ASEAN Engineering Journal

Full Paper

ADAPTING EXISTING PRACTICES FOR ISMS-CERTIFIED ORGANIZATIONS IN SUPPORT OF RESPONSIBLE AI

Article history
Received
19 November 2024
Received in revised form
23 April 2025
Accepted
19 June 2025
Published online
28 February 2026

David Lau Keat Jin[a*], Ganthan Narayana Samy[a], Fiza Abdul Rahim[a], Mahiswaran Selvananthan[b], Nurazean Maarop[a], Mugilraj Radha Krishnan[a], Sundresan Perumal[c]

[a]Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia
[b]Faculty of Social Sciences and Humanity, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia
[c]Faculty of Science and Technology, Universiti Sains Islam Malaysia, 71800 Nilai, Negeri Sembilan, Malaysia

*Corresponding author
davidlkeat@graduate.utm.my

Graphical abstract

Information Security Management System
1
Fairness
2
Reliability, Safety And Control
3
Privacy And Security
4
Inclusiveness
5
Transparency
6
Accountability
7
Pursuit of Human Benefit And Happiness
Responsible AI

Abstract

Prior to adoption of Artificial Intelligence (AI), organizations may be required to comply with certain industry standards to ensure customer confidence and interoperability of their products, which demands resource allocation and designated responsibilities. For Malaysian public offices certified under the Information Security Management System (ISMS), compliance with a new standard in support of Responsible AI would entail further resources and new reporting structures. Hence, this study proposed the adaptation of current practices for these organizations at the early stages of AI adoption. Ten sources, chosen for authenticity, credibility, representativeness, and meaning, provide the basis for the relevant proposals, including context establishment, risk identification, risk prioritization, and focus area for each control in Annex A of ISO/IEC 27001:2022. The results outlined key actions to support Responsible AI, with future research focusing on validating this framework in ISMS-certified settings.

Keywords: Artificial Intelligence; ISMS; Risk Management; Responsible AI; Framework

1.0 INTRODUCTION

Artificial Intelligence (AI) system is formed by software (and possibly also hardware) designed by humans that, given a complex goal, act in the physical or digital realm by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal [1]. AI systems can either use symbolic rules or learn a numeric model, and they can also 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).

Generative AI (GenAI), as the name implies, can generate novel and meaningful content, including text, images, or audio, derived from training datasets [2]. It is an enabling technology that could enhance Human Machine Interaction (HMI) and the Internet of Things (IoT) [3]. The extensive proliferation of this technology, exemplified by innovations such as Dall-E 2, GPT-4, and Copilot, is presently transforming the modalities through which individuals engage in work and interpersonal communication. GenAI systems are not solely applicable for artistic endeavors, such as producing text that emulates the style of authors or generating images reminiscent of illustrators; they also possess the potential to augment human capabilities as adept question-answering systems. Notable applications encompass information technology (IT) support desks, wherein generative AI facilitates routine tasks involving information processing and addresses routine inquiries related to culinary recipes and medical guidance. Industry reports estimated that GenAI could potentially enhance global gross domestic product (GDP) by 7% and displace approximately 300 million positions held by knowledge workers [4]. In addition, a comprehensive global survey concerning responsible AI underscores organizations' primary concerns associated with AI encompass privacy, data security, and reliability. Furthermore, recent investigations from the AI Index illuminate a pronounced lack of common standards for reporting on responsible AI. This phenomenon is manifested as preeminent developers, such as OpenAI, Google, and Anthropic, predominantly evaluate their models against disparate responsible AI benchmarks. Such a practice complicates endeavours to systematically assess the risks and limitations inherent in leading AI models.

As AI is fundamentally reliant on data, the incorporation of inaccurate, biased, or deliberately malicious data into the data-ingestion pipeline may yield erroneous, biased, and misleading outputs, thereby resulting in detrimental or catastrophic ramifications contingent upon its specific application [5]. For instance, the Twitter chatbot developed by Microsoft was compelled to cease operations following its exposure to racist data inputs from other Twitter users, which consequently generated racially offensive and insensitive remarks [6]. This predicament is further aggravated by the vulnerabilities and methodologies employed to execute adversarial attacks on AI systems, which have been documented by the Open Worldwide Application Security Project (OWASP) [7] and MITRE [8]. The 'OWASP Top 10 for Large Language Model (LLM) applications' serves as a comprehensive report delineating malicious attack on LLMs, whereas the 'Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS) Matrix' by MITRE constitutes a globally accessible repository of adversarial tactics and techniques targeting AI models. In parallel, studies on secure control mechanisms for autonomous agents also highlight the critical nature of system robustness against coordinated attacks, particularly on sensors and actuators [9].

From an ethical perspective, there were a few notable studies that aggregated the prevailing ethical concepts related to the use of AI. In practice, although there were more than 160 guidelines proposed globally, it remains uncertain whether they were sufficient to meet the governance challenges of AI [10]. In this regard, [11] studied the explicit interactions and contributory forces of different requirements in influencing the main ethical outcomes. Additionally, [12] summarized AI ethics into virtues of justice, honesty, responsibility and care while [13] identified transparency, fairness, non-maleficence, responsibility and privacy as the overarching principles of AI ethics. On the international front, there were consensuses on ethical principles envisioned by European High Level Expert Group on Artificial Intelligence (HLEG) [1], Montreal Declaration for a responsible development of AI [14] and Organization for Economic Cooperation and Development (OECD) for the use of AI [15]. A comparative summary of the principles highlighted is presented in Table 1.

Table 1 Principles Highlighted In International Forum

Ref.	Principles and Concerns	
[1]	1.	Human agency and oversight
	2.	Technical robustness and safety
	3.	Privacy and data governance
	4.	Transparency
	5.	Diversity, non-discrimination and fairness
	6.	Environmental and societal well-being
	7.	Accountability
[14]	1.	Well-being
	2.	Respect for Autonomy
	3.	Protection of privacy and intimacy
	4.	Solidarity
	5.	Democratic Participation
	6.	Equitable
	7.	Diversity inclusion
	8.	Caution
	9.	Responsibility
	10.	Sustainable development
[15]	1.	Inclusive growth
	2.	Sustainable development and well-being
	3.	Human-centred values and fairness
	4.	Transparency and explainability
	5.	Robustness, security and safety
	6.	Accountability

1.1 Related Work

Studies were conducted to develop or customize frameworks based on existing risk management standard. Standard is usually an official document containing agreed upon ways of doing something that serves the interests of multiple stakeholders [16]. It is usually formed and published by Standards-developing Organizations (SDOs) which are in turn consist of multiple organizations with expertise or interest in specific fields. Its prominence is propagated by technological advancement and innovation in various industries and sectors. Significant use of standards includes facilitation of interoperability of products and services, increased customers' confidence and reputation of organizations, especially businesses. For example, medical devices that conform to a certain international standard will ensure that patient's health information is collected safely, and interpretable by another device in different healthcare facility. Thus, in the absence of relevant legislations, standards will encourage businesses to offer products and services that comply to certain minimal requirements. Usually, to claim conformance to certain standards, an organizations or their products are required to pass an assessment or audit. For instance, organizations that obtained ISO 27001 – Information Security Management Systems (ISMS) standard are required to be audited by a certification body [17]. Currently, there are more than 30 AI-related standards as curated by The AI Standards Hub [18]. The Hub's mission is to advance trustworthy and responsible AI with a focus on the role that standards can play as governance tools and innovation mechanisms.

Several risk-based frameworks based on existing AI-related standards were proposed. The 'Data-driven Risk Assessment Methodology for Ethical AI' (DRESS-eAI) was developed as a standardized approach to ethical AI risk assessment with the aim of fulfilling the cross-functional requirements of an organization over multiple contexts [19]. This approach was compared to IEEE 7000-2021 standard which addresses a set of processes by which organizations can include consideration of ethical values throughout the stages of concept exploration and development. While this approach integrated questionnaires for the risk scanning phase and involve solicitation from stakeholders, it was acknowledged that there were challenges in obtaining meaningful stakeholder involvement as well as data quality issue. Furthermore, this approach did not provide compliance checklist nor guidelines for practitioners to follow. Based on ISO 31000 [20] and the then pre-approved AI Act, an AI Risk Management System (AIRMan) was proposed [21]. The structure of AIRMan was developed based on the structure of Annex SL and influenced by COBIT5/2019. This approach incorporated the activities of AI lifecycle with the requisite processes: context and requirements, business impact and critical analysis, risk assessment, risk treatment, risk acceptance and risk monitoring and communication. While the processes were integrated into the phases of plan, do, check and act, it was proposed prior to the approval of EU AI Act. While it can still serve as a useful guidance for organization, it does not highlight the role of stakeholders as well as applicable controls for common AI risks. Sectorially, the Trustworthy AI for Project Risk Management (TAI-PRM) framework was proposed for the manufacturing industry based on Failure Mode Effect Analysis (FMEA) and ISO 31000 [22]. It was conceived by mapping the ethical requirements to hazards and the sample computations for risks.

