Formulation of AI Governance and Ethics Framework to Support the Implementation of Responsible AI for Malaysia

By

Aini Suzana Ariffin
Perdana Center STI Policy, Razak Faculty of Technology and Informatics
Email: ain.isuzana.@.utm.my

Mathew Maavak
Science, Technology, Engineering & Innovation Policy Asia and the Pacific Network (STEPAN), 902-4, Jalan Tun Ismail, 50480 Kuala Lumpur, Malaysia

Rozzeta Dolah
Engineering & Technology Department, Razak Faculty of Technology, and Informatics, Universiti Teknologi Malaysia, UTM KL, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia

Mohd Nabil Muhtazaruddin
Naglus Industries Sdn. Bhd. (1359337-D), i-13, i-Cube Incubator, Block J3, Level 2, Universiti Teknologi Malaysia, UTM KL, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia

Abstract

The Malaysian Government, through the Ministry of Science Technology & Innovation (MOSTI), has taken active measures to enhance the adoption of Artificial Intelligence (AI) via the development and implementation of the Malaysia National Artificial Intelligence Roadmap (2021-2025). AI technologies have the potential to increase general productivity, solve manufacturing and operational issues, increase standards of living as well as boost national economic growth. For the purpose of this study, AI is defined as “a suite of technologies which enable machines to demonstrate intelligence and is endowed with an ability to adapt to new circumstances in order to amplify human ingenuity and intellectual capabilities through collective intelligence across a broad range of challenges". This definition comports with the terminology used by Malaysia’s AI roadmap. While new AI technologies have the potential to create immense socio-economic benefits, they may also be prone to unanticipated negative ramifications. With AI gaining rapid adoption across many sectors, there is a critical need for an ethical framework to guide its development, particularly since current social systems are internally fragile and sensitive to external influences and risks. Based on UNESCO’s AI ethics recommendations and a desiderata of relevant data, AI’s adoption also raises a number of questions over its ethical impacts over broad areas ranging from technology to social behavior. Additionally, the dynamic and imbalanced nature of the global trade and business structures has given rise to a variety of complex and systemic problems. These may significantly impact vulnerable links in global systems which incorporate AI. Therefore, there is an urgent need for an AI governance system and ethical framework that will help materialize the development of AI in general, and the goals of the national AI Roadmap in particular. The AI ethics and governance framework should oversee the growing scales and complexities of interconnections between relevant stakeholders as well as manage emerging uncertainties and risks in the socioeconomic and Industry 4.0 realms. This study has adopted qualitative and quadruple helix approaches. A series of interviews involving 30 respondents from government agencies, academia, industries and civil society were conducted over a period of four months in the capital city of Kuala Lumpur and the administrative capital of Putrajaya. Subsequently a benchmarking exercise on Governance and Implementation of AI Principles was undertaken with five government agencies and four universities in Tokyo and Kyoto, Japan. The paper's research outcome advocates a more dynamic approach to the formulation of an appropriate AI ethics and governance framework that will maximize the adoption of AI among various Malaysian stakeholders e.g., MOSTI, high tech companies, researchers and consumers. The framework can also be incorporated into the ASEAN AI Ethics and Governance Framework (AI-EGF) for regional implementation and crisis management.

Acknowledgement

This paper was made possible through a research grant from the Toshiba International Foundation Organization (TIFO). Research Title "Formulation of Ethics and Governance Framework to Institutionalize Artificial Intelligence (AI) Adoption among stakeholders through smart partnership with Japanese consortium in enhancing Execution of the Malaysia AI-Rmap". Grant no: R.K130000.7356.4B753 with reference no: PY/2022/00641.

Keywords: Artificial Intelligence (AI), Malaysia National Artificial Intelligence Roadmap, AI Governance, Responsible AI; AI Principles; AI Ethics; AI Ethics Framework

1. Introduction

Malaysia's economic development was facilitated by the country's STI capability and capacity development. The development and implementation of the National Policy on Science, Technology, and Innovation (NPSTI) 2021-2030, the 10-10 Malaysia Science, Technology, Innovation, and Economy (MySTIE) framework, and the Malaysia Digital Economy Blueprint (2021-2030), all focus on leveraging STI for wealth creation, socioeconomic development and achieving Sustainable Development Goals (SDGs). Artificial Intelligence (AI) was recognized as a key driver in the MySTIE framework as well as one of five foundational technologies penciled by the National Fourth Industrial Revolution (4IR) Policy. This 4IR policy, which focuses on multifaceted connectivity, machine learning, automation, and real time data, regards AI acts as a catalyst for economic growth, especially in the agriculture, manufacturing, education, financial and business sectors.

AI is no longer a futuristic concept. During the global COVID-19 pandemic from 2020-2022, AI played a major role in deploying autonomous vehicles that could deliver laboratory samples for rapid processing and testing in the United States alone. AI and its deep learning abilities to predict probabilities of old and new drugs or treatments has been credited for curbing the spread of the SARS- CoV-2 virus. Drones and AI-driven robots also assisted in minimizing contacts between humans. Business entities which leveraged AI's capabilities soon surged ahead in a highly-competitive market. AI-led industries may eventually contribute up to USD15.7 trillion to the global economy by 2030 (PwC's Global Artificial Intelligence Study, 2021). In Malaysia, AI-led industries are expected to increase the national gross domestic product by 1.2%.

Recognizing the importance of AI for economic growth and sustainable development, the government of Malaysia, through the Ministry of Science, Technology and Innovation (MOSTI), introduced and launched the Malaysia National AI Roadmap 2021-2025 (AI-Rmap) in August 2022. AI-Rmap is also aligned with critical national policies such as the Shared Prosperity Vision 2030 and the 12th Malaysia Plan (2021-2025) which emphasizes the role of technological advancements – particularly digitization and public sector service delivery – on nation-building by strengthening overall security, social wellbeing, and inclusivity. The AI-Rmap proposed six strategic initiatives and 22 action plans that will "Make Malaysia a nation where Artificial Intelligence augments jobs, drives national competitiveness, encourages innovation and entrepreneurship to bring economic prosperity, social good and improves people's well-being". The six strategic initiatives are:

1. Establishing AI Governance;
2. Advancing AI R&D;
3. Escalating Digital Infrastructure to Enable AI;
4. Fostering AI Talents
5. Acculturating AI and
6. Kickstarting a National AI Innovation Ecosystem.

AI will radically change our societies. One of example of AI adoption in Malaysia is ELYA, the Employees Provident Fund (EPF)'s chatbot. ELYA is capable of responding to EPF client inquiries rapidly and efficiently by processing a variety of information. Other than that, Malaysia is the home of

A.D.A.M (Accelerated Devices Always Massive), the 88th fastest supercomputer in the world as ranked by the High-Performance Conjugate Gradient (HPCG) 2019. A local company named Twistcode® self- assembled A.D.A.M in 2018 and launched MATA which uses deep learning to accelerate data processing by harnessing AI in sectors such as healthcare, agriculture, transportation, finance, manufacturing, maritime, and oil and gas production. Although AI technologies will generate significant socioeconomic benefits, they are not without attendant risks. Cyberattacks and associated fraudulent transactions and data breaches have been ranked among the Top 10 long-term threats by the WEF Global Risk annual reports.

In 2015, the Future of Life Institute (FLI) published a notable open letter that was signed by the likes of Elon Musk and several top scientists from the Massachusetts Institute of Technology (MIT), University of Harvard, and the Association for the Advancement of Artificial Intelligence (AAAI) on the need to ensure that AI will be solely used for the benefit of humans (Future of Life Institute, 2015). Elsa González Esteban (2022) also warned that AI was a double-edged sword that can both benefit and threaten mankind. An increasing number of tech pioneers are advocating a more sensible approach to AI and its capabilities. They are primarily concerned by the fact that AI is not sufficiently regulated.

Balancing responsible utilization of AI while simultaneously mitigating attendant risks will become a salient task for national policy-makers. Under the Malaysia Digital Economy Blueprint, the government has mapped out a thrust (Thrust 6) to build a trusted, secure, and ethical digital environment for the adoption of new emerging technologies including AI. At the same time, the AI-Rmap has introduced a Responsible AI framework with seven key AI principles. Currently, there are no AI Governance and Ethics Framework (AI-GEF) available to institutionalize the Responsible AI and AI Principles in Malaysia. According to Mintrom & O'Connor (2020), a comprehensive and overarching policy framework will help avoid divergent policy narratives. Policy makers all over the world are aggressively discussing how to develop an appropriate AI Governance and AI Ethics Framework and mitigate the risks in the process. A few developed nations such as the United States, European Union, Australia, Canada, Singapore, Japan and South Korea have developed AI frameworks and guidelines in order to foster the adoption of ethical standards.

Therefore, this study aims to formulate an AI Governance and AI Ethics Framework (AI-GEF) in order to support implementation of Responsible AI in Malaysia. This study developed four stage-by-stage objectives as follows:

1. To gauge perceptions towards Responsible AI in order to enhance AI adoption among national stakeholders
2. To identify barriers faced in adopting AI among stakeholders
3. To benchmark its AI framework, guidelines and standards with the leading Japanese counterpart, particularly with regards to best practices on AI governance and ethics.
4. To Formulate AI Governance and AI Ethics Framework

The framework developed by the Japanese government, namely The Governance Guidelines for Implementation of AI Principles in 2021, was chosen as the ideal benchmark as Japan was consensually seen as the global leader in the area of constructing and introducing a well-rounded framework for the implementation of AI Principles. By addressing AI Governance and AI Principles issues in 2019, Japan has positioned its business sector to be in the driver's seat in this vital area. While the adoption and deployment of AI is advancing rapidly in Japan, so are social problems associated with AI. To address this problem, Japan currently has created a multi-stakeholder and inter-disciplinary network on AI Governance and intends to lead global conversations on the subject matter.

The objectives of the study are structured by semi-structured interviews that are driven by key questions below:

- What are the barriers to adopting AI among stakeholders?
- How will these barriers be addressed?
- How do the Japanese consortium and stakeholders implement AI Principles?
- What are the best practices for AI governance and what constitutes an optimal AI ethics framework (Responsible AI) which can accelerate AI adoption?

