaving room for future confusion. "Implementation" is a less ambiguous word that captures the essence of this paragraph.

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

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

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

[1.2a] The current Model Framework *does* advocate a core set of ethical principles, which are encapsulated within the "guiding principles" of *explainability*, *fairness*, *transparency* and *human-centricity*. However, the brief explanations in paragraphs 2.5a-b do not convey a concrete idea of what each principle entails. Recommendations on clarifying elaboration will be made after paragraphs 2.5a-b.

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

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

[1.2c] As the societal impacts of AI are a main motivation for the Model Framework, this paragraph should be rephrased to clarify that it excludes discussions of *legal liabilities associated with the societal impacts of AI*, and not discussions of the societal impacts themselves.

2. INTRODUCTION

Objectives

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

[2.1] Since many of the harms and risks that AI can pose to society can be counterintuitive and unexpected, we advocate for a more extended and comprehensive discussion specially dedicated for harms and risks.

One useful classification of harms was presented by Kate Crawford from the AI Now Institute. The Model Framework focuses on what Crawford has termed *harms of allocation*, or outcomes of AI systems which specially rewards or punishes a particular group.[13] We suggest that the Model Framework explicitly highlight examples of another major category: *harms of representation*, which involve stereotyping, denigration, under-representation, and other ways of reflecting a picture of human identity that perpetuates inequalities along class, race and other demographic lines. Examples of harms of representation include AI that reinforces racist standards of beauty,[14] and AI that is trained to understand and speak in only American or British accents. These harms do not directly inflict tangible losses on particular groups, but progressively degrade mutual tolerance and the social fabric by influencing beliefs and attitudes about groups of people. These erroneous beliefs and attitudes fuel harms of allocation in a vicious cycle.

The above harms are often amplified by the un-interpretability or autonomy of AI, or its deployment at speed or scale. The harms may also manifest out of left field due to the lack of oversight. Misguided optimization of AI objectives that were not specified carefully could also engender harmful side effects.[15]

2.2 The Personal Data Protection Commission (PDPC), Infocomm Media Development Authority (IMDA), with the advice from the Advisory Council on the Ethical Use of AI

[13] Aarthi Kumaraswamy (2017), "20 lessons on bias in machine learning systems by Kate Crawford at NIPS 2017", *Packt Hub*. Retrieved on June 15, 2019 from https://hub.packtpub.com/20-lessons-bias-machine-learning-systems-nips-2017/

[14] Elena Cresci (2017), "FaceApp apologises for 'racist' filter that lightens users' skintone", *The Guardian*. Retrieved on June 15, 2019 from https://www.theguardian.com/technology/2017/apr/25/faceapp-apologises-for-racist-filter-which-lightens-users-skin tone

[15] Jeffrey Dastin (2018), "Amazon scraps secret AI recruiting tool that showed bias against women", Reuters. Retrieved on 17 June, 2019 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

and Data ("Advisory Council"), proposes for consultation this first edition of a voluntary Model Framework as a general, ready-to-use tool to enable organisations that are deploying AI solutions at scale to do so in a responsible manner. This Model Framework is not intended for organisations that are deploying updated commercial off-the-shelf software packages that happen to now incorporate AI in their feature set.

[2.2] We believe that the Model Framework should provide some important guidelines for deploying off-the-shelf packages, especially because many organizations use "AI as a Service" (AIaaS) from external vendors,[16] and there is no guarantee that AI will behave as expected in new use cases. Companies should be held accountable for their use of externally-sourced AI packages, [17] and their internal governance should build in procedures to conduct due-diligence checks of off-the-shelf AI products that are considered for procurement. Such due diligence is part of the company's responsibility to consumers, and would also boost consumer confidence.

Standard tests for externally-sourced software should continue to serve a gatekeeping function when the software is AI; such tests include black-box testing and simulation environments. However, we recognize some limitations to simulation environments to the extent that AI may only function properly if it interacts with the actual human population. In addition, most of the precautions and tests from the "Operations Management" section could be applied to off-the-shelf AI solutions, especially tests for statistical bias or prejudicial treatment. Companies should find out what tests, monitoring, and safeguards are part of the internal governance of vendors of off-the-shelf AI. Preference could be given to vendors whose AI products are transparent and give explainable outputs.

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

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

[2.3a] The Model Framework should emphasise the fundamental responsibility that companies have to proactively avoid causing harm to society through their operations. We recommend that paragraph 2.3 explicitly mention the ***social responsibility of companies beyond simply building consumer confidence and following procedures or best practices.***

[16] For instance, Amazon has controversially sold its facial recognition service to law enforcement agencies in the United States. See Elizabeth Dwoskin (2018), "Amazon is selling facial recognition to law enforcement — for a fistful of dollars", *Washington Post*. Retrieved on 18 June 2019 from https://www.washingtonpost.com/news/the-switch/wp/2018/05/22/amazon-is-selling-facial-recognition-to-law-enforcement-for-a-fistful-of-dollars

[17] See point 8 of the Summary of Principles in MAS (2019), "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector". Retrieved on 16 June 2019 from http://www.mas.gov.sg/News-and-Publications/Monographs-and-Information-Papers/2018/FEAT.aspx

As a sample, the *OECD AI Recommendations* state that companies that develop or deploy AI should, like all other stakeholders in the AI ecosystem, "proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of under-represented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being."[18] A useful parallel can be drawn with the notion of "Extended Producer Responsibility"[19]: a company that develops or deploys AI should bear some of the responsibility and costs of harm to society caused by their AI.

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

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

Guiding Principles

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

[2.5] The guiding principles are defined very briefly. Although there are further elaborations deep in the body of the Model Framework, we strongly advocate for dedicated paragraphs that explain each guiding principle more clearly and in depth, with examples. The Model Framework could borrow from the Glossary (which has currently been marked as not necessarily applicable to the Model Framework), or the Summary of Principles in the *MAS FEAT Principles*, which breaks down each principle into constituent parts, each substantiated with examples.

[18] Paragraph 1.1 in OECD (2019), *Recommendation of the Council on Artificial Intelligence*. Retrieved on 16 June 2019 from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
[19] OECD, "Extended Producer Responsibility". Retrieved on 16 June 2019 from http://www.oecd.org/environment/extended-producer-responsibility.htm

Elaborating the principles would also align the framework with emerging international standards, such as the *OECD AI Recommendations*, the *EU Guidelines for Trustworthy AI*, the *Beijing AI Principles*, and China's *AIIA Joint Pledge on AI Industry Self-Discipline*.