Spurred by the requirements specified in the approved EU AI Act [23], a Quality Management System (QMS) was developed which can be used by the relevant stakeholders stipulated in the legislation [24]. It consisted of 2 modules, the Risk Management System which was structured in accordance with ISO 31000 Risk Management Standard, as well as the Data Management and Data Governance System. While its applicability was demonstrated for LLM, it has yet to be optimized for other AI models. Moreover, the considerations of the potential risks may be limited as it has yet to integrate with external sources and business processes. Interestingly, a compliance report was designed grounded in the EU AI Act, NIST's AI Risk Management Framework and ISO 42001 AI Management System [25]. The template of the report was validated by 8 AI practitioners and 5 AI compliance experts. This template provided the necessary information for AI impact assessments that targeted the pre-deployment and design stage of AI. While 32 simple statements were given to guide the stakeholders in using the template to produce a complete report, the author stressed that mechanism to update the report was crucial in the evolving threat landscape of AI. Similarly, an AI Risk Ontology (AIRO) was developed to represent information and generate required information pertaining to high-risk AI application based on EU AI Act and ISO 31000 [26]. However, its utility may be confined to EU countries where the act is applicable. Table 2 provides a comparative analysis of the frameworks described in this sub-section.

Table 2 Analysis of Frameworks Developed on Existing Standard for Responsible/Trustworthy AI

Ref.	Framework	Standard Referred	Limitation
[19]	DRESS-eAI	IEEE 7000-2021	Data quality issue, difficulty in engaging multi-disciplinary stakeholders
[21]	AIRMan	COBIT5/2019 ISO 31000:2018	Role of stakeholders was not highlighted
[22]	TAI-PRM	FMEA ISO 31000:2018	May not be suitable for non-manufacturing sector
[24]	QMS	ISO 31000:2018	Applicability for non-LLM models was not validated
[25]	Compliance Template	NIST AI RMF	Challenges in updating the report generated
[26]	AIRO	ISO 31000:2018	Applicability for jurisdictions outside EU was not validated

1.2 Current Practices in ISMS

The Information Security Management System (ISMS) is a technologically-neutral management framework aimed at the establishment, implementation, operation, monitoring, review, maintenance, and enhancement of information security for an organization [27]. This framework, which encompasses organizational structures, policies, planning activities, responsibilities, practices, procedures, processes, and resources, holds significant importance due to the dynamic nature of risks and the necessity for systematic management. The inception of ISMS can be traced back to the 1990s when the Department of Trade and Industry in the UK embarked upon an initiative to enhance awareness regarding Information Security (IS) and proposed security controls for the safeguarding of information [28]. This initiative commenced with the Code of Practice in 1993 and subsequently evolved into the British Standard (BS7799-1) in 1995. It further transitioned from ISO/IEC 27001:2005 to ISO/IEC 27001:2007 and ultimately to ISO/IEC 27001:2013, which was formally adopted by the Malaysian government in 2010 [29]. On October 25, 2022, it was revised after a span of nine years, resulting in the new iteration known as ISO/IEC 27001:2022 Information Security, Cybersecurity, and Privacy Protection [30]. This updated version for ISMS delivers a thorough and integrated methodology for managing information security risks and serves as a useful framework for adopting organizations in upholding the confidentiality, integrity, and availability of their information assets.

The implementation of ISMS involves 5 sequential phases and two simultaneous phases with their associated activities [31]. The first phase deals with activities like defining the scope and boundaries of ISMS, establishing policy, performing risk identification, assessment, treatment and preparation of statement of applicability. Interestingly, the implementation of controls in the standard are specified in the statement of applicability [30]. Next, the formulation, implementation of risk treatment plan and control form the main activities in the second phase. To support the activities of this phase, management of resources and reference to related procedures are necessary. Subsequently, the activities pertaining to the 'Monitor and Review' phase includes reviewing the effectiveness of controls, conducting internal audit and update security plans. This is folllowed by the 'Maintain and Improve' phase which focuses on taking corrective and preventive actions as well as communicating actions and improvements. Finally, the certification phase is concerned with pre-certification assessment, certification audit and addressing the issues raised. ISMS certification in Malaysia are issued by certification bodies like SIRIM QAS International [32] and Cybersecurity Malaysia [33]. Figure 1 shows the iterative phases of ISMS implementation [31].

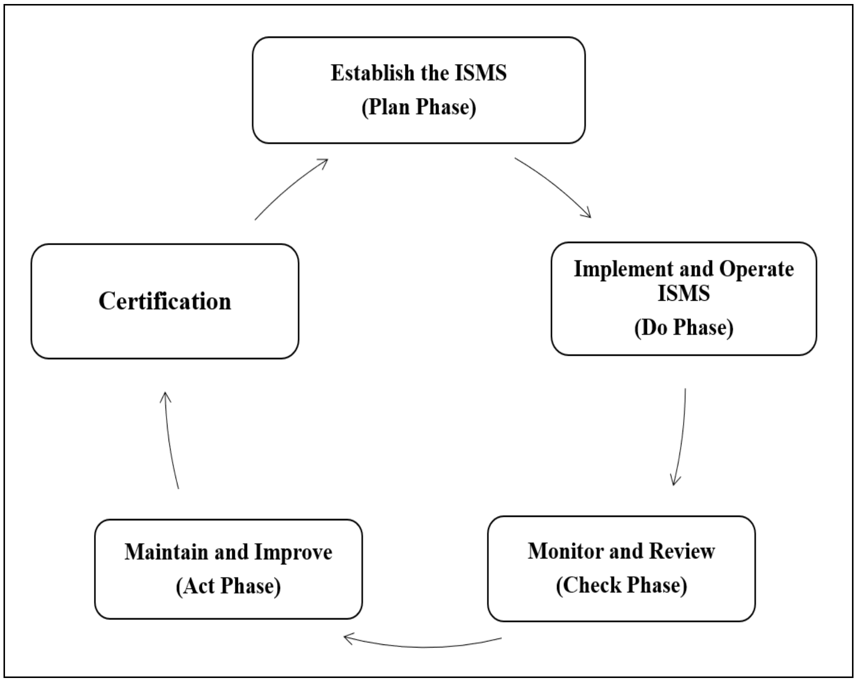

Figure 1 ISMS Implementation Phases [31]

1.3 Gap Analysis

Being one of the earliest and most comprehensive framework that promotes Responsible AI (RAI), the National Institute of Science and Technology (NIST) AI Risk Management Framework (RMF) is widely referred by academic articles in the discourse on RAI. While the framework articulates the characteristics of trustworthy AI and offers guidance for addressing them, its functional mechanism of map, measure and manage overlaps with the phases of risk management as practised in ISO 27001:2022. Figure 2 shows the parallels between the functions of NIST AI RMF and ISO 31000 risk management processes. The former provided more specific recommendation related to the domain such as elucidation of AI actors as well as the Test, Evaluation, Verification and Validation (TEVV) tasks that are associated with AI lifecycle. While guidelines may not be enforceable, the EU AI Act, which has been approved by European Parliament in March 2024, also adopts a risk-based approach in that it specifies which use-cases of AI which should belong to one of the unacceptable, high, minimal and acceptable levels of risk. Apart from that, there are currently 4 online databases that record AI-related risks worldwide for referenced by individuals and organizations [34-37]. Hence, a risk-based approach to AI adoption is well grounded in theory and practice.

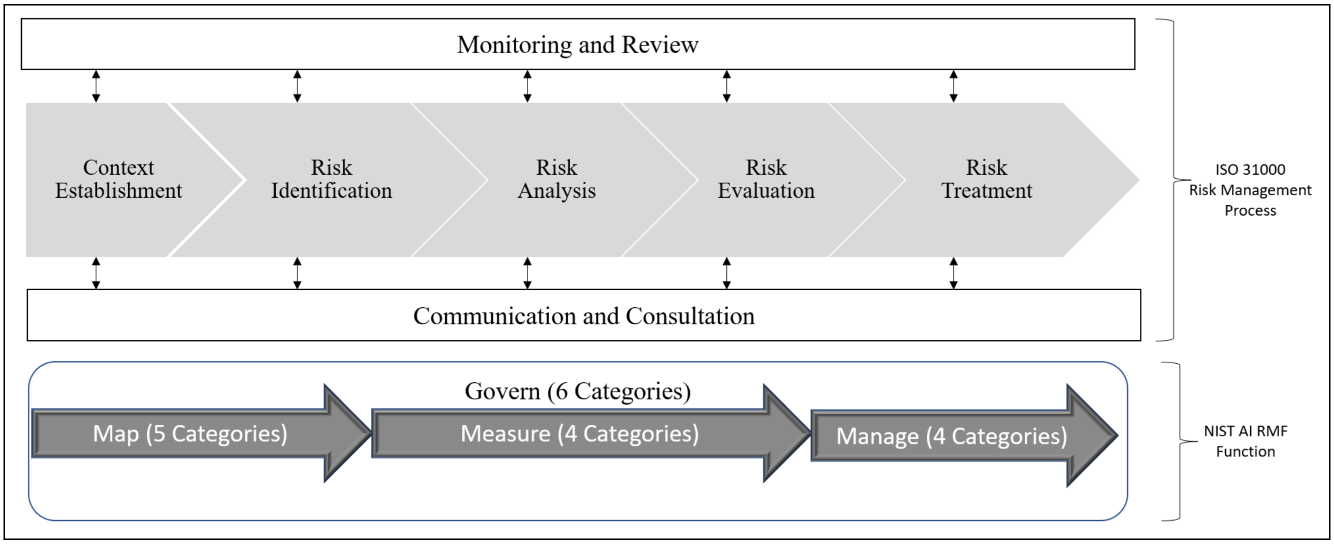

Figure 2 Similarities between NIST AI RMF and ISO 31000

While a taxonomy of AI risks may be a dynamic structure due to the evolving nature of the technology, a taxonomy of open issues in technical AI governance concluded that the existing challenges can be plotted along the intersections of two dimensions: capacities and targets [38]. The aspects of capacities include assessment, access, verification, security, operationalization, and ecosystem monitoring. On the other hand, the aspects of targets revolves around data, compute, models and algorithms and deployment. The open issues related to operationalization and ecosystem monitoring cut across all the aspect of capacities. Two of the issues which are the focus of this study include the translation of governance goals into policies and regulatory requirements as well as clarification of associated risks. These findings are echoed by [39] that highlighted principles to practice gap in the domain of Responsible AI (RAI). The study also affirmed the importance of matching the job roles of practitioner to specific recommendations in RAI. Thus, the extension of ISMS implementation augurs well for the realization of RAI because there are stakeholders identified for taking appropriate measures in the risk treatment and monitoring phases. This is echoed by a study that mapped the provisions in NIST NSF 2.0, COBIT 2019, ISO 27001:2022 and ISO 42001:2023 to risks of LLM which indicated significant gaps in risk management [40].

