ANNEX A

FOR REFERENCE: A COMPILATION OF EXISTING AI ETHICAL PRINCIPLES

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

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

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

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

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

5. **Fairness:**

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

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

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

[12] These include the Institute of Electrical and Electronics Engineers ("**IEEE**") Standards Association's *Ethically Aligned Design* (https://standards.ieee.org/industry-connections/ec/ead-v1.html), Software and Information Industry Association's *Ethical Principles for Artificial Intelligence and Data Analytics* (https://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2017-11-06-160346-990) and Fairness, Accountability and Transparency in Machine Learning's *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms* (http://www.fatml.org/resources/principles-for-accountable-algorithms). There is also the European Commission's *Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions - Building Trust in Human-Centric Artificial Intelligence* (https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58496), and the OECD's Recommendation of the Council on Artificial Intelligence (https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449). They also include principles raised through consultation feedback from the industry.

6. **Human Centricity and Well-being:**

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

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

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

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

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

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

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

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

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

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

10. **Responsibility, accountability and transparency:**

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

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

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

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

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

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

ANNEX B

ALGORITHM AUDITS

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 (as part of a forensic investigation) 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. Ultimately, algorithm audits should normally be used when it is reasonably clear that such an audit will yield clear benefits for an investigation.

2. The following factors may be relevant when considering 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 vary. When the audience consists of **individuals**, providing information on the decision-making process and/or how the individuals' data is used in such processes will achieve the objective of explainable AI more efficaciously. When the audience consists of **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 operation.

   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. For example, the algorithm may be a trade secret or may embody business rules that are trade secrets. If a technical audit is contemplated, corresponding mitigation measures should also be considered (e.g. non-disclosure agreements).

ACKNOWLEDGEMENTS

The PDPC expresses its sincere appreciation to the following individuals and organisations for their valuable feedback to the Model Framework (in alphabetical order):

A*STAR

Accenture

AIG Asia Pacific Insurance Pte. Ltd.

Apple

Asia Cloud Computing Association

AsiaDPO

BSA | The Software Alliance

Cambrian AI

CUJO.AI

Data Synergies

DBS

Element AI

Emerging Technologies Policy Forum

Facebook

Fountain Court Chambers

Google

Grab

Great Eastern

GSK

IBM Asia Pacific

LawTech.Asia

Mastercard

Microsoft Asia

MSD International GmBH (Singapore branch)

Non-Profit Working Group on AI

OCBC Bank

PwC

pymetrics

Salesforce

Singtel

Standard Chartered Bank

Suade Labs

Symphony AyasdiAI

Telenor Group

Temasek International

Tookitaki

UCARE.AI

Untangle AI

#SGDIGITAL

Singapore Digital (SG:D) gives Singapore's digitalisation efforts a face, identifying our digital programmes and initiatives with one set of visuals, and speaking to our local and international audiences in the same language.

The SG:D logo is made up of rounded fonts that evolve from the expressive dot that is red. SG stands for Singapore and :D refers to our digital economy. The :D smiley face icon also signifies the optimism of Singaporeans moving into a digital economy. As we progress into the digital economy, it's all about the people - empathy and assurance will be at the heart of all that we do.

BROUGHT TO YOU BY

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

www.imda.gov.sg

pdpc PERSONAL DATA PROTECTION COMMISSION SINGAPORE

www.pdpc.gov.sg