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

[2.5a] "Fairness" needs to be defined more clearly in the main text to solidly frame the ethical dimensions of this Model Framework. For example, the Model Framework could elaborate on statistical notions of fairness (sometimes called group fairness) and individual notions of fairness. Points 5.4a-c in the Glossary, which elaborate on "fairness", are well-put and critical to upholding the value of fairness. Thus we recommend that points 5.4a-c, or equivalent elaboration, be brought into the explanation of the guiding principles in this paragraph. In particular, it is important to explicitly link fairness together with avoiding discrimination early on, which can be achieved using point 5.4a.

The Model Framework should also recommend that organizations develop metrics to quantify explainability, transparency, and fairness, so that organizations are able to track and demonstrate accountability to these guiding principles. This point is elaborated in our response to paragraph 3.5.

Finally, to underscore the fundamental objective of this Model Framework, which is to avoid harm to society, we suggest amending the final sentence of this paragraph in a way similar to "This helps build trust and confidence in AI, and more importantly, helps to ensure that AI does not lead to undue discrimination or harm."

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

[2.5b] While the above principle is laudable when described in full, the term "human-centric" by itself is extremely vague and lends itself to creative (mis)interpretation: almost any kind of technology could be described as "human-centric" in some way, but still lead to harm (e.g. understanding the preferences of smokers very well when advertising cigarette brands to them) or simply give no real benefit. Other terms could be used that are similarly broad without allowing for misinterpretation. For example, a useful reference point is the *Belmont Report,*[20] whose principles of **beneficence** and **respect for persons** in research on human subjects are much

[20] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979), *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Retrieved on 16 June 2019 from: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

harder to creatively misinterpret in ways that harm people. Similarly, China's *AIIA Joint Pledge on AI Industry Self-Discipline* also commits to more concrete top-level principles than being "human-oriented", emphasising the importance of **enhancing well-being**, **fairness and justice**, and **avoiding harm**, all of which are harder to misinterpret.[21]

As such, the scope of human-centricity should be more clearly defined, to preclude situations where any company can label their AI solutions as "human-centric" just because they contribute to some part of human well-being in some way, regardless of other issues with the AI solution. A useful question to ask is, "what kind of AI system might a well-intentioned company build that is not human-centric?" If it is difficult to answer this question, this suggests that "human-centric" is too vague and broad a term to be useful for governance.

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

[2.6] The above appears to be an attempt to identify certain concerning qualities that characterize AI and separates it from earlier non-AI technologies. As such, it might be best moved into the definition of AI in paragraph 2.12. Organization aside, this paragraph appears to conflate several (interacting) features often found in contemporary AI technology: **un-interpretability**, partial or full **autonomy**, and **speed/scale**. It would be best to state these features explicitly, so that organizations can better identify when this Model Framework applies to the technologies they are using, and in order not to conflate them. A recommendation to this effect will be made after paragraph 2.12.

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

Assumptions

[21] Graham Webster (2019), "Translation: Chinese AI Alliance Drafts Self-Discipline 'Joint Pledge'". *New America*. Retrieved on 25 June 2019 from https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-ai-alliance-drafts-self-discipline-joint-pledge/

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

[2.8] Much of the Model Framework—including the risk management, assignment of responsibility, and customer relations management—is actually useful to govern AI that is not data-driven, such as those based on decision trees or small datasets. Hence we recommend that the Model Framework expand its scope to include non-data-driven AI. This paragraph could be rephrased to read: "The Model Framework includes a discussion of good data management practices, but also addresses AI that is not data-driven. It is equally applicable to big data AI models, pure decision-tree-driven AI models, or small dataset AI methods."

2.9 The Model Framework does not address the risk of catastrophic failure due to cyberattacks 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.

[2.9] There is scope for the Model Framework to productively address cyberattacks and catastrophic failure. Other than the standard precautions and countermeasures to guard software systems against cyberattacks and widespread failure, there are AI-specific measures that can be taken. Indeed, the OECD AI Recommendations state that "AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose an unreasonable safety risk."[22]

Examples of countermeasures include keeping abreast of the latest techniques in adversarial attacks on AI, or the hiring of AI security consultants where system failure would incur a huge social cost to consumers. The Model Framework could draw from the large body of literature on technical AI safety and assurance to provide useful guidance to companies, or refer companies to related resources. For instance, Google DeepMind broke down technical AI safety into the areas of *specification*, *robustness,* and *assurance*, and gave concrete recommendations for each.[23] Bringing these technical risks and precautions to the attention of companies could go a long way to forestalling catastrophic AI failure in the private sector, which can cut across jurisdictions and all sectors of society due to the scalability of AI. Safeguards could be taken in proportion to risk, as judged using the matrix of probability and severity of harm (paragraph 3.11). The Model Framework should institute procedures for these safeguards.

A further measure to guard against malicious attacks is to predict the ways in which the AI system can be attacked or abused, as recommended by DJ Patil *et al.*[24]

---

[22] Paragraph 1.4(a) in OECD (2019), *Recommendation of the Council on Artificial Intelligence*. Retrieved on 16 June 2019 from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[23] Pedro A. Ortega, Vishal Maini, and the DeepMind safety team (2018), "Building safe artificial intelligence: specification, robustness, and assurance", *Medium*. Retrieved on 16 June 2019 from https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1

[24] See Chapter 2 in DJ Patil, Hilary Mason, Mike Loukides (2018), *Ethics and Data Science*, O'Reilly.

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

Definitions

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

AI Solution Provider Organisation Individuals

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

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

[2.12] This definition of AI, while broad, is closer to an academic definition of AI, and does not highlight the policy-relevant applications and dimensions of AI. We recommend using a definition that highlights both of these aspects, so as to clarify which technologies fall under the umbrella of the Model Framework, and to better distinguish AI from other technologies that do not share the same risk profile. No doubt, there is a spectrum of such technologies, some of which are closer to 'AI' than others. It is thus important for a policy-facing definition of AI to emphasize the attributes that make AI and AI-adjacent technologies risky, so that the Model Framework can be applied even when a similarly risky technology is not conventionally called 'AI'.

In constructing such a definition, we believe both the *MAS FEAT Principles* and *OECD AI Recommendations* to be good reference points.[25] The OECD definition is a good foundation because it satisfies the following criteria, as researched and argued for by Krafft *et al.* after comprehensive surveys of both AI researchers and policy-makers[26]: "a good definition of AI should include both current and future applications of AI, be accessible to laypersons, and be implementable in policy through reporting and oversight."