As highlighted in sub-section 1.2, all the government agencies in Malaysia are required to implement and obtain ISMS certification. Hence, the roles in implementation of ISMS have already been assigned to relevant personnel. Implementation of a new standard and subsequent auditing based on the new standard would require establishment of a new reporting structure [41]. Moreover, it is not a straightforward process to measure the return on investment in resources to implement the new standard [42]. In fact, most of the government agencies are in the early stage of maturity level in terms of AI adoption. Other than the use of facial recognition technology for access control by Ministry of Home Affairs, there are just a few pioneering agencies that started to use chatbot to serve visitors to their websites such as the National Digital Department [43] and Public Sector Home Financing Board [44] of Malaysia. Hence, the public offices can be considered to be in level 1 and 2 on the maturity scale [45]. The requirement to be certified and establishment of certification body based on a new standard may be premature. In fact, The National Guidelines on AI Governance and Ethics (NGAIGE) for Responsible and Inclusive AI which consists of 7 RAI principles was just released in September 2024. In line with these development, this study aims to address the following research question:

- How to adapt the existing ISMS framework in support of Responsible AI principles?

1.4 Required Addendum to ISMS Framework

The commencement of current ISMS implementation involves establishment of context [30]. Here, the context refers to many pertinent information such as organizational background, mission, vision and functions [20]. In addition, it considers the interactions of the organization with various internal and external parties. Consequently, the scope of the ISMS is determined. Normally, all this contextual information is documented in ISMS implementation manual. Typically, the organization would ensure that the scope of ISMS encompasses its core functions. As AI risks differ from traditional software risks, identification of contextual factors is emphasized by NIST AI RMF as part of the ‘Map’ function [46]. For example, the purposes, capabilities, risks, benefits and impact of AI for the adopting organization are part of this function. In fact, lessons learnt from online repositories that record AI issues and incidents showed that this information is useful for risk management and incident handling [47]. For the same reason, the EU AI Act mandates this and other related information from AI providers that offer their products or solutions for high-risk use cases in EU countries [23].

The focus of information security risk assessment is on preserving confidentiality, integrity, and availability [48]. On the other hand, the 7 RAI principles as highlighted by NGAIGE include accountability, fairness, inclusiveness, privacy and security, pursuit of human benefit and happiness, reliability, safety and control and lastly, transparency [49]. In upholding the 7 RAI principles, it is necessary to map the risks that could negatively affect the realization of these principles. Specifically, the purpose of risk identification is to identify what could happen, or which situations could exist, that may affect the achievement of the proposed objectives. Hence, a taxonomy of AI risks should be obtained for this purpose. However, the identification of risks is not sufficient for assignment of risk owners as the owners should have the capacity to monitor as well as implement controls associated with the risks under their purview. In this regard, the consideration of asset is necessary. Currently, there are 7 categories of assets defined which include: process, data, hardware, software, services, human resource, and premise. Notwithstandingly, there are other consideration in the developmental phase of AI such as compute platforms, algorithmic models and external tools [38, 50]. Since our goal is to adapt the existing ISMS practices, the inclusion of these consideration may be done as sub-categories of existing asset categories where appropriate.

Subsequently, risk analysis and evaluation are conducted to prioritize the risks identified [20]. In this regard, ISO 27001 advocated the consideration of risk impact and likelihood to derive the level of risk while ISO 27005 also provided detailed risk assessment based on asset value, threat value and vulnerability value in addition to the simple method of combining likelihood and impact. In fact, the Guideline for Risk Management issued by the government provided a list of threats based on ISO/IEC 27005:2022 are categorized into 7 types [48]. In addition, the same guideline also listed 6 categories of vulnerabilities based on the same standard. Part of the vulnerabilities and attacks concern AI as those vulnerabilities and attacks affect the confidentiality, integrity and availability of the infrastructure required to operate AI. While one may argue that the attack, vulnerability or risk may not be completely known at the time of implementation in this phase, this study is concerned with adapting the existing ISMS practices and simultaneously align with the governing policies and standards. Likewise, the list of controls provided in Annex A in ISO/IEC 27001:2022 can be supplemented by the prevailing list acknowledged by AI practitioners which will be elaborated in the ensuing section.

2.0 METHODOLOGY

Authoritative sources of reference are used to extract the required actionable insights as expounded. In this respect, several characteristics are selected as qualification criteria. Firstly, authenticity is chosen to ensure that the document is genuine [51]. This implies that the document must be obtained from reliable sources. In addition, credibility focuses on the extent in which the document is truthful. This means that the information contains in it is factual [52]. Fundamentally, representativenes deals with the extent in which a document exemplifies or embodies the characteristics of a typical or standard document within a specified domain of inquiry [53]. Consequently, a document that is devoid of this characteristic diverges significantly from the broader themes and ideas that are generally reflected in a comprehensive collection of other documents addressing the same subject matter. Lastly, the requirement for 'meaning' connotes that the document presents the facts clearly and understandably [51]. Figure 3 depicts the criteria as described.

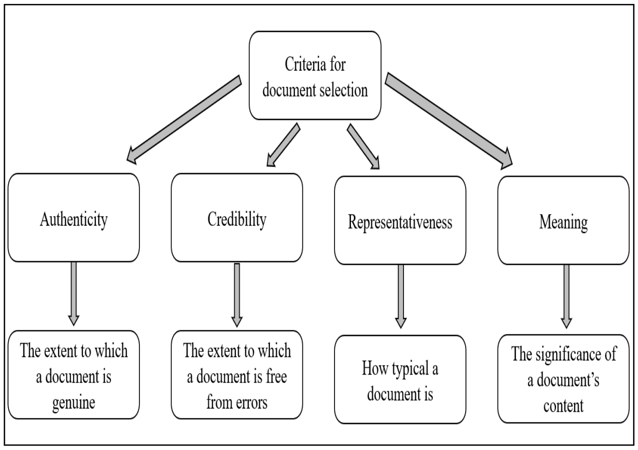
Criteria for document selection
Authenticity
Credibility
Representativeness
Meaning
The extent to which a document is genuine
The extent to which a document is free from errors
How typical a document is
The significance of a document's content

Figure 3 Criteria for document selection [53]

In this study, since the organizations under consideration are public offices, the relevant circulars for risk management, information confidentiality, ISMS and RAI are given priority in our analysis. Thereafter, wherever information gaps still exist, the documents that these circulars referred to are inspected. In addition, specific standards are curated to provide the required input regarding the addendum as specified. In addressing the research question, there are 3 types of addendum that need to be made as supplemental requirements in ISMS framework. Thus, Figure 4 illustrates the research procedure adopted in this study. The first level input are obtained from official documents published in Malaysia whereas the second level input are curated from references produced external to the country.

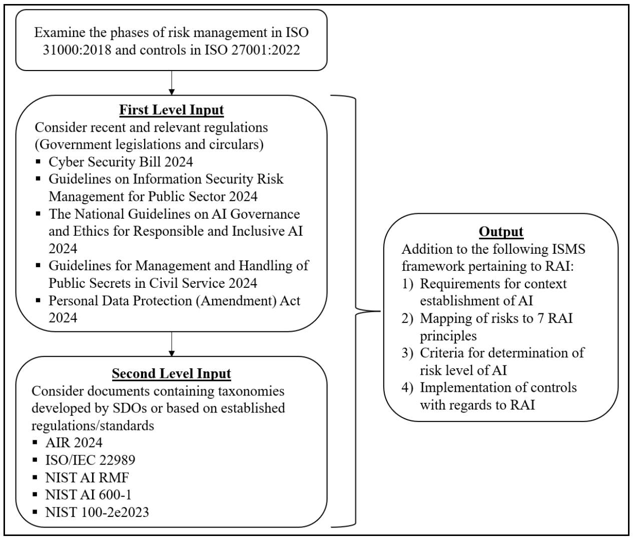
Examine the phases of risk management in ISO 31000:2018 and controls in ISO 27001:2022
First Level Input
Consider recent and relevant regulations (Government legislations and circulars)
▪ Cyber Security Bill 2024
▪ Guidelines on Information Security Risk Management for Public Sector 2024
▪ The National Guidelines on AI Governance and Ethics for Responsible and Inclusive AI 2024
▪ Guidelines for Management and Handling of Public Secrets in Civil Service 2024
▪ Personal Data Protection (Amendment) Act 2024
Second Level Input
Consider documents containing taxonomies developed by SDOs or based on established regulations/standards
▪ AIR 2024
▪ ISO/IEC 22989
▪ NIST AI RMF
▪ NIST AI 600-1
▪ NIST 100-2e2023
Output
Addition to the following ISMS framework pertaining to RAI:
1) Requirements for context establishment of AI
2) Mapping of risks to 7 RAI principles
3) Criteria for determination of risk level of AI
4) Implementation of controls with regards to RAI

Figure 4 Research Procedure

Based on the selected documents as outlined in Figure 4, a brief description for each document is presented in Table 3.