This integrated study also addresses the need and demand of a renewed “Look East Policy” as Japan is globally-recognized for its readiness and preparedness in the area of AI Ethics and Governance. A comparative analysis between the AI principles and action plans of Japan and Malaysia will be evaluated and analyzed. The formulation of an AI Ethics and Governance Framework (AI-EGF) will have to conform to ASEAN’s motto of “As One Vision, One Identity and One Community”. Moreover, mainstreaming the AI-EGF into future development planning, investment and management paradigms will improve life for future generations. It will also help meet goals of the ASEAN Community Vision 2025 which intends to transform the region into a high-growth geographic area. Singapore has recently initiated a project on the Development of ASEAN AI Ethics and Governance Framework and Guidelines to kick off the process.

Literature Review

Artificial Intelligence (AI)

Artificial Intelligence allows a computer to mimic the human thought process. AI can easily surpass humans in certain computational aspects that require a massive crunching of data. According to the Internet Society (2017), human programmers provide computers with certain instructions and rules for the machines to “learn”. Based on inferences gained from the data, algorithms are generated which allow machines to provide new rules and information to solve complicated tasks. An algorithm is defined as “a sequence of instructions used to solve specific problems”.

Human intelligence elements include perception, reasoning, learning, language understanding, comprehension, consciousness, alertness, realization, awareness and intuition.

AI capabilities currently cannot mimic the entire gamut of human abilities but they can excel in generalized learning, reasoning and problem solving. Voice assistant software like Siri and Alexa and large language models such as ChatGPT (Generative Pre-Trained Transformer) are prime examples of success stories in this regard.

There are two types of AI: weak AI and strong AI. Weak AI is programmed to perform only one task or role. Amazon Alexa, for example, acts as a voice-centered assistant that can streamline human control over entertainment applications such as music, provide assistance in managing smart homes; and aid virtual personal shopping. This is an example of weak AI. On the other hand, strong AI is expected to mimic human emotions and self-awareness, making it rather unpredictable. Machine learning is a subset of AI that focuses on getting machines to make decisions through a continuous data feedback loop. Meanwhile, a subset of machine learning called deep learning uses the concept of neural networks to solve complex problems. These are all interconnected facets of AI.

There is no commonly agreed definition of AI. According to the European Commission's High-level Expert Group on AI (AI HLEG): “Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to predefined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analyzing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)” (EC, 2018).

Though comprehensive and lengthy, this definition can be somewhat unwieldy.

The Organization for Economic Co-operation and Development (OECD) has updated its definition of AI as following: “An AI system is a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy.” (OECD, 2019)

Taking cue from these evolving definitions, Malaysia’s National AI Roadmap 2021-2025 defines AI as a suite of technologies that enable machines to demonstrate intelligence, the ability to adapt with new circumstances, and which can be used to amplify human ingenuity and intellectual capabilities through collective intelligence across a broad range of challenges.

AI Governance & Ethics

Proper governance of technology, including AI, is required to maximize socio-economic potentials while mitigating uncertainties and risks. Many public and private sector actors regularly meet to discuss principles, values, mechanisms that can regulate and guide the development and adoption of AI technologies. They recognize that governance is a critical prerequisite for the development of Artificial Intelligence.

According to Zhang and Dafoe (2020), AI governance spans the realm of abstract principles to the world of mass politics. Koene et al (2019) stress that the lack of transparency risks undermining meaningful control and accountability, which is a problem when these systems are applied in the context of decision-making processes that can have significant human rights implications. Who is ultimately responsible for decisions made by machines? It can be problematic to attribute such decisions to humans as a machine itself is learning new problem-solving methods based on an ever-aggregating volume of data. It is also becoming more common for humans to be overwhelmed by multiple workloads. In such situations, it can be difficult to decide against an algorithmic suggestion. In this respect the very concept of responsibility is being fundamentally questioned by the new developments at the moral and legal levels. It will also be problematic to focus primarily on the person signing off on a final decision.

Additionally, the difficulties faced in establishing a comprehensive AI governance system can be seen in the light of i4.0's new technologies and business practices. Organizations and societies may not have the means to deal with these developments. Existing institutional arrangements and administrative silos may no longer be relevant. The drone complex and IoTs, for example, can invade privacy while self- driving autonomous vehicles routinely raise safety concerns. Existing financial systems could be challenged by foreign blockchain competitors. All these disruptions have resulted in a plethora of governance gaps and liabilities.

A holistic and flexible governance system is needed to helm AI technologies. It should be encouraging innovation while minimizing negative disruptive developments. Stability and equilibrium among public and private sectors is a critical component of idea AI governance. This viewpoint has been echoed by the United Nations and several scientists in the field (Mikhaylov et al, 2018). Holistic AI governance includes a variety of frameworks, including intergovernmental strategies, non-binding guidelines, principles, qualifications and certifications (including non-profit feedback), coordination and control systems in accordance with government oversight. Weaving them together into a coherent governance mechanism may be a daunting task but the foundations of such an endeavour should not be delayed due to the exponential advancements in machine learning.

The Government AI Readiness Index 2022, published by the Oxford Insights, has recently rated governments based on their readiness to adopt AI in public services delivery. This global index outlined three fundamental pillars. These which are the Government Pillar; Technology Pillar, and Data and Infrastructure Pillar as shown on Figure 1.

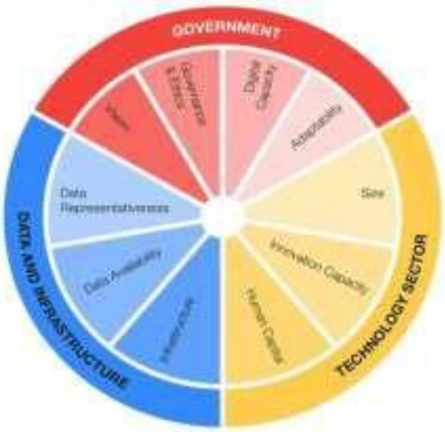


Figure 1: *Three Fundamental Pillars Of Government Ai Readiness Source: Oxford Insight (2022***)**

Governance and Ethics is a primary component of the Government pillar. It predetermines the readiness of a government to adopt AI. Effective AI guidelines should perceive and address evolving ethical concerns. While the United States heads the list of the Government AI Readiness Index 2022, Singapore leads in two out of three pillars established by the index. At the regional level, Western European countries now make up fewer than half of the top 10 countries in the list for the first time as three East Asian countries have advanced into the elite category. In the Government Pillar, AI strategy is dominated by middle income countries while for the Data and Infrastructure Pillar, three East Asian countries top the list as their governments have allocated sufficient budgets to support AI infrastructure and development.

AI Ethics Framework and Implementation

As defined by Nalini (2019), ethics can be denied as moral principles which govern the behavior or actions of an individual or a group. Humans have all sorts of cognitive biases and these are exhibited through behavior. Ethics is therefore a set of moral principles based on shared values that help us determine between the ethical and unethical or between right and wrong. Generally, these principles are simple and universal e.g., “thou shall not steal” but in the AI realm, these principles may have to negotiate between complex or complicated environments. Examples of questionable uses of AI include cyber warfare, weaponized unmanned vehicles and automated propaganda and disinformation campaigns. This also includes the algorithmic manipulation of public perception through social media platforms like Facebook and Twitter (cf. Lazer et al., 2018) and Google’s tendentious search engine optimization (SEO). These platforms enable unmonitored AI-based experimentations on society without informed consent. Another example is mass surveillance which uses a variety of biometric data and tools, including facial recognition software, to pacify a population. (Helbing et al., 2019). China’s social credit scoring system is one such dystopian example.

Maavak (2023) notes that “informed consent to AI-assisted healthcare may be whittled down as patients may either not understand what they are consenting to or they may overly trust an “intelligent and impartial” machine-based system. It gets more ominous when triage needs to be employed in critical situations. Traditionally, triage is determined by the integrity and skills of medical personnel present as well as the resources available in situ. An algorithm however may dispassionately arrange triage based on a patient’s “worth to society”; data-inferred life expectancy; and costs associated with saving the patient’s life.”

ChatGPT is likely to make hundreds of millions of jobs redundant. This will have severe socioeconomic implications. Constant use of AI may also reduce humans into becoming “extensions” of their “machines instead of vice versa” (Maavak, 2023).

To cover all these potential pitfall areas, AI Ethics in this study refers to a set of principles that informs the design and outcomes of AI technology development. The development of an AI-GEF is critical in this regard as it will provide stakeholders with guidance on how to deal with the ethical dimensions of Artificial Intelligence. According to a Capgemini Report (2019), 86 percent of executives across 10 nations claim that they faced ethical issues when adopting AI; 60% of organizations have attracted legal scrutiny; and 22% have faced a customer backlash in the last two to three years because of decisions made by AI systems. Massive pressure to urgently induct AI has also led to a systemic failure to consider ethical implications. Another lacuna noted was the lack of expertise who understand ethical AI systems.

The State of AI 2021 Report emphasizes the criticality of managing AI risks in areas such as cybersecurity, regulatory compliance, personal/individual privacy, explainability, organizational reputation, workforce/labor displacement, equity and fairness, physical safety, national security, and political stability. According to Hongladarom (2021), AI systems must be accountable, responsible, and ethical in facilitating key socioeconomic benefits. Jobin et al (2019) reviewed 84 ethical frameworks and guidelines and proposed 11 principles: 1) Transparency; 2) Justice and Fairness; 3) Nonmaleficence; 4) Responsibility; 5) Privacy; 6) Beneficence; 7) Freedom and Autonomy; 8) Trust;

9) Dignity; 10) Sustainability; and 11) Solidarity. However, they also note the evidence of convergence of these principles into five core principles only namely 1) Transparency; 2) Justice and Fairness; 3) Nonmaleficence; 4) Responsibility; and 5) Privacy.

Ethical technological principles are needed to safeguard humans from negative consequences of digitization as well as guide the development of all such technologies in a better way. Technological development comes with technological problems, and there are also significant ethical issues which these technologies generate. The right mix of values and principles will help technologies handle sensitive confidential information (N. A. Zakaria, Z. Ismail, 2016). Joaquin, et al (2020) stress that decision-making procedures that are not guided by morals and principles or value-based normative judgments may create a significant "philosophical" divide between scientific results and social policies. Many institutions across the world are therefore developing AI ethics frameworks and principles as a prelude to implementing Responsible AI for their nations or organizations. Table 1 showcases some of the more notable undertakings.