[25] Paragraph I in OECD (2019), *Recommendation of the Council on Artificial Intelligence*. Retrieved on 16 June 2019 from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[26] Peter Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo (2019), "Policy versus Practice: Conceptions of Artificial Intelligence". Preprint under review, retrieved on 16 June 2019 from http://people.csail.mit.edu/pkrafft/papers/critplat-policy-vs-practice.pdf

Taking into account all of the above, we propose the following definition of AI, adapted from the *MAS FEAT Principles* and the OECD definition, for consideration by the PDPC: **"Artificial Intelligence (AI) is a set of machine-based technologies designed to assist or replace human capabilities, according to human-defined objectives. These capabilities include, but are not limited to, prediction, recommendation, decision making and content generation. AI technologies are often partially or wholly autonomous, not fully interpretable, and deployed at speed or at scale."**

This proposed definition is more inclusive than the current one, and emphasizes the reality that the goals of AI are those of humans, and are not intrinsic to an AI system.[27] This gives rise naturally to the guiding principles of human-centricity, beneficence, and respect for persons. It also highlights the features we noted in our response to paragraph 2.6 — these are the features we believe distinguish the risk profile of AI from other kinds of technology.

This definition also includes data analytics techniques like regression and hypothesis testing, AI that is not data-driven like expert systems, and AI that does not make decisions, such as content-generating AI.

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

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

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

[Definition of "Bias"] We suggest adding a definition for "Bias" in order to distinguish "bias" as unjustified or unfair treatment from "bias" in the technical sense used in statistics and AI. Given that "bias" has so many meanings, we suggest using the terms "prejudice" or "discrimination"

[27] *Ibid.*

when the former meaning is intended, and "statistical bias" to refer to the technical sense of the term. This distinction is important in practice, because an AI system could be statistically unbiased yet operate with prejudice. This can occur, for example, when the AI is founded on prejudiced assumptions, such as when a categorizing AI assumes that there are only two genders or neglects the existence of minority groups.

A suggested definition is as follows: "**Bias**, in the context of AI governance, can refer both to statistical bias in a dataset or AI algorithm, or to the unjustified or unfair treatment that either results from or occurs alongside statistical bias. To avoid confusion between these two senses of the term, we shall generally use 'statistical bias' to refer to bias in the value-neutral sense, and 'prejudice' and/or 'discrimination' to refer to the latter use of the term."

3. MODEL AI GOVERNANCE FRAMEWORK

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

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

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

[3.1b] Many forms of AI do not fall neatly into the category of "decision-making"; consider search engines, ad placement, chatbots, and image generation. On the other hand, the content in the "Determining AI Decision-Making Model" section—assessing risk with local context in mind, calibrating risk appetite in light of corporate objectives, and tailoring the amount of human oversight to the amount of risk—is not specific to decision-making AI, and can be generalized to all types of AI. Hence we suggest renaming that section to "Assessing Risk and Calibrating Human Oversight", and rephrasing its content to reflect that it applies to all AI, whether it makes decisions or not. Paragraph 3.1b should be amended to read similar to: "**Assessing Risk and Calibrating Human Oversight:** A methodology to aid organisations in setting its risk appetite for use of AI, i.e. determining acceptable risks, and calibrating the appropriate amount of human oversight in implementing AI." References to "decision-making models" elsewhere in the Model Framework should be amended accordingly. More specific recommendations along this vein will be made throughout the section in question.

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

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

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

Internal Governance Structures and Measures

3.3 Organisations should have internal governance structures and measures to ensure robust oversight of the organisation's use of AI. The organisation's existing internal governance structures can be adapted, and/or new structures can be implemented if necessary. For example, risks associated with the use of AI can be managed within the enterprise risk management structure; ethical considerations can be introduced

as corporate values and managed through ethics review boards or similar structures. Organisations should also determine the appropriate features in their internal governance structures. For example, when relying completely on a centralised governance mechanism is not optimal, a de-centralised one could be considered to incorporate ethical considerations into day-to-day decision-making at operational level, if necessary. The sponsorship, support and participation of the organisation's top management and its Board in the organisation's AI governance are crucial.

[3.3] We recommend that the Model Framework include the principle, articulated in the *MAS FEAT Principles*, that the outputs created by AI—decisions, content, recommendations, etc.—are held to *at least* the same ethical standard as similar outputs created by humans.[28]

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

1. <u>Clear roles and responsibilities for the ethical deployment of AI</u>
   a. Responsibility for and oversight of the various stages and activities involved in AI deployment should be allocated to the appropriate personnel and/or departments. If necessary and possible, consider establishing a coordinating body, having relevant expertise and proper representation from across the organisation.

[3.4(1)a] The assignment of AI responsibility to a subset of the organization needs to be balanced by creating broader awareness of the use and implications of AI in the company and management. The MAS FEAT Principles highlighted the need for "proper due diligence so that approving authorities have sufficient understanding of the data and model logic used for decision-making."[29] The "materiality or complexity" of certain decisions about the use of AI may also require higher management approval than normal,[30] resulting in a need for upper management to be informed about AI.

   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.

[3.4(1)b] We strongly recommend adding that personnel should undergo AI ethics training, and not just AI technical training. In particular, (AI) engineers are often not trained to think about the social context and impact of their work, and it is important that they be able to think critically about these issues on a day-to-day basis, rather than rely upon the reasoning that they are "just

[28] See paragraph 6.2 in MAS (2019), "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector". Retrieved on 16 June 2019 from http://www.mas.gov.sg/News-and-Publications/Monographs-and-Information-Papers/2018/FEAT.aspx
[29] Ibid., paragraphs 7.5 and 7.6
[30] Ibid., paragraph 7.5

engineers". Kate Crawford has identified the "just-an-engineer" syndrome as a major issue in the ethics of AI.[31]

[Diversity in the workforce and leadership] A report by the AI Now Institute decried the lack of gender and racial diversity in the AI industry.[32] A lack of diversity in a company's workforce and leadership could result in a narrower view of societal problems, leading to AI systems being designed without the needs of minority groups in mind, or with avoidable avenues for discrimination. A case in point is how a now-defunct recruiting tool designed by Amazon discriminated against female applicants because it was trained on a collection of resumes that reflected the male dominance of the industry.[33] Therefore companies should promote gender, race, age, and other forms of diversity in hiring and promotion.

c. Key roles and responsibilities that should be allocated include:
   i. Using any existing risk management framework and applying risk control measures (See further "Risk management and internal controls" below) to

      - Assess and manage the risks of deploying AI (including any potential adverse impact on the individuals, e.g. who are most vulnerable, how are they impacted, how to assess the scale of the impact, how to get feedback from those impacted, etc.)

      - Decide on appropriate AI decision-making models.