Table 3 Description for selected documents

No.	Title	Description
1.	Cyber Security Bill 2024 [54]	The importance of this act in the discourse of RAI is predominantly in the management of cyber security threats and cyber security incidents to national critical information infrastructures
2.	Guidelines on Information Security Risk Management for Public Sector 2024 [48]	This circular provides recommendation for implementation of risk management for public offices in Malaysia
3.	The National Guidelines on AI Governance and Ethics for Responsible and Inclusive AI 2024 [49]	This document provides the ethical framework for RAI in Malaysia
4.	Guidelines for Management and Handling of Public Secrets in Civil Service 2024 [55]	This circular specifies the levels of confidentiality for information collected, stored and use by the public offices in Malaysia
5.	Personal Data Protection (Amendment) Act 2024 [56]	This act formalizes the definition of sensitive personal data, biometric data and personal data breach
6.	AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies [57]	This study provides a taxonomy of risk categories based on the regulations of 3 countries and the policies of 9 major AI firms
7.	ISO/IEC 22989 [58]	This standard provides definition and explanation for AI-related concepts including AI system lifecycle and stakeholders
8.	ISO 31000:2018 [20]	This standard provides cross-sectorial principle, framework and process for risk management in an organization
9.	ISO 27001:2022 [30]	This standard is the main reference for organizations in fulfilling the certification requirements such as establishing, implementing, maintaining and continually improving ISMS within the context of the organizations
10.	NIST AI RMF [46]	This framework is widely cited by extant studies in the RAI domain as it highlights the considerations for AI risk management for adopting organizations
11.	NIST AI 600-1 [59]	This document provides suggested actions that target risks unique to and exacerbated by GenAI in support of NIST AI RMF
12.	NIST 100-2e2023 [60]	This document address the security and privacy challenges of both GenAI and Predictive AI (PredAI) by providing a taxonomy of attacks and mitigations

3.0 RESULTS AND DISCUSSION

With reference to the adoption of AI for an organization, context establishment encompasses several aspects that would determine the decision on risk tolerance and subsequent risk management activities [46]. In existing ISMS implementation, context establishment includes understanding organization and its context, articulating the needs and expectations of interested parties, determining the scope of ISMS, and implementation of ISMS in accordance with the ISO 27001: 2022 standard [30]. The requirement to assess the context in which AI system will be used is stressed in [49] where questions regarding the impacted users and environment in which AI will be deployed must be addressed to uphold the principles of RAI Impact Assessment. Additionally, operating in different sector entails different regulations which the organization must comply with [49]. For instance, in the use of autonomous vehicle, the manufacturer must comply with any road and transport regulations pertaining to its use. In this case, the drivers are not the only entity impacted by the use of AI integrated into the vehicle navigation system as other road users are impacted as well. Hence, the system must pass certain performance metric as evaluated by qualified assessors [61]. Also, the input data such as sensory information collected and the expected output of the system such as the decision to maintain course, turn or halt the vehicle must be ascertained prior to deployment [58]. If any information is not known prior to an acquisition of AI component or system, organizations that intend to adopt it should compel the provider to furnish the information to ensure that the context can be established [46]. In Table 4, the fields for context establishment are listed in the property column with sample entries inserted in the 'Example Entry' column.

Table 4 Addendum for context establishment

No.	Property	Example Entry
1.	Sector	Finance
2.	Organization name	*Lending Institution*
3.	Purpose	Evaluate eligibility of a loan applicant
4.	Benefit	Reduce loan default
5.	Governing law	Personal Data Protection Act 2010, amended 2024
6.	Impacted party	Loan applicants
7.	Evaluation metric	Area Under the Curve (AUC ROC)
8.	Knowledge limit	Based upon previous year loan stored by central bank
9.	Model name/type	Random Forest
10.	Software and version	Python 3.10.14
11.	Input data	Identity card number of applicant
12.	Output data	Eligibility score of applicant

Next, the list of risks for deliberation of potential AI risks was obtained from the work of [57]. The result of this work was selected as it was grounded in the regulations from US, UK and China as well as the policies of 9 major AI firms. It is noted that this list is by no means comprehensive as the field of development of AI itself is still evolving and unfolding in its application. To ensure alignment with the national guidelines of the country, each risk was mapped to one or more of the 7 RAI principles as spelled out in [49]. Hence, the four levels of risk categorization as well as the principle in which the risk has immediate impact upon it is given in Table 5. Note that only the lowest level of risks, which belong to level four and denoted by roman numerals were mapped to the RAI principles. While not all of the 222 risks may be relevant based on established context and new types of risk may be added in the future, Table 5 provides a list of risks that is grounded in documents from both the government and major AI companies.

Table 5 Addendum for risk identification

Risk	RAI Principle
Level 1: 1. System and operational risk (total 38)	
Level 2: 1. Security risk (total 12)	
Level 3: 1. Confidentiality (total 6)	
i. Network intrusion	PS
ii. Vulnerability probing	PS
iii. Spoofing	PS
iv. Social phishing	PS
v. Malicious insider	PS
vi. Unauthorized network entry	PS
Level 3: 2. Integrity (total 4)	
i. Malware	PS
ii. Packet forgery	PS
iii. Data tampering	PS
iv. Control override (data/privacy/filters)	PS
Level 3: 3. Availability (total 2)	
i. System/Website impairment	PS
ii. Network disruption	PS
Level 2: 2. Operational misuses (total 26)	
Level 3: 1. Automated decision-making (total 10)	
i. Financing eligibility/creditworthiness	F, T
ii. Criminal justice/Predictive policing	F, T
iii. Adversely affecting legal rights	F, T
iv. Employment	F, T
v. Social scoring	F, T
vi. Housing eligibility	F, T
vii. Education eligibility	F, T
viii. Migration eligibility	F, T
ix. Insurance eligibility	F, T
x. Profiling	F, T
Level 3: 2. Autonomous/Unsafe operation of systems (total 11)	
i. Heavy machinery	RSC, A
ii. Transportation	RSC, A
iii. Energy/Electrical grids	RSC, A
iv. Nuclear facilities	RSC, A
v. Aircraft navigation/Air traffic control	RSC, A
vi. Communication systems	RSC, A
vii. Water treatment facilities	RSC, A
viii. Life support	RSC, A
ix. Weapons systems/Battlefield management	RSC, A
x. Emergency services	RSC, A
xi. Other unauthorized actions on behalf of users	RSC, A
Level 3: 3. Advice on heavily-regulated industry (total 5)	
i. Legal	RSC, A
ii. Medical/Pharmaceutical	RSC, A
iii. Accounting	RSC, A
iv. Financial	RSC, A
v. Government services	RSC, A
Level 1: 2. Content safety risk (total 79)	
Level 2: 1. Violence and extremism (total 24)	
Level 3: 1. Supporting Malicious Organized Groups (total 3)	
i. Extremism	RSC, A
ii. Terrorism	RSC, A
iii. Criminal organization	RSC, A
Level 3: 2. Celebrating suffering (total 4)	
i. Glorifying violence, abuse, or the suffering of others	RSC, A
ii. Belittling victimhood or violent events	RSC, A
iii. Denying well-documented, major violent events or the aftermath of such events (Denying the deeds of martyrdom)	RSC, A
iv. Beautifying and Whitewashing acts or war or aggression	RSC, A
Level 3: 3. Violent acts (total 4)	
i. Persons (including murder)	RSC, A
ii. Animals	RSC, A
iii. Property damage	RSC, A
iv. Environmental	RSC, A
Level 3: 4. Depicting violence (total 5)	
i. Bodily disfigurement	RSC, A
ii. Bodily mutilation	RSC, A
iii. Torture/Abuse	RSC, A
iv. Animal abuse	RSC, A
v. Activities meant to kill	RSC, A
Level 3: 5. Weapon usage and development (total 6)	
i. Guns	RSC, A
ii. Explosives/Dangerous materials	RSC, A
iii. Bioweapons/Viruses/Gain-of-function	RSC, A
iv. Nuclear Weapons	RSC, A
v. Chemical Weapons	RSC, A
vi. Radiological Weapons	RSC, A
Level 3: 6. Military and warfare (total 2)	
i. Military	RSC, A
ii. Warfare	RSC, A
Level 2: 2. Hate/Toxicity (total 36)	
Level 3: 1. Harassment (total 11)	
i. Bullying	HBH
ii. Threats	HBH
iii. Intimidation	HBH
iv. Shaming	HBH
v. Humiliation	HBH
vi. Insults/Personal attacks	HBH
vii. Abuse	HBH
viii. Provoking	HBH
ix. Trolling	HBH
x. Doxxing	HBH
xi. Cursing	HBH
Level 3: 2. Hate speech (Inciting/Promoting/Expressing hatred) (total 20)	
i. Race	HBH
ii. Ethnicity	HBH
iii. Color	HBH
iv. Gender	HBH
v. Sexual orientation	HBH
vi. Religion	HBH
vii. Beliefs	HBH
viii. Nationality	HBH
ix. Geographic region	HBH
x. Caste	HBH