Table 1: *Ethic Principle and Ethical Framework*

Institution/Year	Objective	AI Ethics Framework/AI Principle
European Commission’s High-Level Expert Group on AI (European Commission, 2019)	The guideline is designed to guide the AI community in the development and use of “trustworthy AI” (i.e., AI that is lawful, ethical, and robust). This guideline emphasizes four principles; respect for human autonomy, prevention of harm, fairness and explicability.	Respect Agency and oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Fairness, Accountability, Diversity, Societal and Environment
Australia’s Ethics Framework (Dawson et al. 2019)	This ethics framework highlights the ethical issues that are emerging or likely to emerge from AI technologies and outlines the initial steps toward mitigating them. The goal of this document is to provide a pragmatic assessment of key issues to help foster ethical AI development in Australia.	Generates Net-benefits, Regulatory and Legal Compliance, Fairness, Contestability, Do No Harm, Privacy Protection, Transparency and Explainability, Accountability
Institute of Electrical and Electronics Engineers (IEEE, 2019)	The proposed design lays out practices for setting up AI governance structure, including pragmatic treatment of data management, effective computing, economics, legal affairs, and other areas.	Human Rights, Well-being, Data Agency, Effectiveness, Transparency, Accountability, Awareness of Misuse, Competence

Model AI Governance Framework, Singapore, 2019	The purpose of this model is to accommodate critical concerns regarding AI. It is guided by a set of AI principles and actions that can be taken and Implemented. The Model Framework categorizes AI principles into two high- level guiding categories which aim to encourage Trustworthy AI and awareness on leveraging AI technologies.	Accountability, Accuracy, Auditability, Explainability, Fairness, Human Centricity and Well-being, Human Rights Alignment, Inclusivity and Progressiveness
Principles on Artificial Intelligence OECD (2019)	AI Principles here promote use of AI that is innovative and trustworthy and that respects human rights and democratic values. Also set standards for AI that are practical and flexible enough to stand the test of time.	Inclusive Growth, Sustainable Development and Well-being, Human- Centred Values, Fairness, Transparency and Explainability, Robustness, Security and Safety, Accountability
Governance Guidelines for Implementation of AI principles. Japan (2021)	The purpose is to facilitate the deployment of AI. Provide hypothetical examples of implementation corresponding to each action target and practical examples of gap analysis between AI governance goals and current state of AI.	Social Principles of Human-Centric AI: Human-centric, Education/Literacy; Privacy Protection; Ensuring Security; Fair Competition, Fairness, Accountability and Transparency; and Innovation
Recommendations of AI Ethics. (UNESCO, 2021)	The proposed ethical principles aim to provide decision-makers with criteria set that extends beyond purely economic considerations.	Human Dignity, Value of Autonomy, Value of Privacy, “Do no harm” Principle, Principle of Responsibility, Value of Beneficence, Value of Justice

Consolidated by Authors (2023)

In order to coordinate the European Union Member States’ National AI strategies, the Council of the European Union endorsed the European Commission’s Coordinated Plan on Artificial Intelligence in 2018. Subsequently, in April 2019, the Commission published the Ethics Guidelines for Trustworthy AI. The guidelines focus on two components: Ethical AI and Robust AI.

The foundations of Trustworthy AI were laid out by outlining a fundamental, rights-based approach where AI Principles will ensure Ethical AI and Robust AI. The AI Principles are respect for human autonomy; prevention of harm; fairness, and explicability. These AI Principles were translated into seven non-exhaustive key requirements that should be implemented throughout the life cycle of AI systems. These key requirements are evaluated and addressed continuously using technical methods and non-technical methods. Practical guidance for the AI practitioners is also laid out in the form of an assessment list for Trustworthy AI implementation. The assessments were tailored specifically to AI applications. The European Union also recognized recommendations published by the Organization for Economic Co-operation and Development (OECD) i.e., the OECD Principles on Artificial Intelligence 2019. The realization of Trustworthy AI is shown on Figure 2. Figure 3 shows how AI Systems are linked to AI Principles.

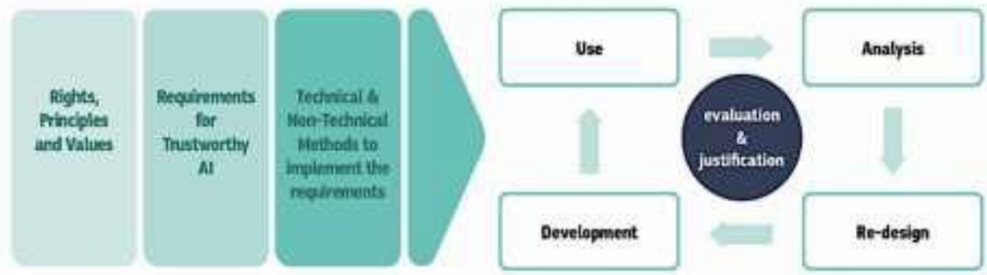

Rights, Principles and Values
Requirements for Trustworthy AI
Technical & Non-Technical Methods to implement the requirements
Use
Analysis
evaluation & justification
Development
Re-design

Figure 2: *Realizing Trustworthy AI throughout a system's entire life cycle (Source: EU's Ethics Guidelines for Trustworthy AI)*

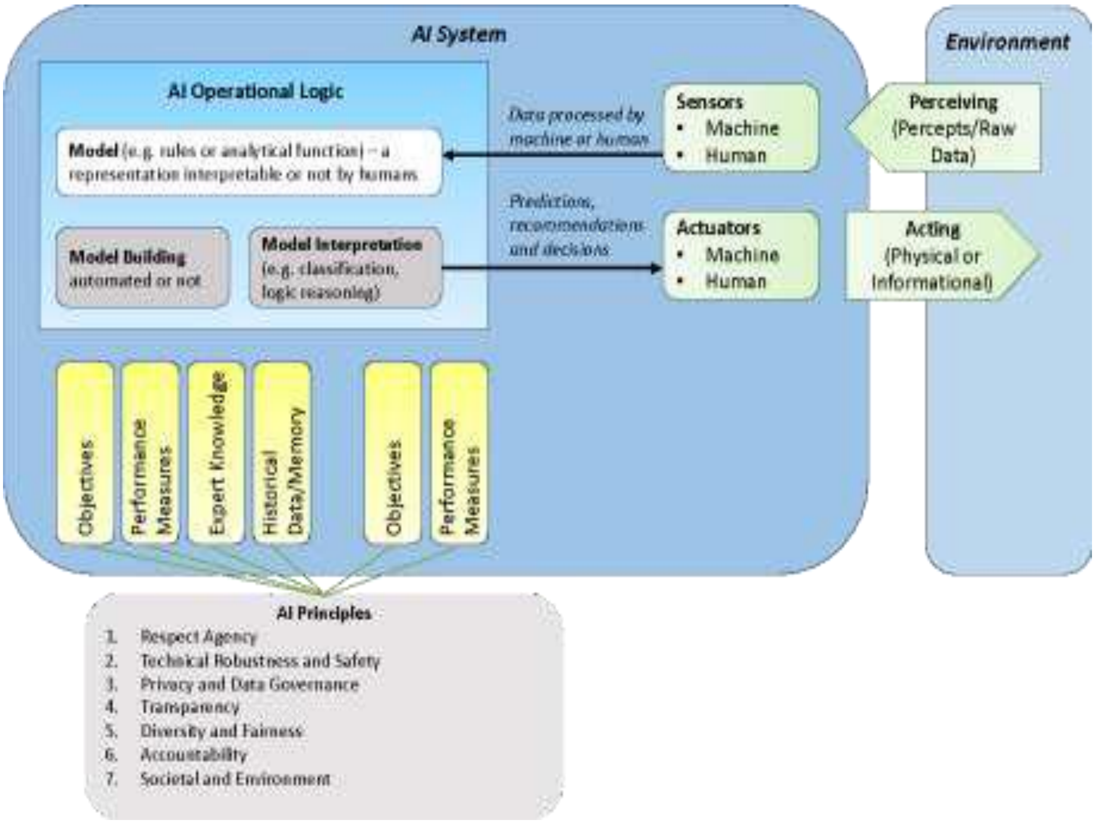

AI System
Environment
AI Operational Logic
Model (e.g. rules or analytical function) – a representation interpretable or not by humans
Model Building automated or not
Model Interpretation (e.g. classification, logic reasoning)
Data processed by machine or human
Predictions, recommendations and decisions
Sensors
Machine
Human
Actuators
Machine
Human
Perceiving (Percepts/Raw Data)
Acting (Physical or Informational)
Objectives
Performance Measures
Expert Knowledge
Historical Data/Memory
Objectives
Performance Measures
AI Principles
1. Respect Agency
2. Technical Robustness and Safety
3. Privacy and Data Governance
4. Transparency
5. Diversity and Fairness
6. Accountability
7. Societal and Environment

Figure 3: *Linking the AI System to the AI Principles (Source: OECD Principle on AI, 2019)*

AI Governance & Ethics in Malaysia

The government of Malaysia, via the Economic Planning Unit (EPU) under the Prime Minister's Department, launched several policies on digitalization and advancement of technologies. These include the National Fourth Industrial Revolution (4IR) Policy in 2019, 12th Malaysia Plan (2021-2025) and Malaysia Digital Economy Blueprint (MyDIGITAL) in 2021. Concurrently, the National Science, Technology and Innovation Policy (NSTIP) 2021-2030 was launched by the Ministry of Science, Technology and Innovation (MOSTI). All these strategic policies play a crucial role in driving a technology-based economy in Malaysia. It is also intended to support the Malaysia Shared Prosperity Vision 2030 policy which emphasizes the well-being of the people. Subsequently, the Malaysia National Artificial Intelligence (AI) Roadmap 2021-2025 (AI-Rmap) was unveiled by MOSTI to specifically direct the adoption of AI in Malaysia by ensuring a sustainable AI innovation ecosystem. There are seven (7) AI Principles for Responsible AI outlined in the AI-Rmap. Table 2 and 3 explains the objectives and descriptions of all seven AI principles. Figure 4 shows the Principle of Responsible AI and outlines the description of each principle.