[3.4(1)c(i)] Since the intention behind discussing "decision-making models" seems to be the calibration of human oversight in proportion to risk,[34] we suggest restating "Decide on appropriate AI decision-making models" as "Decide the appropriate degree of human oversight".

      - Manage the AI model training and selection process.

We recommend changing the above sentence to 'Manage the processes for AI model development, training, section', so as to include non-ML AI technologies that do not involve training on data.

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

---

[31] Kate Crawford (2019), "The Politics of AI", a talk at the Royal Society. Retrieved on 16 June 2019 from https://www.youtube.com/watch?v=HPopJb5aDyA

[32] Sarah Myers West, Meredith Whittaker, and Kate Crawford (2019), "Discriminating Machines: Gender, Race, and Power in AI", *AI Now Institute*. Retrieved on 21 June 2019 from https://ainowinstitute.org/discriminatingsystems.pdf

[33] Jeffrey Dastin (2018), "Amazon scraps secret AI recruiting tool that showed bias against women", Reuters. Retrieved on 17 June, 2019 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[34] See our comment after paragraph 3.4(1)b.

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

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

2. Risk management and internal controls

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

b. Such measures include:

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

ii. Establishing monitoring and reporting systems as well as processes to ensure that the appropriate level of management is aware of the performance of and other issues relating to the deployed AI. Where appropriate, the monitoring can include autonomous monitoring to effectively scale human oversight. AI systems can be designed to report on the confidence level of their predictions, and explainability features can focus on why the AI model had a certain level of confidence, rather than why a prediction was made.

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

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

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

[Confidential reporting of misuse and malpractice] Companies should also establish and strengthen internal mechanisms for employees to confidentially report issues related to AI sourcing, development, and deployment, as well as data management. Employees who make such

reports should be protected against reprisal. We also recommend that procedures be established in internal governance to escalate problems to higher management when necessary.

Determining AI Decision-Making Model

[Governing AI that does not make decisions] As noted in our comment after paragraph 3.1b, not all AI technologies can be readily described as being used to replace or assist human decision-making. For example, automated image generation (e.g. thispersondoesnotexist.com), image modification (e.g. Snapchat and Meitu), caption generation, text completion (e.g. Gmail, TalkToTransformer.com), machine translation, voice recognition, or chat-bots, all use AI to directly provide media and text services to users. These classes of AI cannot be naturally described as making decisions. Even technologies like AI-driven recommendations and search engine optimization do not fall neatly into the "decision-making" mould.

In order to include these increasingly common AI technologies within this framework, we suggest reframing and rewording this section to include the risks of non-decision making AI. In particular, while non-decision-making AI is less likely to result in *harms of allocation* (e.g. wrongly or unfairly denying a service to an individual based on an AI decision), it can still ready result in *harms of representation* (e.g. AI that is trained to only understand and speak in American or British accents, or AI that reinforces Eurocentric beauty and gender norms).[35] These harms and risks should also be discussed in the model framework. This section should also be renamed to **Assessing Risk and Calibrating Human Oversight** in order to preserve its general thrust while making it inclusive of non-decision-making AI.

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

[3.5] The specification of objectives for AI needs utmost care. Many AI systems have already caused unexpected negative social impacts through misguided optimization for objectives that are not specified well.[36] Data analytics systems can create toxic feedback loops that perpetuate adverse conditions which increase measurements of progress towards their objectives.[37] "Goodhart's

[35] Harms of allocation and harms of representation are explained in further detail in Aarthi Kumaraswamy (2017), "20 lessons on bias in machine learning systems by Kate Crawford at NIPS 2017", *Packt Hub*. Retrieved on June 15, 2019 from https://hub.packtpub.com/20-lessons-bias-machine-learning-systems-nips-2017/

[36] Jeffrey Dastin (2018), "Amazon scraps secret AI recruiting tool that showed bias against women", Reuters. Retrieved on 17 June, 2019 from https://www.reuters.com/article/us-amazon-com-jobs-idUSKCN1MK08G

[37] Cathy O'Neil criticized the Level of Service Inventory-Revised (LSI-R), a criminal recidivism risk model used in the United States, for predicting high-risk for individuals who probably lack employment and lived in an area with frequent encounters with law enforcement, thus increasing their sentences, making it harder for them to find jobs after release and thus more likely for them to be imprisoned again. This toxic cycle helps the LSI-R fulfill its own prophecy

Law" observes that complex systems that are scored on any one variable will definitely be "gamed",[38] so AI systems may need more metrics to optimize to avoid harm.[39] Therefore, AI systems must be explicit on their objectives and include some mitigation of harms in their objectives. These objectives need to be reviewed if unexpected side effects crop up over the course of deployment.

Hence, **the Model Framework should recommend that metrics for the success of AI deployment be formulated.** These could include quantitative metrics (e.g. the amount of money, time or manpower saved, debt reduction among customers) qualitative metrics (e.g. customer satisfaction, increase of customer autonomy), as well as metrics that lie in between (e.g. fairness, lack of discrimination, consistency and stability of operation). The AI deployed should be regularly measured, quantitatively or qualitatively, against these metrics and metrics, and both the AI and the metrics should be updated to match corporate goals. The results of metrics that impact society, such as metrics that measure demographic bias, could be made available to the public or the government. For instance, the National Environment Agency (NEA) collects figures from companies about the amount and material makeup of products and packaging sold.

[False positives and negatives] An important class of risks to consider are the likelihood and severity of impacts arising from any false positives or false negatives that result from AI. The company could compare this with the risks and impacts from false positives or negatives when humans make the decisions, set the firm's appetite for these kinds of risks, and use it to calibrate the amount of human oversight.

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

[3.6] Highlighting the diversity between countries should be balanced by pointing out the diversity within countries. Organisations should try to avoid making static assumptions or generalizations about individual and social preferences within each country or region. Hiring local consultants and research firms can aid in striking an appropriate balance.

---

and boosts the "accuracy" of the LSI-R, lending it a veneer of legitimacy. More details are available in Cathy O'Neil (2017), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Broadway Books (New York), pages 23-27.

[38] Jack Clark and Dario Amodei (2016), "Faulty Reward Functions in the Wild", OpenAI Blog. Retrieved on 17 June 2016 from https://openai.com/blog/faulty-reward-functions/

[39] Theo O'Donnell (2018), "AI in Impact Investing: Goodhart's Law, Unintended Consequences, and the Dangers of Blunt Metrics", SILO.AI. Retrieved on 17 June 2019 from https://silo.ai/ai-in-impact-investing-blunt-metrics/

We also recommend that the example about recommendation engines and algorithmic trading would best be used in a separate paragraph because it concerns "sector-specific" risks, in contrast to the "culture-specific" or "country-specific" risks in the rest of the paragraph.