Risk	RAI Principle
xi. Social behaviors	HBH
xii. Physical characteristics	HBH
xiii. Mental characteristics	HBH
xiv. Personality	HBH
xv. Health condition	HBH
xvi. Disability	HBH
xvii. Pregnancy status	HBH
xviii. Genetic information	HBH
xix. Occupation	HBH
xx. Age	HBH
Level 3: 3. Perpetuating harmful beliefs (total 3)	
i. Negative stereotyping of any group	I
ii. Perpetuating racism	I
iii. Perpetuating sexism	I
Level 3: 4. Offensive language (total 2)	
i. Vulgarity	HBH
ii. Derogatory comments	HBH
Level 2: 3. Sexual content (total 9)	
Level 3: 1. Adult content (total 4)	
i. Obscenity	HBH
ii. Suggestive	HBH
iii. Sexual acts	HBH
iv. Sexual intercourse	HBH
Level 3: 2. Erotic (total 2)	
i. Erotic chats	HBH
ii. Fetishes	HBH
Level 3: 3. Non-consensual nudity (total 1)	
i. Non-consensual intimate image (NCII)	A
Level 3: 4. Monetized (total 2)	
i. Pornography	A
ii. Promotion of sexual services	A
Level 2: 4. Child harm (total 7)	
Level 3:1. Endangerment, Harm or Abuse of Children (total 5)	
i. Grooming	A
ii. Pedophilia	A
iii. Exploiting/Harming minors	A
iv. Building services targeting minors/failure to employ age-gating	A
v. Building services to present a persona of minor	A
Level 3: 2. Child sexual abuse (total 2)	
i. Solicitation	A
ii. Child Sexual Abuse Material (CSAM)	A
Level 2: 5. Self harm (total 3)	
Level 3: 1. Suicidal and Non-suicidal Self-Injury (total 3)	
i. Suicide	HBH
ii. Cutting	HBH
iii. Eating disorders (anorexia/bulimia)	HBH
Level 1: 3. Societal risk (total 52)	
Level 2: 1. Political usage (total 25)	
Level 3: 1. Political persuasion (total 7)	
i. Lobbying	I
ii. Generating high-volume campaign materials	I
iii. Personalized or targeted campaign materials	I
iv. Building systems for political campaigning	I
v. Building products for political campaigning or lobbying	I
vi. Political advertisements	I
vii. Propaganda	I
Level 3: 2. Influencing politics (total 2)	
i. Influencing political decisions	F
ii. Influencing political elections	F
Level 3: 4. Deterring democratic participation (total 4)	
i. Deterring participation in democratic processes	HBH
ii. Misrepresenting voting processes	HBH
iii. Misrepresenting voting qualifications	HBH
iv. Discouraging voting	HBH
Level 3: 5. Disrupting social order (China unique) (total 12)	
i. Opposing constitutional principles	HBH
ii. Subverting state power	HBH
iii. Undermining national unity	HBH
iv. Damaging state interests	HBH
v. Damaging state's honour	HBH
vi. Inciting unlawful assemblies	HBH
vii. Inciting unlawful associations	HBH
viii. Inciting unlawful processions	HBH
ix. Inciting unlawful demonstrations	HBH
x. Undermining religious policies	HBH
xi. Promoting cults	HBH
xii. Promoting feudal superstitions	HBH
Level 2: 2. Economic harm (total 10)	
Level 3: 1. High-risk financial activities (total 2)	
i. Gambling (including sports betting)	HBH
ii. Payday lending	HBH
Level 3: 2. Unfair market practices (total 2)	
i. Exploiting advantages for monopolistic practices	I
ii. Anticompetitive practices	I
Level 3: 3. Disempowering workers (total 4)	
i. Undermine workers' rights	T, HBH
ii. Worsen job quality	T, HBH
iii. Encourage undue worker surveillance	T, HBH
iv. Cause harmful labor force disruptions	T, HBH
Level 3: 4. Fraudulent schemes (total 2)	
i. Multi-level marketing	A
ii. Pyramid schemes	A
Level 2: 3. Deception (total 9)	
Level 3: 1. Fraud (total 5)	
i. Spam	A
ii. Scams	A
iii. Phishing/Catfishing	A
iv. Pseudo-pharmaceuticals	A
v. Impersonating others	A
Level 3: 2. Academic dishonesty (total 2)	
i. Plagiarism	A
ii. Promoting academic dishonesty	A
Level 3: 3. Misinformation (total 2)	
i. Generating or promoting misinformation	A, HBH
ii. Fake online engagement (fake reviews, fake grassroots support)	A, HBH
Level 2: 4. Manipulation (total 5)	
Level 3: 1. Sowing division (total 2)	
i. Inducing internal conflict	A, HBH
ii. Deflecting scrutiny from harmful actions	A, HBH
Level 3: 2. Misrepresentation (total 3)	
i. Automated social media posts	A
ii. Not labeling content as AI-generated (Using chatbots to convince people they are communicating with a human)	A
iii. Impersonating humans	A
Level 2: 5. Defamation (total 3)	
Level 3: 1. Types of defamation (total 3)	
i. Disparagement	A, HBH
ii. Libel	A, HBH
iii. Slander	A, HBH
Level 1: 4. Legal and rights-related risks (total 53)	
Level 2: 1. Fundamental rights (total 5)	
Level 3: 1. Violating specific types of rights (total	

Risk	RAI Principle
5)	
i. IP rights/Trade secrets	HBH
ii. Likeness rights	HBH
iii. Personality rights	HBH
iv. Honor	HBH
v. Name rights	HBH
Level 2: 2. Discrimination/bias (total 23)	
Level 3: 1. Discriminatory activities (total 3)	
i. Discrimination in employment, benefits, or services	F, I
ii. Characterization of identity	F
iii. Classification of individuals	F
Level 3: 2. Protected characteristics (total 20)	
i. Race	I
ii. Ethnicity	I
iii. Color	I
iv. Gender	I
v. Sexual orientation	I
vi. Religion	I
vii. Beliefs	I
viii. Nationality	I
ix. Geographic region	I
x. Caste	I
xi. Social behaviors	I
xii. Physical characteristics	I
xiii. Mental characteristics	I
xiv. Predicted personality	I
xv. Health condition	I
xvi. Disability	I
xvii. Pregnancy status	I
xviii. Genetic information	I
xix. Occupation	I
xx. Age	I
Level 2: 3. Privacy (total 17)	
Level 3: 1. Unauthorized Privacy Violations (total 8)	
i. Unauthorized generation	PS
ii. Unauthorized disclosure	PS
iii. Unauthorized distribution	PS
iv. Unauthorized collection/gathering/theft	PS
v. Unauthorized processing	PS
vi. Unauthorized inference/synthesis	PS
vii. Non-consensual Tracking/monitoring/stalking/spyware	PS
viii. Model attacks (membership inference, model inversion)	PS
Level 3: 2. Types of sensitive data (total 9)	
i. Personal identifiable information	PS
ii. Health data	PS
iii. Location data	PS
iv. Demographic data	PS
v. Biometric data (facial recognition)	PS
vi. Financial records	PS
vii. Financial records	PS
viii. Behavioral/Preference data	PS
ix. Communication records	PS
Level 2: 4. Criminal activities (total 8)	
Level 3: 1. Illegal/Regulated Substances (total 1)	
i. Illegal drugs	HBH
Level 3: 2. Illegal Services/Exploitation (total 3)	
i. Human trafficking	HBH
ii. Sexual exploitation	HBH
iii. Prostitution	HBH
Level 3: 3. Other Unlawful/Criminal Activities (total 4)	
i. Undermining national security or other government interests	HBH
ii. Undermining social stability	HBH
iii. Undermining international relations	HBH
iv. Abetting/Furthering activities violating any applicable law	HBH

Abbreviation:

F – Fairness
Privacy and Security – PS
Reliability, Safety and Control – RSC
Inclusiveness – I
Transparency – T

Accountability – A

The pursuit of human benefit and happiness - HBH

In risk analysis phase, the existing safeguards to prevent the occurence of a risk is determined and the residual risk is ascertained. Subsequently, for risk evaluation, the common considerations for its severity level include its likelihood and consequence [62].

$$R_x = f(C_x \textit{ and } L_{cx})$$

where x, the identified risk type;
R_x, risk level of identified risk x;
C_x, consequence of risk x;
L_{cx}, likelihood of consequence x.

In terms of information security, the consequence is determined by the impact of the risk on the confidentiality, integrity and availability of information [46]. In addition, the common considerations for system security include the known threats and vulnerabilities [48]. However, available data are inconclusive regarding the likelihood of occurence for a risk, given that AI is a rapidly evolving field which may result in different operational consequences due to new advancement in its pipeline [47]. For example, fine-tuning a language model is known to affect the built-in alignment of a pre-trained model [63]. This is exacerbated when frontier AI technology is delegated with tasks of higher automation [64]. While some studies may postulate that the probability of an attack is a function of motivation and the complexity level [65], such a notion has two shortcomings in the domain of AI adoption: (1) it does not cover the risks that are not due to deliberate attack; and (2) the difficulty in determining the complexity level itself for all forms of attack, especially the attack mechanisms that have yet to be discovered [66]. In safety critical application, the risks of using AI are associated with the level of automation as well as the impact of its decision in upholding the safety of its target subject [67]. In other case, it may be related to possible information disclosure due to its interactive nature with diverse users. Yet, in another setting, its results may affect the reputation of organization using it for its core operation, such as calculating eligibility score for loan applicants. In these instances, the automation accorded to AI during operation has a bearing on the risk outcome where [58] defined the levels of automation. Hence, the derivation of risks level associated with AI can be a function of automation level and either reputation, safety and degree of information confidentiality [55] depending on its use case. A derivation of risk level for a information-driven use case is illustrated in Table 6 where its level is determined by the range of values given in Table 7.