Table 2: *Ethics Principle and Ethical Framework*

Institution/Year	Objective	AI Ethics Framework/AI Principle
Responsible AI and Principles of AI (MOSTI 2022)	The purpose of the framework is to institutionalize Responsible AI via adoption and stakeholder participation.	7 AI Principles: Fairness; Reliability, Safety and Control; Privacy and Security, Inclusiveness; Transparency, Accountability, Pursuit of Human Benefit and Happiness

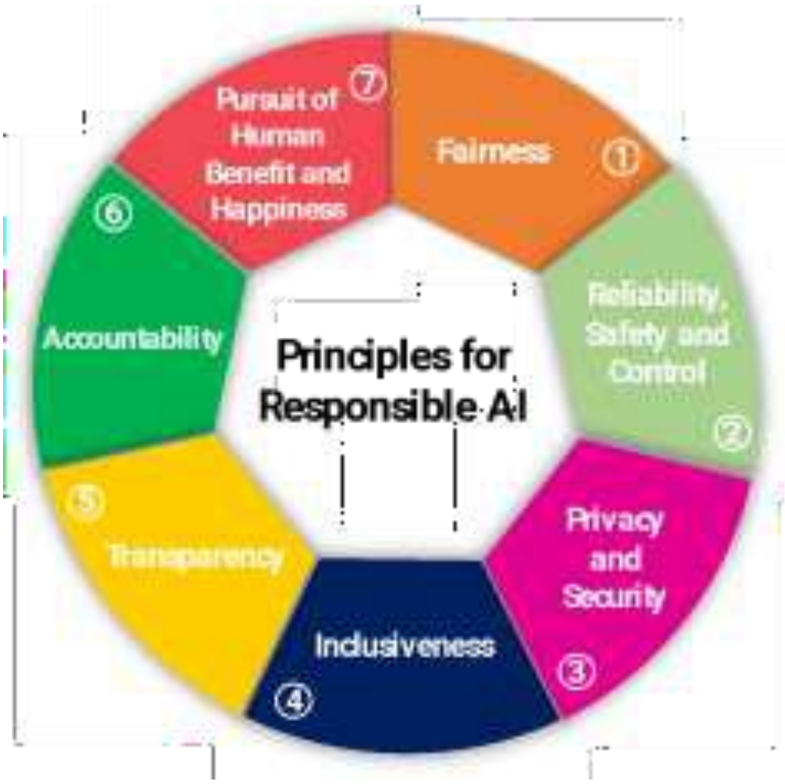

Principles for Responsible AI
Fairness ①
Reliability, Safety and Control ②
Privacy and Security ③
Inclusiveness ④
Transparency ⑤
Accountability ⑥
Pursuit of Human Benefit and Happiness ⑦

Figure 4: *Principles for Responsible AI*

During the formulation stage, a few significant questions were raised over AI risk mitigation. These questions also gauged perceptions towards the implementation of Responsible AI and the seven AI Principles among stakeholders. Table 3 presents the description of each principle and the questions raised. In order to gauge the effectiveness of the proposed AI governance system and its implementation as well as the impact of responsible AI on stakeholders, it was suggested to conduct an Impact Assessment exercise. Several Impact Assessment questions also were raised during the formation of the AI-GEF. They are broadly outlined below:

- Any periodical audit or revision by any parties/government for AI?
- Does any government agency carry out Impact Assessments of AI Principles?
- Which agency will develop and monitor Ethical Impact Assessments, if any?
- Who is responsible for AI monitoring and surveillance?
- How to understand the positive and negative impacts of AI?
- Is there any Data Ethics Centre being established?
- Are there any Measuring Indices/Principles for Responsible AI?

Table 3: *Ai Principles Descriptions and Questions*

AI Principle	Description	Significant Questions
1. Fairness	It is essential that AI does not limit opportunities for anyone – fairness is the foundation for treating people with dignity and respect. AI systems must provide guidance on medical treatment, loan applications or employment and these should make the same recommendations to everyone with similar symptoms, financial circumstances, or professional qualifications without bias.	How to quantify Fairness in AI practice? Do AI fairness/bias issues that concern developers revolve around datasets, or algorithms/models, or both, and how are the problems different for each? What methods are being developed to define, detect and correct potential bias? Are there any standardization undertakings for multiple AI fairness guidelines, best practices, tools and platforms? How can we ensure that the datasets used for AI training are fair and balanced?
2. Reliability, Safety & Control	AI systems should perform reliably and safely. The complexity of AI technologies has fueled fears that AI systems may cause harm in the face of unforeseen circumstances, or that they can be manipulated to act in harmful ways. Trust in AI systems will depend on whether they can be operated reliably, safely, and consistently even under unexpected conditions, especially for applications in fields affecting both lives and livelihoods such as healthcare, financial and other services where consequential decisions are involved.	How Reliable is Artificial Intelligence? “Is today’s AI technology really as reliable and world- changing as we imagine?” Can AI be dangerous? (Views and opinions) What are the main threats of AI reliability? The cyber security of AI – who controls AI Safety?
3. Privacy & Security	People will not want to share their data if they do not believe that it will be stored securely, used safely, and for a good end. It is essential that AI systems comply with applicable privacy laws i.e. on the collection, use and storage of data. The systems must be designed to protect personal data from bad actors who may steal private information or inflict harm.	How can we protect privacy as automated systems increasingly track us? How can we control personnel data that will be used for AI training? What is the state of AI in security? How do we embrace the use of AI technology but still preserve our rights to privacy? What constitutes invasion of privacy? What legal protections apply to AI have and under which Act?

4. Inclusiveness	AI systems should benefit everyone and address a broad range of human needs and experience, inclusively. For example, these technologies can become tools of empowerment for people who are physically or cognitively disabled, enabling them to gain access to opportunities that they may not have had before, in education, employment, and citizen services, thereby improving their overall health, socioeconomic situation, quality of life, and participation in society.	What can agencies/companies do to build more inclusive AI? How to ensure AI-related diversity and inclusiveness? What is the measurement of AI inclusivity?
5. Transparency	Transparency is crucial as otherwise it leads to suspicion and reluctance. The Malaysian public places significant value in organizations being transparent about what they do with people's data. Compared to the global average, Malaysians are more receptive to their data being used by organizations – both private and government – and they want to understand the risks involved.	What are the examples of transparency in AI? Is transparency a challenge in AI? Why is transparent AI important? How to ease the AI bureaucracy to ensure its transparency? Transparency between AI Facilitator-Developer-AI operator-AI user?
6. Accountability	People who design and deploy AI systems must be accountable for how their systems operate. To establish norms and best practices, we can draw upon experience in other sectors such as healthcare. Internal review boards can provide oversight and guidance on which practices should be adopted during the development and deployment of AI systems.	Accountability in Artificial Intelligence: What is it and how does it work? Who's responsible when AI goes wrong? How to Build Accountability into your AI system? What is the Control Plan of AI accountability?
7. Pursuit of Human Benefits & Happiness	AI is first and foremost a tool. The purpose and objective of this tool is to promote human well-being. By enshrining the goal of elevating human happiness and quality of life in our national AI Ethics charter, we can begin addressing each AI principle in order to solve people's problems and improve the overall quality of life in the nation.	How is AI being used for the benefit of humanity? How is AI helping us today? Will AI pose a particular threat to humans? Is there a Happiness Index of AI being developed? How to overcome the shortage of employability due to AI displacement of humans? AI vs humans: what are the dimensions?

Developed by Authors 2023

Research Methodology

Research methodology is an important aspect of any research project. The selection of a methodology depends on the research objectives, questions and the nature of data required. It helps ensure that the data gathered is reliable and valid. In this study, the methodology also provided guidance on the ethical considerations involved in research and how to communicate research findings effectively. This study adopted a qualitative research mode via three main approaches. They were an Elite Interview session (involving top or senior management position), Focus Group and Benchmarking. Qualitative research methods are useful when exploring complex phenomena and require an in-depth understanding of the subject. The Sage Handbook of Qualitative Research by Denzin and Lincoln (2018) highlights the importance of using a wide range of data collection techniques to capture the richness of the data. The research design commenced with gathering of secondary data such as existing National AI Policies and strategies, Malaysia's National AI Roadmap 20021-2025 and other related Malaysia national policies. Global and select country reports and documents related to AI as well recent peer-reviewed articles were also studied. These included articles on the Principles for Responsible AI, AI Governance and AI Ethics Framework, AI Technology Adoption and Performance measure of AI Ethics.

A sample, a subset of a population, representing elements of the national population were selected to participate in the study. This sampling offered the opportunity to study actual experiences and to make an in-depth investigation of the population through an exploratory process. Therefore, purposeful sampling techniques were utilized in the selection for interviews and benchmarking. Participants had to be appropriate experts with well-developed views on the research topic. The selection of experts for this interview was based on a Quadruple-Helix engagement comprising Government authority body (policy maker), Industry, Academics and Society; qualifications; and experience in managing and handling AI projects or matters. This study therefore presented a unique opportunity to align national and global agendas on AI governance and ethics adoption among stakeholders. As this study was also aimed at exploring best practices on Responsible AI implementation, the Japanese government and relevant agencies and universities were also selected as participants. Japan was seen as a champion in AI readiness and preparedness at the policy making and implementation levels.

There were seventy-five (75) individuals who participated in this study. Interviewees included ten (10) senior officers from government agencies, including lawyers and policy makers, eight (8) industry experts from different sectors, ten (10) senior lecturers and researchers, two (2) senior librarians from universities and two (2) heads of associations. The study also included round table discussions with eight (8) senior officers from Malaysian government agencies, namely Malaysia External Trade Development Corporation (MATRADE), Malaysia Investment Development Authority (MIDA), Civil Services Department (JPA), Prime Minister's Office and the Malaysian Embassy in Tokyo, Japan.