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

[3.7] We suggest adding another example of emergent behavior which has had high impact, and is still fresh in the popular imagination: social media newsfeed optimization can affect social polarization and electoral outcomes.

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

[3.8] To promote the guiding principle of fairness, the Model Framework should recommend that companies list, for the sake of internal accountability, the protected categories that they will try to avoid discriminating against. This list can be informed by corporate values and context. For example, the *Charter of Fundamental Rights of the European Union* forbids discrimination based on the protected categories of "sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation".

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

[3.9] The risk impact assessment should include the quantitative or qualitative measurement of those risks, and continual tracking of those risks over the course of AI deployment.

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

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

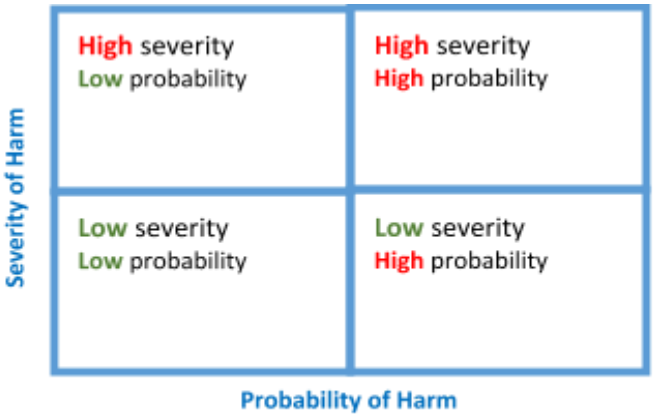
b. **Human-out-of-the-loop.** This model suggests that there is no human oversight over the execution of decisions. AI has full control without the option of human override. For example, a product recommendation solution may automatically suggest products and services to individuals based on pre-determined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

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

[3.10] Human oversight should not only be calibrated only for AI decision-making, but also for other capabilities of AI such as recognition, content creation or prediction. Hence the scope of paragraphs 3.10a-c should be broadened by focusing on general AI outputs instead of only decisions. Examples from AI that do not make decisions should be given to affirm that calibrating human oversight applies generally to those forms of AI as well. For example, a human-in-the-loop facial recognition AI may suggest candidate identities, and associated confidence levels, computed from a certain face detected from an image, and leave a human operator to make the final selection of the identity. A human-over-the-loop AI to generate prose may add constraints during the generation process, such as a request for particular emotional expression. A human-over-the-loop global climate simulator might experiment with different regional climate interventions as the simulation proceeds, to test how various sets of interventions will influence the global climate.

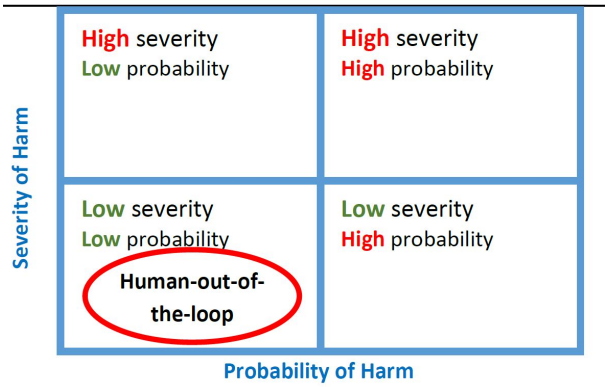
3.11 The Model Framework also proposes a matrix to classify the probability and severity of harm to an individual as a result of the decision made by an organisation about that individual. The definition of harm and the computation of probability and

severity depend on the context and vary from sector to sector. For example, the harm associated with a wrong diagnosis of a patient's medical condition will differ from that associated with a wrong product recommendation for apparels.



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

Illustration:
An online retail store wishes to use AI to fully automate the recommendation of food products to individuals based on their browsing behaviours and purchase history. The automation will meet the organisation's commercial objective of operational efficiency.
Severity-Probability Assessment: The definition of *harm* can be the impact of making product recommendations that do not address the perceived needs of the individuals. The *severity* of making the wrong product recommendations to individuals may be low since individuals ultimately decide whether to make the purchase. The *probability of harm* may be high or low depending on the efficiency and efficacy of the AI solution.
Degree of human intervention in decision-making process: Given the low severity of harm, the assessment points to an approach that requires no human intervention. Hence, a human-out-of-the-loop model is adopted.

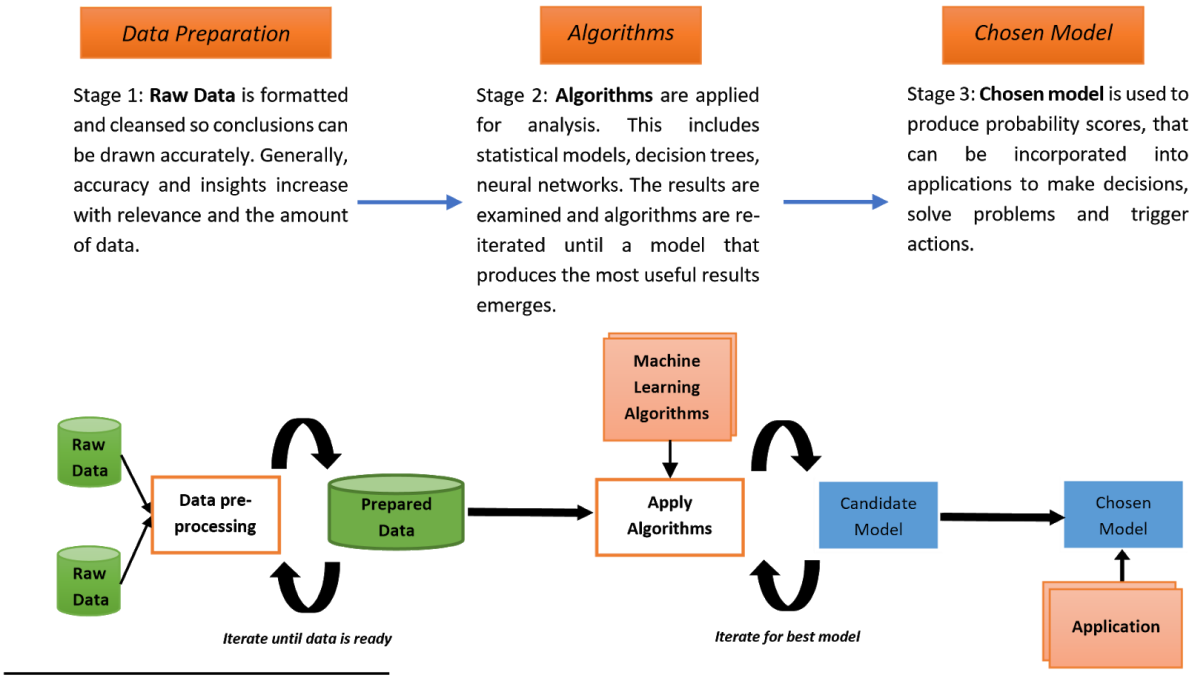


Regular review: The organisation can review this approach regularly to assess the severity of harm and as societal norms and values evolve. For example, the product recommendation solution may consistently promote sugary drinks to certain individuals. With heightened concerns about diabetes, the organisation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