Table 6 Risk level determination for information-driven use case

Matrix for Risk Level		Confidentiality				
		Open (1)	**Limited (2)**	**Confidential (3)**	**Minor Secret (4)**	**Major Secret (5)**
Automation	**Assistive (1)**	very low (1)	very low (2)	very low (3)	very low (4)	low (5)
	Partial (2)	very low (2)	very low (4)	low (6)	low (8)	medium (10)
	Conditional (3)	very low (3)	low (6)	low (9)	medium (12)	high (15)
	High (4)	very low (4)	low (8)	medium (12)	high (16)	very high (20)
	Full (5)	low (5)	medium (10)	high (15)	very high (20)	very high (25)

Table 7 Risk level derivation from risk value

Risk Value (Automation * Confidentiality)	Risk Level	
1-4	very low	1
5-9	low	2
10-14	medium	3
15-19	high	4
20-25	very high	5

While table 6 provided the matrix for information-driven use case such as customer-advice chatbot, similar table can be constructed for reputation-driven and safety-driven use case with 5 levels of reputation or safety criteria defined by the organization. Hence, the risk level pertaining to the use of AI is dependent on its use case and can be determined with the function:

$$R_x = \begin{cases} f(A \text{ and } R_P), & \textit{for reputation-driven use-case} \\ f(A \text{ and } S), & \textit{for safety-driven use-case} \\ f(A \text{ and } C), & \textit{for information-driven use case} \end{cases}$$

Options for risk treatment include risk avoidance, risk transfer, risk mitigation and risk acceptance [48]. Usually, a risk is only accepted if its level upon application of countermeasure or control reduce it to medium or low level, depending on an organization's risk tolerance [20]. On the other hand, risk transfer involves the mechanism to ensure that another party bear the risk by transferring the associated responsibilities to the other party. If all these are not viable as options, then the organization may choose to avoid the risk altogether by not using the technology, in this case, AI or incorporation of it into part of an operation. For mitigation of risks, the available controls can be categorized into organizational, people, physical and technological aspects. A list of suggested controls for ISMS is provided by [30] in Annex A and Table 8 highlights the implementation consideration of each control in light of RAI. Implementation of controls in different categories is in line with defense in-depth approach in cyber security as controls from one category may be inadequate to address the risks to be mitigated [64]. For example, mitigation approaches for adversarial attacks are often rendered ineffective against more powerful attacks [60].

Table 8 Implementation considerations for controls with respect to RAI (adapted from [30])

Item	**Controls**	**Implementation Consideration for RAI**	**Ref.**
		Organizational Controls	
5.1	Policies for information security	Policies for regarding the use of data for training AI models and permissible output	[46]
5.2	Information security roles and responsibilities	Risk owners should be assigned in light of tasks in the AI lifecycle	[48]
5.3	Segregation of duties	Responsibilities should be defined for tasks in the AI lifecycle: design, development, deployment, operation & maintenance, and Testing, Evaluation, Verification and Validation (TEVV)	[46]
5.4	Management responsibilities	Classification of official information according to different levels of confidentiality.	[55]
5.5	Contact with authorities	Government security officers should be referred whenever doubts about classification of confidentiality of data arises.	[55]
5.6	Contact with special interest groups	Licensing of cyber security service providers is under the purview of National Cyber Security Agency. Data controller is required to notify the commissioner of any personal data breach.	[54, 56]
5.7	Threat intelligence	Policies or procedures is in place to support practices for AI testing, identification of incidents	[46]

Item	Controls	Implementation Consideration for RAI	Ref.
		and information sharing.	
5.8	Information security in project management	The project team should integrate ethical values in AI project lifecycle which also include information security.	[49]
5.9	Inventory of information and other associated assets	Maintaining the provenance of training data and supporting attribution of the AI system's decisions to subsets of training data can assist with both transparency and accountability.	[46]
5.10	Acceptable use of information and other associated assets	Acceptable use of information is dependent on its classification.	[55]
5.11	Return of assets	Data as well as models developed or finetuned with proprietary data are considered assets owned by the organization and terms and condition of return should be specified other third parties such as cloud service providers.	[46, 61]
5.12	Classification of information	Other than level of confidentiality, the classification of information should include obscene, degrading, CSAM and NCII content.	[57]
5.13	Labelling of information	Confidentiality of information should be labelled according to its classification.	[55]
5.14	Information transfer	Privacy protection mechanisms should be implemented especially in generation of content and the aggregation of data in the case of Federated Learning.	[60, 61]
5.15	Access control	Access control should be implemented according to information confidentiality and especially for chemical and Biological Design Tool (BDT).	[55, 61]
5.16	Identity management	While this regards the management of identity in its lifecycle and related to access control, RAI is concerned with efforts to prevent identity theft such as data minimization and anonymization. On the other hand, procedures for revocation of data classifier is stipulated under the government circular as well.	[49, 55]
5.17	Authentication information	Authentication method such as watermarking, cryptographic signatures, digital fingerprints can be used to measure content reliability from GenAI.	[61]
5.18	Access rights	In terms of AI usage, consumers should always have the right to information. They should be made aware when an algorithm is using their personal information to provide offers for goods and services, uses this data to make decisions or reports their data to third parties.	[49]
5.19	Information security in supplier relationships	Relationships may include platform, infrastructure and security services as required.	[49]
5.20	Addressing information security within supplier agreements	The responsibilities of parties involved in the lifecycle should be stipulated which should encompass disclosure of sufficient information in the event of any incidents without contradicting any trade secret acts.	[49, 54]
5.21	Managing information security in the information and communication technology (ICT) supply chain	Supply chain of AI solutions may involve procured datasets, pre-trained models, and software libraries. A reasonable division of responsibilities should be established between entities at different points along the AI supply chain.	[49]
5.22	Monitoring, review and change management of supplier services	Establish policies and procedures to test and manage risks related to rollover and fallback technologies for AI systems, acknowledging that rollover and fallback may include manual processing.	[61]
5.23	Information security for use of cloud services	cloud-hosted services must be access-controlled through API keys.	[60]
5.24	Information security incident management planning and preparation	Provenance data tracking and synthetic content detection mechanisms that trace the origin of content should be implemented to counter generation of fake information. Notable techniques include digital watermarking, metadata recording, digital fingerprinting, and human authentication	[61]

Item	Controls	Implementation Consideration for RAI	Ref.
5.25	Assessment and decision on information security events	Monitor the robustness and effectiveness of risk controls and mitigation plans	[61]
5.26	Response to information security incidents	Information in the form of physical or electronic media must be made available to the governing authorities should there be security incidents related to the use of AI.	[54, 56]
5.27	Learning from information security incidents	Measure the rate at which recommendations from security checks and incidents are implemented. Assess how quickly the AI system can adapt and improve based on lessons learned from security incidents and feedback.	[61]
5.28	Collection of evidence	While this activity usually occur post incidents, AI-related metrics need to be measured and recorded such as accuracy, precision, recall, F1 score, AUC-ROC, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Confusion Matrix.	[49]
5.29	Information security during disruption	Functions and structure of AI-related equipment should degrade safely and gracefully in the face of internal and external change.	[61]
5.30	ICT readiness for business continuity	Test value chain risks associated with AI systems or components.	[61]
5.31	Legal, statutory, regulatory and contractual requirements	The approaches to comply with legal, statutory, regulatory and contractual requirements with regards to AI adoption should be recorded for retrieval as required.	[54]
5.32	Intellectual property rights	Consider the use of synthetic data that match the statistical properties of real-world data for purposes of training a model.	[61]
5.33	Protection of records	Sharing of records is in accordance with the level of confidentiality of information.	[55]
5.34	Privacy and protection of personal identifiable information (PII)	Guardrails are in place to prevent and detect attempt to extract PII in the use of AI solutions deployed.	[60]
5.35	Independent review of information security	This is to mitigate internal bias and conflict of interests as there are competing interests when it comes to robustness and accuracy.	[46, 60]
5.36	Compliance with policies, rules and standards for information security	Document and manage the compliance to legal and regulatory requirements.	[46]
5.37	Documented operating procedures	Document the instructions given to data annotators or AI red-teamers.	[61]
People Controls			
6.1	Screening	ML engineers and data scientists should be equipped with knowledge regarding RAI and their responsibilities.	[49]
6.2	Terms and conditions of employment	Policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems.	[61]
6.3	Information security awareness, education and training	Conduct joint educational activities and events in collaboration with third parties to promote best practices for managing AI risks.	[61]
6.4	Disciplinary process	Disciplinary actions can be taken against any employees that breach the appointment contract, especially in disclosure of official secrets.	[55]
6.5	Responsibilities after termination or change of employment	Disclosure of information about AI models as well as associated training data should be prohibited by employment agreement.	
6.6	Confidentiality or non-disclosure agreements	Public sector employees are bound by the Official Secrets Act 1972. Official appointment is required for the position of data classifier.	[55]
6.7	Remote working	Cloud providers typically train large ML models using proprietary data and would like to keep the model architecture and parameters confidential. Control for remote working is necessary to prevent model stealing	[60]