On the Japanese side, twenty (20) respondents contributed views and ideas during the benchmarking exercise. Ten (10) comprised of five (5) directors and three (3) senior management members from government agencies namely the Japan International Cooperation Agency (JICA), the Japan Science and Technology Agency (JST), RISTEX, Research Institute of Science and Technology for Society, and the Japan External Trade Organization (JETRO) as well as five (5) individuals from Mitsubishi Electric Corporation, Tokyo Gas Group and AI & Naglus Inc in Tokyo and five (5) Professors from Meiji University, Keio University and Kyoto University.

Based on all data gathered, a Focus Group was also conducted with MOSTI in October 2022 at Bangi, Selangor. There were fifteen (15) respondents who participated, including four (4) senior officers from MOSTI, six (6) senior members from other government departments namely the Prime Minister's Office, Malaysian Administrative Modernization and Management Planning Unit (MAMPU), representatives from the Attorney General of Malaysia, Malaysia Digital Economy Corporation (MDEC), MIMOS Berhad and four (4) individuals from industries and one (1) lawyer from a local university.

All interview sessions, benchmark, round-table and focus group took place face-to-face were bilingual either Malay and English, fully transcribed, coded and drafted in the written manuscript in English. During the interviews, other considerations were adhered to ensure it accordance with ethical research standards such authorization for access to venues, time, recording interviews and taking photographs, request of additional relevant documents as well as note-taking. In every social science research study, the content of the interview questions should be validated by experts in order to obtain good interview content. However, during interviews, although respondents have expressed their views freely, they might be affected by bias. Thus, to increase reliability, the details of interviews need to be validated through the transcription process. In addition, supporting evidence from other sources or cross-checking their view with evidence from the literature and expert judgement also help to minimize bias.

This study adopted Thematic and Content Analysis methods for data gathering and analysis. Immediately after each interview session, audio recordings were transcribed into verbatim texts to facilitate analysis. According to Tan (2018), transcription helps identify the main concepts in a conversation for further analysis. It also makes researchers aware of participant behaviors during discussions, such as moments when conversations get heated. After transcribing, data patterns were studied, with repeated issues described by respondents and diverse codes emerging from the findings.

These findings led to the formation of a data pattern obtained from the first interview which was then used as a guideline for collecting data for the second and subsequent interviews. The codes developed were assigned based on the text, images and recording. The list of code categories and descriptions were finalized into a codebook. According to Creswell (2018), the researcher should start with open coding for related significant information in the first interview analysis. This is an essential initial step before moving to the second interview analysis stage. The existing open codes from the first interview and new codes from the second interview are repeated for the subsequent analyses. In this process, open coding can be shared among different categories. Several categories were developed by clustering related codes according to particular categories. These categories are then broken into sub-themes by selecting or grouping several related categories into a designated topic. At this stage, each category clearly had its own different sub- theme thus, the theme formation was both exclusive and mutually exclusive i.e., the sub-theme was not related to other sub-themes (Yin, 2011). The final stage included a process of validation and reliability. Creswell and Poth (2017) describe validation as an endeavor to evaluate the accuracy of the results as best described by the researcher, participants and reviewers.

The Focus Groups involved an open discussion among stakeholders who comprised policy makers and experts from industries and local universities. Their opinions or perceptions on the formulation of AI- GEF were sought. Researchers also shared the proposed findings from previous interviews, roundtable and benchmark exercises carried out earlier to stimulate discussions. There were 15 experts from the AI field alone and an additional desiderata of six well-informed individuals. According to Tan (2018), such a mix allows a good exchange of views, ideas and recommendations. The advantage of this method is that it produces a large amount of information from many people in a short time. One of the key benefits of FGDs is that they provide an opportunity for participants to share their perspectives and experiences in a supportive and non- threatening environment, allowing for a deeper and more nuanced understanding of AI governance and ethical matters. It also provided an opportunity for project team members to observe and analyze the dynamics of group interaction, which generated valuable insights into this study.

Findings and Recommendations

Based on four research objectives, semi-structured questions raised and data gathered during all the sessions and activities, the following section highlights the results and recommendations of this study.

Background of Respondents

As mentioned earlier, the total number of respondents were 75, with 20 being females (26%) and 55 (74%) males. A total of 53 respondents (70%) were aged between 25-45 while the rest were aged between 46-65 (30%). There were two categories for duration of service: more than 10 years and less than 10 years. The majority of respondents (85 %) had more than 10 years of working experience while the remaining 15% comprised the rest. Most (65%) respondents had a doctorate degree while the remaining 35% had Masters and Bachelor's degrees. The respondents hailed from various disciplines namely Information Technology, Computer Science, AI, Science, Technology and Innovation Policy, Trade & Investment, R&D Management, Human Resources, Management, Public Administration, International Business Management and Law. A few of them had served more than twenty years in fields such as IT, Technology Management and Business Management.

Perceptions towards of Responsible AI /AI Principles and implementation

This section is based on an interview session with 32 respondents. The majority of respondents (95%) had a positive perception/optimistic view of AI technology. Nonetheless, they also believed that AI may pose societal risks if its development is not guided properly. The biggest risk posed was perceived to be the loss of human lives and/or human civility. Over time, AI may pose serious risks to society, particularly if it becomes sentient and tries to "emancipate itself" from its preset parameters. Bostrom and Cirkovic (2011) place global risks into three categories: risks from nature, risks from unintended effects, and risks from hostile acts. AI was shown in two risk categories: risk from unintended effects like climate change and pandemics, and risks from emerging technologies such as particle accelerators as well as social collapse.

In the Malaysian context, 28 interview respondents (90%) had limited knowledge on the topic of Responsible AI, AI Principles, AI Ethics Standards as well as risks associated with AI technology. Twenty-two (22) of them (70%) claimed that they used AI daily, particularly in the area of communications and financial transactions, but most of them did know or were aware of the AI components embedded into such systems.

Many respondents also broadly echoed risks posed by large language model tools such as ChatGPT. While this study was conducted when ChatGPT version 3 (ChatGPT-3) was in its embryonic stage, the risks associated with more advanced evolutions (versions 4 and 5) were also discussed. The specter of job displacement, existential threats to humans, and misleading analyses leading to human error were deliberated.

Table 4: *Principles for Responsible Ai: Respondent Breakdown*

	Agreement Level					
	Low		Moderate		High	
Fairness (32)	2	6.25%	8	25 %	22	68.75%
Reliability, Safety, & Control	3	9.3%	13	40.7%	16	50%
Privacy & Security	1	3.1 %	11	34.4%	20	62.5%
Inclusiveness	3	9.3%	4	12.5 %	25	78.2%
Transparency	2	6.25 %	9	28.15 %	21	65.6 %
Accountability	2	6.25 %	6	18.75 %	24	**75 %**
Pursuit of Human Benefits & Happiness	2	6.25 %	3	9.3%	27	84.45 %

Among the three highest scoring Principles for Responsible AI were: Pursuit of Human Benefits and Happiness; Inclusiveness; and Accountability. The former was prioritized over the latter two. Reliability, Safety, and Control had the lowest agreement levels (lowest confidence levels) among respondents.

Barriers faced in adopting AI among stakeholders in Malaysia

Inputs were gathered from 55 Malaysia respondents over barriers faced in adopting AI in Malaysia. Four major barriers noted were in the areas of governance; finance; infrastructure (including communications platform); and knowledge.

In the area of Governance, there are currently no AI-GEF, AI Guidelines as well as an institution to govern, control and monitor AI implementation and performance. Lack of commitment by quadruple helix stakeholders (government, society, industry, academia) were also noted during the interview session. In the area of Finance, insufficient budget allocation for the implementation of Responsible AI and AI Principles were singled out. Respondents also noted the lack of investments for specialized talent in the area of AI Principles implementation. Similar gaps were noted in the area of physical infrastructure and technological capacities to implement AI Principles. A centralized communication platform for effective implementation has yet to be established. The majority of respondents were not even aware that a Virtual Reality platform towards this end was launched by MOSTI in 2021 as part of an integrated communication platform for all stakeholders on any technology-related subject. The lack of knowledge on AI Principles, particularly due to the lack of training and related AI courses on AI Principle as well as limited local champions who can steer the implementation of AI Principles were also discussed. Respondents also singled out the lack of communications on AI initiatives by MOSTI.

Benchmarking with the Japanese Governance on AI Principles Implementation

Currently in Japan there are no cross-sectoral laws and regulations applicable to AI. However, given that AI often involves the use of personal information, compliance with the AI Principles is deemed necessary. Japan is a global champion in harnessing innovative and emerging technologies. Its AI Strategy outlines four (4) goals and six (6) programs, one of which is developing and implementing AI Social Principles. In the area of practical implementation of AI principles, particularly as it relates to the practical implementation of the privacy principle, the Japanese understand that they need to pay attention to not only AI model building and their output, but also how they manage input data to the AI model.

It is critical to understand that the implementation of AI Principles cannot be undertaken without creating an AI Model for Governance. This model was unveiled during the AI-Rmap exercise in Malaysia. This model was based on the Strategic Foresight Model (SFM) that was formulated by one of the authors during his doctoral research (Maavak, 2019).

A framework for AI principles, the Social Principles of Human-Centric AI, was launched by the Japan Cabinet Secretariat in 2019. An overview of the document is shown on Figure 5.

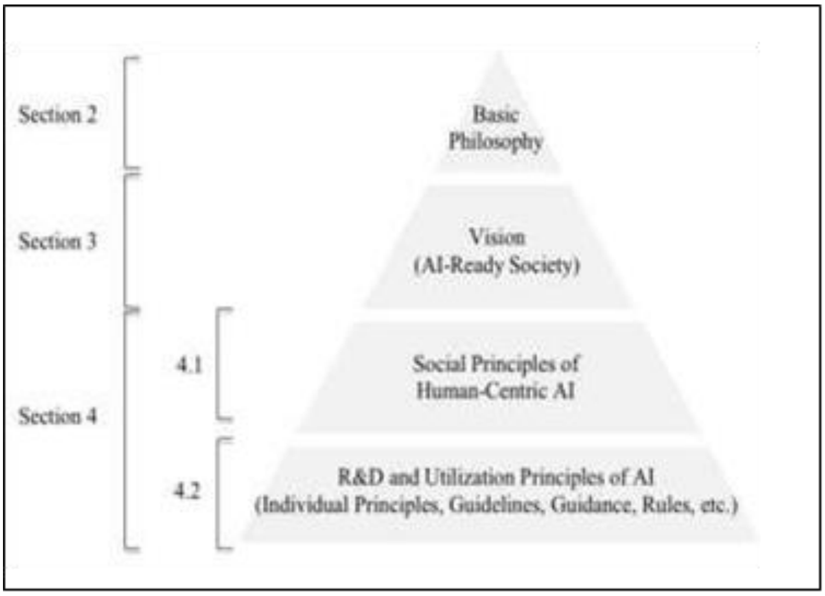

Section 2
Section 3
Section 4
4.1
4.2
Basic Philosophy
Vision (AI-Ready Society)
Social Principles of Human-Centric AI
R&D and Utilization Principles of AI (Individual Principles, Guidelines, Guidance, Rules, etc.)