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

Operations Management

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



[1] Adapted from Azure

[3.13] This description of AI deployment applies specifically to machine learning (ML) algorithms, which do not include all AI technologies. Given that that non-ML-based AI is still widely used, and is unlikely to be entirely replaced by ML in the near future, we recommend that the "Operations Management" section expand its scope to include guidelines for a broader range of AI technologies and algorithms. The following are some sample operations management guidelines relevant to AI that does not fall strictly within the ML paradigm:

1. **Expert systems and knowledge engineering** (e.g. most chatbots today) require careful curation of the sources of expertise (professionals, books, etc.) used to inform model development.
2. **Autonomous systems** (e.g., service or manufacturing robots, self-driving vehicles) require safety guarantees, oversight, and continual testing.
3. **Large-scale real-time web-based systems** (e.g. Google search, Facebook's newsfeed, algorithmic trading platforms) require continual oversight, planning to consider systemic risks, and the capacity to respond quickly and effectively to errors and failures.

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

Data for Model Development

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

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

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

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

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

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

   Keeping a **data provenance record** allows an organisation to ascertain the quality of the data based on its origin and subsequent transformation, trace potential sources of errors, update data, and attribute data to their sources. The Model Framework recognises that in some instances, the origin of data could be difficult to establish. One example could be datasets obtained from a trusted third-party who may have commingled data from multiple sources. Organisations should assess the risks of using such data and manage them accordingly.

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

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

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

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

iv. How recently the dataset was compiled or updated.

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

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

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

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

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

[3.16c] Following our comment after the "Definitions" section, we advocate for clearly distinguishing between "bias" as a form of unjustified differential treatment, and "bias" in its technical, value-neutral sense. We suggest referring to the former as "prejudice" or "discrimination", and the latter as "statistical bias". This distinction matters because even statistically unbiased AI systems can give rise to prejudicial treatment, as demonstrated in our earlier comment.

Unfortunately, paragraph 3.16c and its subparagraphs appear to conflate statistical bias with prejudice. Selection bias on its own is a neutral, statistical term, but if it leads to prejudicial treatment of Asian people, then it becomes a form of unjustified differential treatment. If it simply

leads to under-representation of the number of bicycles in Singapore in the context of modelling transportation trends, then this is a problem of inaccuracy, and not of unjustified differential treatment. To address this conflation, We have accordingly suggested amendments to the aforementioned paragraphs.

The Model Framework should also caution that just because an AI system learns a trend from data in a statistically unbiased way, it may not be fair or human-centric to operate based on that trend. For example, Amazon's recruitment tool scanned resumes submitted to the firm, learned the trend of male predominance in the software industry, and consequently favoured male candidates.[40] For recruitment tools to be fair and human-centric, they cannot simply replicate and reinforce such trends.

i. **Selection bias.** This statistical 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*~~ *over-representation bias*. Omission bias describes the omission of certain characteristics from the dataset, e.g. a dataset of Asian faces only will exhibit omission bias if it is used for facial recognition training for a population that includes non-Asians. This kind of statistical bias may then lead to prejudicial treatment against Asians. A dataset of vehicle types within the central business district on a weekday may exhibit ~~stereotype~~ over-representation 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. This may then lead to inaccurate modelling outcomes.

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

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

d. **Different datasets for training, testing, and validation.** Different datasets are required for training, testing, and validation. The model is trained using the training data, while the model's accuracy is determined using the test data. Where

[40] Jeffrey Dastin (2018), "Amazon scraps secret AI recruiting tool that showed bias against women", Reuters. Retrieved on 17 June, 2019 from https://www.reuters.com/article/us-amazon-com-jobs-idUSKCN1MK08G

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

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

[3.16e] Regarding dataset review, specific factors should be identified that would trigger reviews of datasets, such as algorithm malfunction or egregious errors.[41]

Companies should certainly hold data that is up to date, but the Model Framework should caution that it may not always be fair or human-centric to use, draw conclusions from, or otherwise give weight to very recent data. Consider the example of an insurance firm whose premiums are computed by an AI algorithm that takes into account the exercise patterns of the insured. If an insured individual does not exercise for a week, the algorithm may raise the premium to reflect an expected decrease in fitness. However, this period without exercise may not reflect any lasting change in exercise habits, so premium increase may not be fair. Extrapolating a recent fluctuation into a trend may not respect human self-determination, and thus may not be human-centric. Companies should consider if the proper functioning of their AI systems would require retaining older versions of data, and how the old data should be used alongside the new.

[Governing data content: personal attributes] We support a recommendation from the *MAS FEAT Principles* for a content-specific measure: justifying any use of personal attributes as input to an AI system, as a safeguard against discriminating along demographic lines.[42] However, we caution that removing personal attributes from the inputs may not prevent such discrimination, which could still happen based on proxy variables that are strongly correlated to the personal attributes. Our comment after paragraph 3.22b elaborates on this possibility.

[41] See paragraph 5.7 in MAS (2019), "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector". Retrieved on 16 June 2019 from http://www.mas.gov.sg/News-and-Publications/Monographs-and-Information-Papers/2018/FEAT.aspx
[42] *Ibid.*, paragraph 5.5

Algorithm and Model

[Reviewing algorithms and models] Building on a recommendation from the *MAS FEAT Principles*,[43] algorithms and models should undergo regular software verification and validation to ensure that they perform to specifications and satisfy the needs of the end-users. Such testing could include comparing the outputs of the AI systems with what humans would have produced for the same task.

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

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

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

3.20 It should be noted that technical explainability may not always be enlightening, especially to the man in the street. Implicit explanations of how the AI models' algorithms function may be more useful than explicit descriptions of the models'

[43] *Ibid.*, paragraph 5.6 and 5.7

logic. For example, providing an individual with counterfactuals (such as "you would have been approved if your average debt was 15% lower" or "these are users with similar profiles to yours that received a different decision") can be a powerful type of explanation that organisations could consider.