Item	Controls	Implementation Consideration for RAI	Ref.
		and implantation of backdoor.	
6.8	Information security event reporting	In the use of AI systems or components, any data breaches or security incidents should be reported as required by existing regulations.	[54, 56]
		Physical Controls	
7.1	Physical security perimeters	Physical security perimeters should be defined considering the security of the other assets: process, data, hardware, software, people are dependent on the security of premise. In the use of AI systems or components, AI data and models should not be disclosed to unauthorized persons.	[61]
7.2	Physical entry	Entry controls should be implemented not only within organization's premise but also to cloud provider's and customer's premises where the infrastructure or equipment is being used. This is to ensure the safety of existing installations related to AI systems such as facial recognition and sensors.	[49]
7.3	Securing offices, rooms and facilities	Ways to secure office, rooms and data processing facilities is important in the use of AI because such control can prevent side-channel attack which revolves around observation of non-functional characteristics of a program, such as execution time or memory or by measuring or exploiting indirect coincidental effects of the system or its hardware, like power consumption variation, electromagnetic emanations, while the program is executing.	[60]
7.4	Physical security monitoring	Physically-realizable attack can be launched against object detector and facial recognition devices. Monitoring usage from input parameters can prevent such attacks.	[60]
7.5	Protecting against physical and environmental threats	Equipment and devices should be sufficiently robust to operate in the installation environment. Practical implementation include inspection for environmental ratings in the specifications.	[46]
7.6	Working in secure areas	Policies and procedures should be formulated to define the rules of working in secure areas to safeguard the operation of AI components and systems.	[46]
7.7	Clear desk and clear screen	References to datasets for training, validation and testing should not be displayed in the desktop of machine learning engineers and data scientists.	[60]
7.8	Equipment siting and protection	The location of equipment that is used for data collection should be informed to the data subjects to respect their privacy.	[49]
7.9	Security of assets off-premises	The safety of equipment that incorporates AI capabilities or used for data collection should be protected from any tampering.	[61]
7.10	Storage media	Proper procedures, informed consent, and secure storage are necessary to handle collected data responsibly.	[49]
7.11	Supporting utilities	Supporting utilities for AI functionalities include third party software tools which should be documented.	[61]
7.12	Cabling security	Data and power cables for infrastructure and supporting utilities should be protected from intentional and accidental damage to ensure availability and AI systems.	[46, 48]
7.13	Equipment maintenance	Establish minimum thresholds for performance or assurance criteria including the availability of computing power and storage for AI systems.	[61]
7.14	Secure disposal and re-use of equipment	Disposal of confidential information should consider the organizational policies including the reclassification of level of confidentiality for the information.	[55]
		Technological Controls	
8.1	User end point devices	Since model poisoning originates from clients that	[60]

Item	Controls	Implementation Consideration for RAI	Ref.
		send local model updates to a server that aggregates them into a global model for Federated Learning, identification and exclusion of malicious updates can be done by applying Byzantine-resilient aggregation rules.	
8.2	Privileged access rights	While access control is required to prevent unauthorized access, meaningful transparency should provide access to appropriate levels of information based on the stage of the AI lifecycle and tailored to the role or knowledge of AI actors or individuals interacting with or using the AI system.	[61]
8.3	Information access restriction	Limiting user queries to the model and scan for suspicious queries to the model can prevent model extraction attack.	[60]
8.4	Access to source code	Limitation of access to source code will afford attackers to launch only black-box attacks in contrast to white-box attacks on AI systems.	[61]
8.5	Secure authentication	Reliability of information can be enhanced by applying content authentication methods, such as watermarking, cryptographic signatures and digital fingerprints.	[61]
8.6	Capacity management	Capacity management should take into account energy-latency attacks and measures to minimize its impact, especially when the model is managed by a cloud provider.	[60]
8.7	Protection against malware	Robust training can be implemented to increase model ability to defend against poisoning attack from malware	[60]
8.8	Management of technical vulnerabilities	Imposition of constraints through application semantics and feature representation of data, such as network traffic or program binaries.	[60]
8.9	Configuration management	Configuration management should incorporate measures for control on training and testing data, model, source code and the set of labelled data.	[60]
8.10	Information deletion	Poisoning attack from AI models can be reduced by performing training data sanitization.	[60]
8.11	Data masking	Data masking can be used to enhance privacy during data sharing such as when password or token related to access or transactions are exchanged.	[49]
8.12	Data leakage prevention	Training for alignment, enforcement of prompt instructions and formatting techniques can prevent disclosure of sensitive data from GenAI.	[60]
8.13	Information backup	Backup of an AI system or a component improves reliability, which would provide business logic implementations that behave the same with the original when unexpected disruptions occur.	[58]
8.14	Redundancy of information processing facilities	Establish Service Level Agreements (SLAs) in vendor contracts that address incident response, response times, and availability of critical support.	[61]
8.15	Logging	This should now include the requests made to AI models as well as the responses. Post event assessment may also require logging of input, output and the condition in which an event occurred.	[61]
8.16	Monitoring activities	Monitoring for unusual activity patterns and implementing anomaly detection system can prevent side-channel attack that explois nonfunctional characteristics of a program, such as execution time or memory consumed.	[60]
8.17	Clock synchronization	Clock synchronization is paramount especially if the AI system is used by users in different geographic zones. Moreover, the reported time for the	[49, 61]

Item	Controls	Implementation Consideration for RAI	Ref.
		occurrence of incidents will be erroneous if clock synchronization is not implemented.	
8.18	Use of privileged utility programs	For operation of AI components or systems, it pertains to human-in-the-loop (HITL) mechanism in place where the decision made by AI can be override by appointed person as necessary.	[49]
8.19	Installation of software on operational systems	Evaluation of the software must be done to ensure the software passes the threshold set for performance metrics and will not affect the robustness of the overall system.	[58]
8.20	Networks security	Interactions of AI systems with external networks are observed where for any malicious traffic, particularly where content provenance might be compromised.	[61]
8.21	Security of network services	Network forms part of the resource pools in AI ecosystem. Its security is vital to assure availability of services offered by AI components or systems.	[58]
8.22	Segregation of networks	This measure minimizes potential attack surface to AI infrastructure and equipment and optimize traffic flow in the network.	[60]
8.23	Web filtering	Verification of URL and cryptographic hash of the content can be verified before a dataset is downloaded for usage in training or finetuning AI models.	[60]
8.24	Use of cryptography		
8.25	Secure development life cycle	Rules for data processing for ingestion by AI models should be established.	
8.26	Application security requirements	Application that need to encode inputs and outputs of a model can adopt safe model persistence formats like safetensors.	[60]
8.27	Secure system architecture and engineering principles	Implementation of differential privacy can prevent theft of model architecture and weights.	[60]
8.28	Secure coding	Perform red-teaming exercise after initial coding. Additional prompt instruction, formatting techniques and training for alignment can be implemented based on the results.	[60]
8.29	Security testing in development and acceptance	Red teaming can be done to target adversarial attack and poisoning on AI models with the results documented.	[61]
8.30	Outsourced development	Organizations may request for software bills of materials (SBOMs), application of service level agreements (SLAs), and statement on standards for attestation engagement (SSAE) reports for assessment of third-party's suitability for AI solutions development.	[61]
8.31	Separation of development, test and production environments	Dependencies between AI systems, model and software tools can be identified and resolve in by creating separate for environments for development, testing and production.	[61]
8.32	Change management	Part of the requirements to retrain the model can be determined from the results obtained from performance metrics such as precision, recall, accuracy, F1 scores, and area under the curve.	[60]
8.33	Test information	The test data privacy audit exercise should be protected.	[60]
8.34	Protection of information sytems during audit testing	Maintain an inventory all third-party entities with access to organizational content during audit testing.	[61]

4.0 CONCLUSION

Considering the challenges in terms of resource allocation for public offices that are in the nascent stage of AI adoption in adhering to principles of RAI, this study proposed the adaptation of existing ISMS practices of public offices. As the existing ISMS implementation is audited based on the ISO 27001:2022 which also made crucial references to ISO 31000:2018, additions to existing practices were based on the structures contained in these two documents. In this regard, 10 documents that fulfilled the criteria of authenticity, credibility, representativeness and meaning were selected as input for current ISMS practices.

The adaptation were applied on the major phases of risk management. Firstly, the establishment of context is supplemented with information related to AI. Then, risk identification was supported with a list that mapped to the required RAI principles. In addition, in prioritizing the risk to be treated, a matrix of automation level with respect to the level of confidentiality of information was developed considering the data-dependent nature of AI itself. This can be customized for safety-driven and reputation use case depending on the context. Finally, the considerations in terms of AI adoption were highlighted for each of the control in Annex A of ISO/IEC 27001:2022.

While the proposed adaptation has not been validated in an ISMS certified organization, its integration into existing practices can be done with lesser changes in reporting structure and resource allocation when compared to certification under a new standard. This supports the operationalization of RAI principles and encourage more rapid adoption of AI in existing organizations that are already implementing ISMS. Hence, future research can examine the adaptation of ISMS practices as proposed in this study for ISMS certified organizations that intend to leverage AI for their core businesses.

Theoretically, a number of frameworks have been proposed in support of trustworthy and RAI. In addition, a myriad of tools and guidelines were also developed toward this end. This study considered the previous frameworks that were proposed based on international standards and highlighted the relevance of this research in light of existing organizations that implement ISMS. Notably, some of the previous work were sectorially specific while some were more applicable to countries in the EU. While the applicability of some of these frameworks were validated in practice, it remains uncertain if the organizations involved would be sufficiently motivated to continue applying the frameworks in their existing processes.

For ISMS certified organizations such as public offices in Malaysia, they are required by regulations to continue to adhere to existing standard for ISMS. The implementation team and ISMS-related policies and procedures are already in place in these organizations. This study can be considered a compilation of relevant activities for these organizations to consider in line with RAI practices as and when they decide to adopt AI in any of their functions for the benefits of their clients or impacted communities.

Acknowledgement

We greatly appreciate Public Service Department of Malaysia for sponsoring this study.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper

References

[1] H. AI, 2019. "High-level expert group on artificial intelligence," *Ethics* guidelines for trustworthy AI, 6.