Figure 5: *AI Social Principles of Human-Centric*

The Social Principles of Human-Centric AI describes how AI should be governed, so that various companies and economic sectors can be guided by its concepts. For instance, in February 2019, the Japan Business Federation (Keidanren) put out the “AI Utilization Strategy: For an AI-Ready Society" which was effectively a strategic blueprint for AI utilization. The phrase “AI-Ready Society’ is also used in the Social Principles of Human-Centric AI document. In spring 2019, companies like Fujitsu, NEC, and NTT Data launched their own AI Principles based on the document's recommendations. These multinationals and a company called ABEJA also set up ethics committees to discuss how AI should be governed and distinguishing the right from the wrong. The main thrusts behind these human- centric AI Principles framework are sustainability, diversity and inclusion.

The OECD principles for AI encourage shared prosperity, sustainable growth, and well-being. Japan desires to promote an AI-based socioeconomic environment that is equitable, where different parties can leverage AI and data. This form of Responsible AI will be used to engage in socio-economic activities that are compensated in accordance with accomplishments. It should give participants a sense of achievement, have more time to relax, etc. Thus, Japan's approach is congruent with OECD AI Principles. In response to the Social Principles of Human-Centric AI, the Ministry of Economy, Trade and Industry (METI) of Japan launched its AI Governance Guidelines in 2021. METI characterizes AI Governance as the planning, development, and implementation of technological, organizational, and societal structures by stakeholders with the goal of minimizing risks associated with the adoption of AI while maximizing the benefits. The Governance Guidelines for Implementation of AI Principles highlighted future issues such as coordination between standards and policies and guidance on the adoption of AI by the government. Overviews of both AI Governance and AI Governance Guidelines are depicted on Figures 6 and 7 respectively.

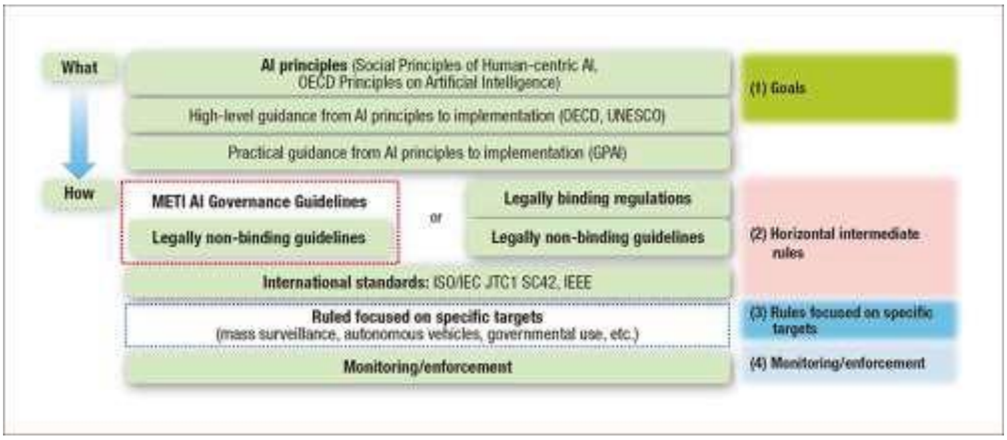

What
How
AI principles (Social Principles of Human-centric AI, OECD Principles on Artificial Intelligence)
High-level guidance from AI principles to implementation (OECD, UNESCO)
Practical guidance from AI principles to implementation (GPAI)
METI AI Governance Guidelines
Legally non-binding guidelines
or
Legally binding regulations
Legally non-binding guidelines
International standards: ISO/IEC JTC1 SC42, IEEE
Ruled focused on specific targets
(mass surveillance, autonomous vehicles, governmental use, etc.)
Monitoring/enforcement
(1) Goals
(2) Horizontal intermediate rules
(3) Rules focused on specific targets
(4) Monitoring/enforcement

Figure 6: *AI Governance*

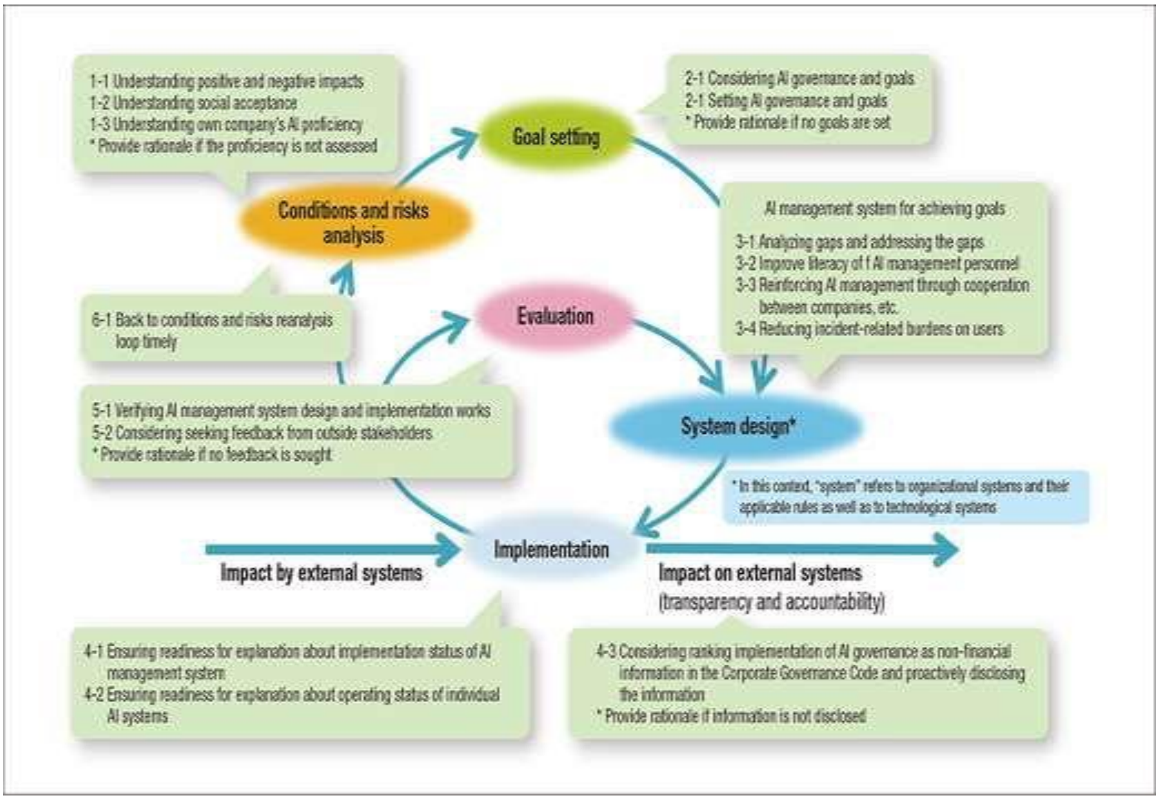

1-1 Understanding positive and negative impacts
1-2 Understanding social acceptance
1-3 Understanding own company's AI proficiency
* Provide rationale if the proficiency is not assessed
Goal setting
2-1 Considering AI governance and goals
2-1 Setting AI governance and goals
* Provide rationale if no goals are set
Conditions and risks analysis
AI management system for achieving goals
3-1 Analyzing gaps and addressing the gaps
3-2 Improve literacy of f AI management personnel
3-3 Reinforcing AI management through cooperation between companies, etc.
3-4 Reducing incident-related burdens on users
6-1 Back to conditions and risks reanalysis loop timely
Evaluation
5-1 Verifying AI management system design and implementation works
5-2 Considering seeking feedback from outside stakeholders
* Provide rationale if no feedback is sought
System design*
* In this context, "system" refers to organizational systems and their applicable rules as well as to technological systems
Implementation
Impact by external systems
Impact on external systems
(transparency and accountability)
4-1 Ensuring readiness for explanation about implementation status of AI management system
4-2 Ensuring readiness for explanation about operating status of individual AI systems
4-3 Considering ranking implementation of AI governance as non-financial information in the Corporate Governance Code and proactively disclosing the information
* Provide rationale if information is not disclosed

Figure 7: *AI Governance Guidelines*

A majority of respondents from agencies and companies were aware of the Japanese guidelines and intend to adhere to it accordingly. Respondents from universities were not aware of the guidelines published by the Japanese government. Many governments therefore have a problem in communicating its AI principles, ethical outlook and guidelines to quadruple helix realms. This is an issue which the Malaysian government must overcome.

Formulation of AI Governance & Ethic Framework (AI-GEF) and Implementation

An AI governance and ethics framework outlines the step-by-step process on how to design, build, process, use, consume, manage data. Effectively, it will be an AI Governance Model.

The Governance and Ethics Framework developed by Japan's Governance Guidelines for Implementation of AI Principles in 2021 serves as an excellent template towards this end.

As outlined by Figure 7, the Goal Setting phase begins with a series of strategic questioning and SWOT analysis. Anticipated internal and external changes will have to be factored as this phase doubles up as a scene- setting stage. Policy-makers and stakeholders deliberating in this stage will have to be realistic over local conditions and capabilities and not be carried away by trends set by “advanced economies”. For instance, many experts are predicting an imminent collapse of the electric vehicle industry as there is not enough lithium and cobalt (amongst others) on earth to support a global EV ecosystem. Recycling material from the renewable energy sector is also prohibitively expensive and such material is frequently dumped in landfills – potentially creating an environmental nightmare for future generations. To avoid technological traps, a full-time foresight component must be instituted at this stage. Foresight will enable all stakeholders to continually re-evaluate AI governance modus operandi, goals before they are future-proofed and “cast in stone”. New AI-powered Technologies will also be constantly evaluated to probe latent risks. Risks detected will be regularly looped back to apex policymakers within the AI governance structure.