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

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

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

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

[3.22b] We suggest moving contrafactual fairness testing out of paragraph 3.22 and into its own paragraph, since it is not really just a subordinate of "repeatability". The Model Framework should warn that this fairness testing is important even if the model does not take demographic information as explicit inputs, because of the propensity for machine learning algorithms to discriminate based on "proxies", or variables which are not directly related to demographics but are strongly correlated with them. As examples, language is often used as a proxy for race in Singapore, while in USA, "redlining" is a practice that systematically withholds goods and services from particular neighbourhoods that are associated with certain races, using geographic location as a proxy for race. Contrafactual fairness testing would help to detect whether an AI system discriminated based on demographic factors or any proxies strongly correlated with them.

We also support DJ Patil *et al.*'s recommendation for a check that tests for the error rates of the AI system when applied to different subgroups of the target population.[44] Disparate error rates might result, such as when certain facial-recognition algorithms had much higher error rates

[44] See Chapter 2 in DJ Patil, Hilary Mason, Mike Loukides (2018), *Ethics and Data Science*, O'Reilly.

for individuals with darker skin.[45] Depending on the context in which facial recognition is applied, this could lead to misidentification of criminal suspects, or the malfunction of secure access technology based on facial recognition, all disproportionately problematic for darker-skinned people.

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

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

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

[3.22] An additional possible measure to enhance explainability is to make part of the AI system available for testing by the public as a black box, so that members of the public can query the system, read responses, and test for fairness.

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

3.24 Practices that promote traceability include:

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

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

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

3.25 Organisations should establish an internal policy and process to perform **regular model tuning** to cater for changes to customer behaviour over time and to refresh models based on updated training datasets that incorporate new input data. Model

[45] Larry Hardesty (2018), "Study finds gender and skin-type bias in commercial artificial-intelligence systems", *MIT News Office*. Retrieved on 21 June 2019 from https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

tuning may also be necessary when commercial objectives, risks, or corporate values change.

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

[3.26] It is important to monitor the deployment of AI systems, even those that keep humans out of the loop. It is already standard practice for businesses to monitor the real-time running of their software systems, especially when runtime failures would incur huge costs to individuals or other companies depending on the systems. Such practices should be extended to AI systems. For instance, instrumentation could be built into AI systems to automatically report certain indicators at key moments to allow humans or other systems to verify that the system is running as intended. Such indicators could include the distribution of outcomes for different demographics or other categories.

The failure modes of an AI system should be identified, and contingency measures should be put in place to mitigate those failure modes. Some contingency measures could be executed by humans, such as "Andon Cords" that allow human operators to shut down AI systems.[46] Others could be executed autonomously, like a monitoring program that shuts down an AI system if key indicators exceed allowed ranges; for example, an AI system might be shut down automatically if it begins to significantly privilege one demographic group above another, or begins to diverge sharply from the past behaviour of the system itself, or its antecedents. Other than shutting down the system, companies could revert to earlier stable versions of an AI system, or switch to a non-AI system to perform the same tasks.

In addition, companies should have standard procedures for investigating into algorithm or model failures, such as by hiring relevant consultants or building in-house capacity. Companies should have frameworks which guide the assignment of responsibility for failures given the outcomes of investigations.

[46] Six Sigma Daily, "What is an Andon Cord". Retrieved on 22 June 2019 from https://www.sixsigmadaily.com/what-is-an-andon-cord/

Customer Relationship Management

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

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

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

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

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

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

3.33 **Human-AI interface**. Organisations should test user interfaces and address usability problems before deployment, so that the user interface serves its intended purposes. Individuals' expectations can also be managed by informing them that they are

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

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

a. Degree of risk/harm to the individuals.

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

c. Availability of alternative decision-making mechanisms.

d. Cost or trade-offs of alternative mechanisms.

e. Complexity and inefficiency of maintaining parallel systems.

f. Technical feasibility.

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

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

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

[3.36a] We advocate for the Model Framework to include mechanisms for consumers to challenge decisions made by AI systems that affect them. This is supported by both the *OECD AI*

Recommendations[47] and *MAS FEAT Principles*.[48] We further recommend that consumers should not only be allowed to challenge AI-made *decisions*, but also other types of outputs that affect them, such as translations, predictions, recommendations, and so on.

b. **Decision review channel**. Apart from existing review obligations, organisations can consider providing an avenue for individuals to request a review of material AI decisions that have affected them. Where a decision is fully automated, it is reasonable to provide an individual review by a human agent upon request, if the impact of the decision on the individual is material. However, should it be partially automated with review prior to confirming the decision, the decision has already been reviewed by a human agent. In the latter scenario, this would be no different than a non-AI decision.

[3.36b] Even if a decision was reviewed by a human before confirmation ("human-in-the-loop"), the options generated by an AI system for the human to choose from may not be the options that a human would generate when making the decision without the help of AI. For example, semi-autonomous vehicle may lead to driver complacence, as tragically demonstrated by the Uber car crash in 2018[49], resulting in higher accident risk that cannot be attributed to the driver alone.

It is thus difficult to clearly demarcate how much involvement by a human would absolve the decision from requests to be reviewed, and we warn against the categorical denial of such reviews by the Model Framework. A pertinent factor in deciding if a "human-in-the-loop" decision can be absolved from review is whether the AI that assisted the decision-making is explainable, and whether the human understood how the AI generated the options it presented.

Conclusion

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

[47] See paragraph 1.3(iv) in OECD (2019), *Recommendation of the Council on Artificial Intelligence*. Retrieved on 16 June 2019 from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
[48] See point 10 of the Summary of Principles in MAS (2019), "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector". Retrieved on 16 June 2019 from http://www.mas.gov.sg/News-and-Publications/Monographs-and-Information-Papers/2018/FEAT.aspx
[49] Laura Bliss (2018), "Behind the Uber Self-Driving Car Crash: a Failure to Communicate", *CityLab*. Retrieved on 22 June 2019 from https://www.citylab.com/transportation/2018/05/behind-the-uber-self-driving-car-crash/561230/

ANNEX A
Algorithm Audits

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

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

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

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

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

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

[4.2] We recommend that specific triggers be identified that could lead to algorithm audits, such as if the AI system began to grossly privilege one demographic group over another.