Foresight is a critical prelude to an AI management system which will involve quadruple helix stakeholders who will participate via a virtual platform on a voluntary basis. Allocations will be needed to set up the virtual platform which will operate on a 24/7/365 basis. The AI management system will help improve AI literacy among stakeholders while reinforcing cooperation between them. This will enable cross-sectoral or systems learning among participants and thereby reduce risks and AI-related burdens (incidents, costs, threats etc.) on users. The first task of the foresight component in an AI-GEF is to harmonize the contrasting pulls of all elements identified in a national AI ecosystem. For instance, if top AI researchers warn of an impending and existential threat posed by AI systems (RT.com, 2023), then it is the task of the foresight component or team to investigate the matter. If such warnings are deemed credible, then it should recommend a cessation of affected AI activities in a vulnerable sector.

Foresight will have to be coupled with systems design. It is imperative that experts on systems theory and foresight be employed on a full-time basis to helm the AI management system. This is to ensure continuity and optimality in the AI Governance structure. As an incentive for AI adoption, the implementation of AI governance should be designated as non-financial information in the corporate governance code.

The AI Governance structure should also continuously update information on thresholds crossed in its AI management as well as progress made by individual entities in adopting AI. At this stage, the number of stakeholders should be increased to allow a higher volume of feedback. These feedbacks will be used to improve or re-align the AI governance structure, mode of operations and goals. This process will be repeated until a mature AI Governance and Ethics Framework (AI-GEF) is developed for implementation. At this stage, performance measures and evaluation metrics would be mature enough for execution. The legal pillars of AI governance as well as Ethical Standards will also be fully- established at this stage. While discussions on legal, ethical standards and measurement indices will commence from the onset of the AI Governance process, it should only be finalized until the governance structure attains robustness and maturity.

The AI-GEF should include an apex decision-making entity that will govern all aspects of AI in a nation. This will avoid policymaking from being swayed by the billion-dollar war chests of Big Tech. Figure 8 illustrates the formulation process of an AI-GEF which will facilitate the implementation of Responsible AI.

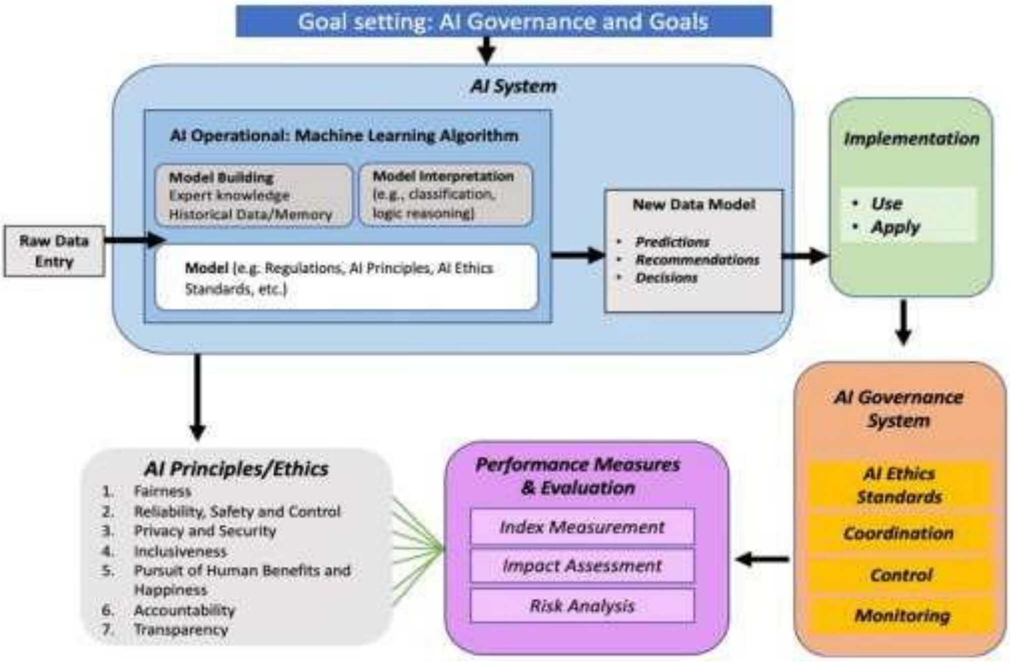


Figure 8: *Artificial Intelligence Governance & Ethics Framework (AI-GEF) Source: Author (2023)*

The AI-GEF model (Figure 8) can be used as a reference for all stakeholders who will include:

- Policymakers/heads of institutions and universities: A person or group that is responsible for ensuring appropriate disclosure of information to various stakeholders including ensuring collaboration, giving overall direction to operation levels in order to establish a management system so that sound ethical standards are adhered to.
- Senior management or operations staff from agencies and companies: A person or group that designs and operates the management system according to directions from top management. This ensures ethical standards and codes of conduct are adhered to while facilitating continuous evaluation of the management system.
- AI system developers: Companies that develop AI systems, AI system operators, companies that operate AI systems and data providers. This includes entities that develop AI systems for their own use or who provide them as a business, including entities conducting AI training and re-training.
- AI systems operator: An entity that operates an AI system for its own use or for the use of others as a business. This includes entities which do not engage in AI systems development, but simply procures and operates an AI system, and is responsible to a certain extent for the operation of the AI systems or maintenance of its performance.
- Data providers: An entity that, as a business, provides others with data collected from a number of unspecified sources, data collected from specific people, data prepared by the data provider itself; a combination of them; or data created by processing the above-mentioned data, for the purpose of AI system training, etc.
- AI customer/consumer: A person that uses AI developed by AI System developers.

This AI-GEF will be a “living document” intended to provide stakeholders with a reasoned approach to judgment and to assist with the documentation of considerations associated with the AI lifecycle. In doing so, this AI-GEF will enable enhanced understanding of goals between AI practitioners and managers while promoting the ethical use of AI.

Recommendations

Build capacity to engage in AI governance

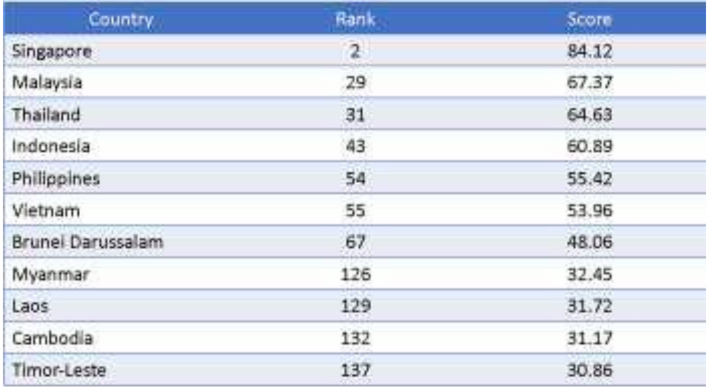
Different types of capacities are needed to facilitate AI Ethics and Governance. The first is an understanding of technologies and their potential implications. Besides training workshops, a respondent suggested that participation in or engagement with AI interest groups would enable interested parties to learn more from technology professionals and advocates. Some of these interest groups actively engage with governments and can be valuable allies in the area of AI governance. Capacity-building will include AI coalitions and the coordination of AI advocacy across local, national, regional, and international levels.

Enhance Collaboration and Network

Partnership between the public and private sectors should be of a "glocal" nature. To adequately adapt to technological developments that are reshaping business strategies and human-machine interactions, a progressive approach is crucial for AI governance. Collaboration and networking with civil society groups will yield benefits in terms of feedback volume and will optimize the use of limited resources. By bridging different societal actors, this approach will facilitate interdisciplinary and cross-sectoral conversations on AI governance. Few respondents recommended that knowledge-sharing sessions be intensified. Establishing a community of scholars and advocates under the rubric of AI governance, possibly through a conference or networking program, will help generate AI literature and projects between academic researchers and the tech community. This will increase cross-disciplinary knowledge sharing and dissemination. Bridging society and academia through guest lectures which include a human rights or societal angle can inspire a future generation of ethical AI engineers.

According to the Government AI Readiness Index 2022, Malaysia was ranked 29 globally or number 2 in ASEAN with a score of 67.37 as illustrated in Table 5 below. Referring to the indices and statistics cited, we can categorize ASEAN countries into three distinct levels. Singapore is in its own league, being AI-ready and competitive at the global level according to the scores listed in the table. The second level comprises Malaysia, Indonesia, Thailand, Philippines, Vietnam, and Brunei Darussalam, which have governments that are playing catch up in building supportive policies and regulations. These nations are increasingly becoming digitally connected and savvy. The third level includes Myanmar, Cambodia, Laos, and Timor-Leste, which have low Internet penetration, and whose governments lack the capacity to plan for and support the adoption of AI technologies. These contextual differences need to be taken into account when analyzing the benefits and risks of developing an ASEAN AI-EGF.

Table 5: *Government Ai Readiness of Asean In 2022*



Country	Rank	Score
Singapore	2	84.12
Malaysia	29	67.37
Thailand	31	64.63
Indonesia	43	60.89
Philippines	54	55.42
Vietnam	55	53.96
Brunei Darussalam	67	48.06
Myanmar	126	32.45
Laos	129	31.72
Cambodia	132	31.17
Timor-Leste	137	30.86

Source: *Oxford Insights (2022)*

Capability building training program on AI application

Currently, there is a lack of awareness programs on AI as well as insufficient studies on AI applications and their social impacts. This makes it more difficult to construct evidence-based AI governance and advocacy for Malaysia. Currently most studies on the impacts of AI technologies are based on the experiences of developed nations such as the United States, EU nations, Singapore and Japan. Malaysian researchers tend to replicate studies on AI Ethics rather than modifying them to suit national imperatives. Examples of research projects proposed by respondents include a specific AI product from development, roll-out, to real-world usage, and a repository of portfolios of major companies supplying AI technologies and their implications.

Leverage existing capacities on human rights and community work

A few Malaysian ministries and agencies have conducted human rights studies into AI gender equity and privacy protection issues. These studies enable them to contribute to discussions over the ethical and societal dimensions of AI governance. Professional and interest-based associations, charity and mutual aid organizations as well as local and neighborhood groups can also provide a social backbone for AI governance. The AI-GEF should have an in-built horizon scanning facility to tap into these discussions and thereby engage these interest groups in order to improve the governance process.

All seven AI Principles highlighted in this study are inevitably linked to each other. A systems theory approach is therefore needed to situate these principles accordingly for optimal implementation. The goal is a holistic AI governance framework which can monitor how elements of a national AI ecosystem may interact with each other. This will avoid risks like unanticipated “emergence” in a system while identifying hidden opportunities (Maavak, 2019). To coordinate the implementation of AI Principles, it is important for the government to have a centralized command center to plan, execute, monitor, and evaluate the whole implementation process. This command center must be given access and authority over the coordination and management of the entire national AI ecosystem – including the role of participants and stakeholders.

It is impossible to kickstart a national AI-GEF without establishing a command center. It will act as the “brain” for all AI activities within the nation. Furthermore, the proposed command center will be empowered to make decisions whenever competing claims emerge. Without it, the AI-GEF will be stuck in a bureaucratic tangle.

With strategic coordination from the command center, the next step will involve constructing a framework and guideline process for the implementation of AI Principles and Governance. In the process, the command center will spearhead AI development across the public, private and social sectors. The government should also convey these principles among its internal agencies, associated sectors and institutions, and integrate them into government policies (and not only for AI). These principles and their implementation should be flexibly revised in alignment with the evolution of AI technologies, social changes, global changes, and many other factors (e.g., sustainable policy planning).

Conclusion

This study explored the potential receptivity of AI Governance and Principles among key respondents in Malaysia and Japan. The seven AI Principles were seen to address concerns regarding the adoption of AI-led technologies. The literature review process analyzed how AI governance and principles were treated by the government of Japan, UNESCO and OECD.

Later, a methodology was developed to construct the building blocks of an Artificial Intelligence Governance and Ethics Framework (AI-GEF). Several Interview sessions, discussions, and Malaysia's National AI Roadmap (AI-Rmap) helped in identifying key components needed to establish a robust AI-GEF.

The next step is for the government to ramp up the implementation of AI Principles. Governance is the primary enabler of AI's development. It is suggested all ministries and agencies agree to a standard set of guidelines to streamline AI Governance. Government coordination is needed to remove barriers that may obstruct the implementation of AI Principles in the country. MOSTI, along with other ministries and relevant industry players, must intensify efforts to tackle AI risks.

A command center should be situated within the AI-GEF to helm the national AI ecosystem. It will act as the "brain" and coordinator for all AI activities and initiatives in the nation. This will help avoid knee-jerk reactions to new or unexpected developments in AI. Italy, for example, became the first Western country to ban ChatGPT for security reasons. Other governments plan to either ban or regulate this AI-powered chatbot tool. Rapid reaction capabilities – with policy decision powers – should be an inbuilt feature of an AI-GEF.

Once the structure and functions of the AI-GEF is finalized, the implementation of Responsible AI will promote national AI adoption. If the proposed seven principles are not addressed and implemented objectively and successfully, it may affect AI trustworthiness in the long run.

This research project may bring significant contributions to relevant government agencies, research institutions and universities on the subject of AI Governance. Malaysia can serve as a testbed for an expanded AI-GEF for the ASEAN region. This study also addressed the needs and demands of a new "Look East Policy" as Japan is ready to lead the way in the area of AI Ethics and Governance.

References

ASEAN Vision on AI Ethic and Governance Framework (AI-EGF), 2019, Singapore.

Bostrom, N., & Cirkovic, M. M. (2011). Global Catastrophic Risks. Oxford University Press: Oxford.

B., Mirbabaie, M., Lembcke, T.-B., & Hofeditz, L. (2021). Ethical Management of Artificial Intelligence. Sustainability, 13(4), 1974. https://doi.org/10.3390/su13041974

BigIT. 2020. Malaysia AI Blueprint Annual Report 2020.

Capgemini Report (2020), Ethical Issues from the use of AI. https://www.capgemini.com/news/press-releases/ai-and-the-ethical-conundrum-report/

Creswell, J. W. & Poth, C. N. 2017. Qualitative inquiry and research design: Choosing among five approaches, Sage publications.

Creswell, J. W., & Creswell, J. D. (2018). Research design: qualitative, quantitative, and mixed methods approaches. Sage publications.

Competing in the Age of AI 2020, Harvard Business Review Press.

Dasar Sains, Teknologi Dan Inovasi Negara 2021-2030 (2021) Kementerian Sains, Teknologi dan Inovasi.

Deloitte. 2018. National Artificial Intelligence Framework for Malaysia Final Report.

Deloitte Centre for Technology, M. & T. (2020). Deloitte's State of AI in the Enterprise, 3rd Edition.

Denzin, N. K., & Lincoln, Y. S. (2018). The Sage handbook of qualitative research. Sage publications.

EC (2019). A definition of AI: Main capabilities and scientific disciplines. Published by the European Commission High Level Expert Group on Artificial Intelligence, Dec 18, 2018. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.p df

Future of Life Institute. (2015). An Open Letter Research Priorities for Robust and Beneficial Artificial Intelligence.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. Minds and Machines, 30(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Governance of Artificial Intelligence (AI) in Southeast Asia 2021, Swedish International Development Cooperation Agency (SIDA).

Government of Malaysia. (2021). Twelfth Malaysia Plan 2021-2025. Putrajaya: Economic Planning Unit, Prime Minister's Department.

González-Esteban y Patrici Calvo, E. (2022). Ethically governing artificial intelligence in the field of scientific research and innovation. Heliyon, 8(2), e08946. https://doi.org/10.1016/j.heliyon.2022.e08946

Governance Guideline for Implementation of AI Principle (2021), Japan

Government AI Readiness Index (2022). https://www.oxfordinsights.com/government-ai-readiness-index-2023

IEEE (2019). Ethically aligned Design. Retrieved from https://ethicsinaction.ieee.org/

IEEE Standards Association (2020). Statement Regarding the Ethical Implementation of Artificial Intelligence Systems (AIS) for Addressing the COVID-19 Pandemic. IEEE Standards Association.

Helbing, D., Caron, & Helbing. (2019). Towards digital enlightenment. New York, NY: Springer International Publishing.

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society, 22(7), 900-915.

Industry 4wrd National Policy on Industry 4.0.(2018), Ministry of International Trade, Malaysia

Insights, Oxford. 2022. Government AI Readiness Index. https://www.oxfordinsights.com/government-ai-readiness-index-2022

Internet Society. (2017). Artificial Intelligence and Machine Learning: Policy Paper. Internet Society. https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/

Jobin obin, A.; Ienca, M.; Vayena, E. (2019) The global landscape of AI ethics guidelines. Nat. Mach. Intell. 1, 389–399. [CrossRef]

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F.& Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094-1096.

Maavak, M. 2019. Developing a Net-Centric Foresight Model for the Management of
Emerging Risks. Unpublished Doctoral dissertation, Universiti Teknologi Malaysia.

Maavak, M. 2023. Will humans become an extension of their machines? ChatGPT may have irrevocable consequences for learning and decision-making. RT.com, March 13, 2023. https://www.rt.com/news/572583-chatgpt-ai-decision-making/

Malaysia Digital Economy Blueprint 2021 , Economic Planning Unit, Prime Minister Department.

Malaysia Artificial Intelligence Roadmap (2021-2025), 2022, Ministry of Science Technology & Innovation, Malaysia, Mikhaylov, S. J., Esteve, M., Campion, & A. Averill. (2018). Artificial intelligence for the public sector: Opportunities and challenges of cross-sector collaboration. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.

Mintrom, M., & O'Connor, R. (2020). The importance of policy narrative: effective government responses to Covid-19. Policy Design and Practice, 3(3), 205–227. https://doi.org/10.1080/25741292.2020.1813358

Nalini, B. (2019). The Hitchhiker's Guide to AI Ethics. Medium Retrieved from https://towardsdatascience.com/ethics-of-ai-a-comprehensive-primer-1bfd039124b0

N. A. Zakaria, Z. I. (2016). Technology Ethics Attributes in Handling Confidential Information for the Armed Forces. Journal of Advanced Research in Social and Behavioural Sciences, 2(1),

OECD (2019). OECD AI Principles overview https://oecd.ai/en/ai-principles

Paper, White. 2019. "White Paper A Framework for Developing a National Artificial Intelligence Strategy Centre for Fourth Industrial Revolution.

Report, Development. 2018. "China AI Development Report 2018." (July).

RT.com. 2023. 'Everyone on Earth will die,' top AI researcher warns: Humanity is unprepared to survive an encounter with a much smarter artificial intelligence, Eliezer Yudkowsky says, April 1, 2023. https://www.rt.com/news/573972-ai-danger-nuclear-yudkowsky/

Science & Technology Foresight Malaysia 2050: Emerging Science, Engineering & Technology (ESET) 2017. Study. Malaysia, Academy of Sciences.

Singapore Smart Nation: National Artificial Intelligence Strategy. 1. 2019 Smart Nation and Digital Government Office. Singapore.

Tan. W (2018). Research Methods: A Practical Guide for Students and Researchers, Singapore, World Scientific Publishing Company.

The Recommendation on the Ethics of Artificial Intelligence 2021 https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

The National Artificial Intelligence Research and Development Strategic Plan :2019 Update. 2019 Science, National, Technology Council, Information Technology, and Development Subcommittee, Malaysia.

WEF. (2020). The Global Risks Report 2020: Insight Report.

Yeh, S.-C., Wu, A.-W., Yu, H.-C., Wu, H. C., Kuo, Y.-P., & Chen, P.-X. (2021). Public
Perception of Artificial Intelligence and Its Connections to the Sustainable Development Goals. Sustainability, 13(16), 9165. https://doi.org/10.3390/su13169165

Yin. R. (2011). Qualitative research from start to finish. New York, NY. Guilford Publications, Inc.

10-10 MySTIE Framework 2020, MOSTI.