ANNEX B
Glossary

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

5.2 On Accuracy:

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

5.3 On Explainability:

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

5.4 On Fairness:

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

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

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

5.5 On Human Centricity and Well-Being:

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

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

[2] These include Institute of Electrical and Electronics Engineers (IEEE) Standards Association's *Ethically Aligned Design* (https://standards.ieee.org/industry-connections/ec/ead-v1.html), Software and Information Industry Association's *Ethical Principles for Artificial Intelligence and Data Analytics* (https://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2017-11-06-160346-990) and Fairness, Accountability

and Transparency in Machine Learning's *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms* (http://www.fatml.org/resources/principles-for-accountable-algorithms). They also include feedback from the industry in previous rounds of consultation.

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

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

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

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

5.6 On Responsibility, Accountability and Transparency:

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

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

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

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

5.7 On Human Rights

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

5.8 On being Sustainable

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

5.9 On being Progressive

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

5.10 On Auditability

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

5.11 On Robustness and Security

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

5.12 On Inclusivity

a. Ensure that AI is accessible to all.

ANNEX C

Use Case in Healthcare – UCARE.AI

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

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

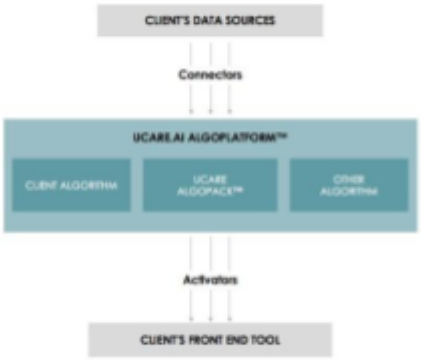
Background

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

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

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

High-Level Architecture of APACHE API



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

Aligning with PDPC's Model AI Governance Framework

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

Trustworthy and Verifiable

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

[Trustworthy and Verifiable] We commend the inclusion of this principle into the evaluation of this case study. However, it does not appear in the guiding principles of the Model Framework or the rest of the document. Given that this case study implies that trustworthiness and verifiability are important aspects for the implementation of AI systems, we strongly recommend that they be mentioned in the main text of the Model Framework, either as part of the Guiding Principles, or in the 'Determining AI Decision-Making Model' or 'Operations Management' sections. This would also ensure coherence of the main text with this case study.

Accountability and Transparency

Prior to data collection, informed consent from stakeholders would have been obtained and approval of the use of data sought via open communication channels. The careful curating and conversion of data into usable format prior to building the models ensures the AI algorithm is kept accountable and coherent to users; this is done in conjunction with Parkway Pantai. The proper storage and repair of previously broken or missing data also serve to provide greater transparency and safety to users by minimising the influence of data gaps in the projection of the result. Careful monitoring of data is key in ensuring service reliability, and therefore detailed and consistent logging across the multiple

components involved is also employed in APACHE, collected in a secure, centralised log storage that is made easily accessible to the development and operations team when required, allowing for prompt debugging and uptime tracking if necessary.

Fairness

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

Human-Centric

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

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

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

[Annex C] We support illustrating the Model Framework with a "use case" scenario that demonstrates how a company could apply the Model Framework to its operations. However, we feel that the justifications of how the UCARE.AI use case promotes fairness and human-centricity are unsatisfactory, and could leave the impression that companies could label their AI solutions as "fair" or "human-centric" without more substantial safeguards. Other important parts of the Model Framework have also gone unmentioned.

The use case explains that "automated prediction... reduces the likelihood of human biases affecting the ultimate judgement of the data and provides an element of consistency across all predictions." Almost every AI system that automates predictions could say the same, but by no means does that necessarily promote fairness. In fact, many of the issues in AI fairness centre around AI systems that claim to remove human bias but instead institutionalize discrimination through biases in training data or through other ways. Indeed, one major issue is the amount of faith placed in AI systems to be "unbiased" because they are "scientific" and "objective", when the bias simply has more subtle and insidious ways of infiltrating an AI system. We should not see statements that automation reduces human bias as any guarantee of fairness.

The use case does not explain how "Discrimination based on income levels and insurance coverage... would be effectively negated." There is no justification that the AI system might not discriminate in that way. It is unclear whether income levels and insurance coverage are included in the data; if they are, there is potential for the AI system to discriminate based on those factors, but even if they are absent from the data, discrimination based on highly-correlated proxy variables has to be tested for and safeguarded against.[50] The use case mentions "basic data validation" and "verification" but does not elaborate on how statistical bias in data is mitigated, according to the Model Framework or otherwise. We have also explained how prejudice in an AI system may arise even when the data is statistically unbiased.[51] Almost none of the measures in the "Data for Model Development" section were alluded to.

The sentence "Although there would be concerns about the use of a 'human out-of-the-loop' [*sic*] system, the algorithm in question is designed to be human-centric" seems to suggest that a system being human-centric can justify putting the human out of the loop in decisions. We think that this is an inaccurate application of the "human-centric" principle, because the degree to which a human is in the loop is an operational decision of the company that does not directly impact whether the overall AI system promotes the well-being of humans or protects their interests—that is, promotes "human-centricity". More precisely, just because a human vets all the decision-making of an AI system doesn't mean that the overall system promotes the well-being and interests of humans; conversely, just because an overall system promotes the well-being and interests of humans, the same might not hold true if a human were to be taken out of the loop.

Other than fairness and human-centricity, other parts of the Model Framework have also been left out of this use case. The use case does not touch on the internal governance structure of UCARE.AI: no personnel or positions are explicitly designated with responsibilities for AI, and no training for employees on AI implementation or ethics is mentioned. There is no comment on a "policy for explanation" (paragraph 3.32), which could affect whether patients understand how cost estimates translate to actual medical costs. There are also no references to the communication channels advocated for in paragraph 3.36.

To conclude, we feel that a more detailed description of a use case is necessary to show how the scenario is truly aligned with each part of the Model Framework. This is required to set a good

[50] See our comment after paragraph 3.22b.
[51] See our comment after paragraph 3.16c.

example for companies, to show that a deeper knowledge of company operations and governance, and their ethical implications, is necessary to actually align with the Model Framework.

ACKNOWLEDGEMENTS

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

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

END OF DOCUMENT

Contributors

Non-Profit Working Group on AI

Cheng Herng Yi
Doctoral Candidate, Department of Mathematics, University of Toronto

Tan Zhi Xuan
Board Member, Effective Altruism SG
Doctoral Candidate, MIT Electrical Engineering & Computer Science

Loke Jia Yuan
Research Associate, Centre for AI and Data Governance, SMU

Jeremy Osborn
Core Lead, DataKind SG

Chan Wai Mun, Raymond,
Chapter Lead, DataKind SG

Pooja Chandwani
Board Member, Effective Altruism SG

Paul Amazona
Core Lead, DataKind SG

Joh Kersey Stapleton
Member, Effective Altruism SG

Zeng Wanyi
Board Member, Effective Altruism SG