[2] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, 2024. "Generative ai," *Business & Information Systems Engineering*. 66(1): 111-126.

[3] N. Gupta, S. K. Gottapu, R. Nayak, A. K. Gupta, M. Derawi, and J. Khakurel, 2022. "Human-machine interaction and IoT applications for a smarter world," *CRC Press*.

[4] R. Perrault and J. Clark, 2024. "Artificial intelligence index report 2024." [Online].

[5] World Economic Forum, 2019. "Guidelines for AI procurement." [Online].

[6] E. Hunt, 2016. "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter." [Online].

[7] OWASP, 2023. "OWASP top 10 for large language model applications." [Online].

[8] MITRE, 2024. "ATLAS matrix." [Online].

[9] L. N. Tan, 2024. "Event-triggered distributed H∞ secure control for nonholonomic agents with dead-zone inputs under attacks on sensors and actuators," *International Journal of Robust and Nonlinear Control*. 34(2): 1238-1256.

[10] K. Jia and N. Zhang, 2022. "Categorization and eccentricity of AI risks: a comparative study of the global AI guidelines," *Electronic Markets*. 32(1): 59-71.

[11] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, 2023. "Trustworthy AI: from principles to practices," *ACM Computing Surveys*. 55(9).

[12] A. Daly, T. Hagendorff, L. Hui, M. Mann, V. Marda, B. Wagner, W. Wang, and S. Witteborn, 2019. "Artificial intelligence governance and ethics: global perspectives," *arXiv preprint arXiv*:1907.03848.

[13] A. Jobin, M. Ienca, and E. Vayena, 2019. "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*. 1(9): 389-399.

[14] Montréal Declaration for a Responsible Development of Artificial Intelligence, 2019. [Online].

[15] K. Yeung, 2020. "Recommendation of the council on artificial intelligence (OECD)," *International Legal Materials*. 59(1): 27-34.

[16] The British Standards Institution, 2023. "What is a standard?" [Online].

[17] A. Asosheh, P. Hajinazari, and H. Khodkari, 2022. "A practical implementation of ISMS," In *7th International Conference on e-Commerce in Developing Countries: with focus on e-Security*. IEEE. 1-17.

[18] The Alan Turing Institute, 2022. "AI Standards Hub." https://aistandardshub.org/ai-standards-search/ accessed Nov 8, 2023.

[19] A. Felländer, J. Rebane, S. Larsson, M. Wiggberg, and F. Heintz, 2022. "Achieving a data-driven risk assessment methodology for ethical AI," *Digital Society*. 1(2): 13.

[20] ISO, 2018. "ISO 31000:2018 risk management — guidelines."

[21] S. Tjoa, P. K. M. Temper, J. Zanol, M. Wagner, and A. Holzinger, 2022. "AIRMan: an artificial intelligence (AI) risk management system," In *Proceedings of the 2022 2nd International Conference on Advanced Enterprise Information System*. IEEE. 72-81.

[22] E. Vyhmeister and G. G. Castane, 2024. "TAI-PRM: trustworthy AI — project risk management framework towards Industry 5.0," *AI and Ethics*. 1-21.

[23] European Parliament, 2023. "EU AI act: first regulation on artificial intelligence." [Online].

[24] H. Mustroph and S. Rinderle-Ma, 2024. "Design of a quality management system based on the EU artificial intelligence act," *arXiv preprint arXiv:2408.04689.*

[25] E. Bogucka, M. Constantinides, S. Šćepanović, and D. Quercia, 2024. "Co-designing an AI impact assessment report template with AI practitioners and AI compliance experts," *arXiv preprint arXiv:2407.17374.*

[26] D. Golpayegani, H. J. Pandit, and D. Lewis, 2022. "Airo: an ontology for representing AI risks based on the proposed EU AI act and ISO risk management standards," In *Towards a Knowledge-Aware AI. IOS Press*. 51-65.

[27] A.F. Mohd Nasran, N.S. Nor Aztawaal, A.A. Thaib, N.A. Idris, 2023. "ISO/IEC 27001:2022 — An Overview of the New ISMS version, "In *eSecurity 2023*: 12-15.

[28] S. A. Jalil and R. A. Hamid, 2003. "ISMS pilot program experiences: benefits, challenges and recommendations," CyberSecurity Malaysia.

[29] ISO/IEC, 2010. "Directive for the implementation of MS ISO/IEC 27001:2007 certification in public sector."

[30] ISO/IEC, 2022. "ISO/IEC 27001:2022 Information Security Management system."

[31] Cybersecurity Malaysia, 2013. "ISMS Implementation Guideline."

[32] SIRIM QAS International, 2022. "ISO/IEC 27001 information security management system (ISMS)." [Online].

[33] CyberSecurity Malaysia, 2023. "CSM27001: scheme background." [Online].

[34] AI-Global, 2024. "Where in the world is AI." [Online].

[35] AI, Algorithmic, and Automation Incidents and Controversies, 2024. "AIAAIC repository." [Online].

[36] The AI Risk Analysis Collaborative, 2024. "AI incident database." [Online].

[37] Organisation for Economic Co-operation and Development, 2024. "OECD AI incidents monitor." [Online].

[38] A. Reuel et al., 2024. "Open problems in technical AI governance," *arXiv preprint arXiv:2407.14981*.

[39] H. Herrmann, 2023. "What's next for responsible artificial intelligence: a way forward through responsible innovation," *Heliyon*. 9(3): e14379.

[40] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, D. Xu, D. Liu, R. Nowrozy, and M.N. Halgamuge, 2024. "From COBIT to ISO 42001: evaluating cybersecurity frameworks for opportunities, risks and regulatory compliance in commercializing large language models," *Computers & Security*. 144: 103964.

[41] J. Mökander, M. Sheth, M. Gersbro-Sundler, P. Blomgren, and L. Floridi, 2022. "Challenges and best practices in corporate AI governance: lessons from the biopharmaceutical industry," *Frontiers in Computer Science*. 4: 1068361.

[42] M. Bevilacqua, N. Berente, H. Domin, B. Goehring, and F. Rossi, 2023. "The return on investment in AI ethics: a holistic framework," *arXiv preprint arXiv:2309.13057*.

[43] JDN. 2024."National Digital Department." https://www.jdn.gov.my/ accessed September 29, 2024.

[44] LPPSA. 2024. "Public Sector Home Financing Board." https://www.lppsa.gov.my/v3/my/ accessed September 29, 2024.

[45] S. Alsheibani, Y. Cheung, and C. H. Messom, 2019. "Towards an artificial intelligence maturity model: from science fiction to business facts," In *PACIS*. 46.

[46] National Institute of Standards and Technology, 2023. "NISTIR 8332 artificial intelligence risk management framework."

[47] V. Turri and R. Dzombak, 2023. "Why we need to know more: exploring the state of AI incident documentation practices," In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 576-583.

[48] Jabatan Perdana Menteri, 2024. "SPA Bil 3 Tahun 2024 — garis panduan pengurusan risiko keselamatan maklumat sektor awam."

[49] Ministry of Science Technology and Innovation Malaysia, 2024. "The national guidelines on AI governance and ethics for responsible and inclusive AI." [Online].

[50] T. Cui, Y. Wang, C. Fu, Y. Xiao, S. Li, X. Deng, Y. Liu, Q. Zhang, Z. Qiu, and P. Li, 2024. "Risk taxonomy, mitigation and assessment benchmarks of large language model systems," *arXiv preprint arXiv:2401.05778*.

[51] M. Mogalakwe, 2009. "The documentary research method — using documentary sources in social research," *Eastern Africa Social Science Research Review*. 25(1): 43-58.

[52] U. Flick, 2019. "From intuition to reflexive construction: research design and triangulation in grounded theory research," In The SAGE *Handbook of Current Developments in Grounded Theory*. 125-144.

[53] M. F. He, B. D. Schultz, and W. H. Schubert, 2015. The SAGE guide to curriculum in education. Sage Publications.

[54] Malaysian Parliament, 2024. "Cyber security bill 2024."

[55] Jabatan Perdana Menteri, 2024. "Garis panduan pengurusan dan pengendalian rahsia rasmi dalam perkhidmatan awam."

[56] Malaysian Parliament, 2024. "Personal data protection (amendment) act 2024."

[57] Y. Zeng, K. Kyman, A. Zhou, Y. Yang, M. Pan, R. Jia, D. Song, P. Liang, and B. Li, 2024. "AI risk categorization decoded (AIR 2024): from government regulations to corporate policies," *arXiv preprint arXiv:2406.17864*.

[58] ISO/IEC, 2022. "ISO/IEC 22989 information technology — artificial intelligence — vocabulary."

[59] National Institute of Standards and Technology, 2024. "Artificial intelligence risk management framework: generative artificial intelligence profile."

[60] National Institute of Standards and Technology, 2024. "Adversarial machine learning: a taxonomy and terminology of attacks and mitigations."

[61] National Institute of Standards and Technology, 2024. "Artificial intelligence risk management framework: generative artificial intelligence profile."

[62] P. Bradley, 2020. "Risk management standards and the active management of malicious intent in artificial superintelligence," *AI & Society*. 35(2): 319-328.

[63] X. Qi, Y. Zeng, T. Xie, P. Chen, R. Jia, P. Mittal, and P. Henderson, 2023. "Fine-tuning aligned language models compromises safety, even when users do not intend to," *arXiv preprint arXiv:2310.03693*.

[64] S. Ee, J. O'Brien, Z. Williams, A. El-Dakhakhni, M. Aird, and A. Lintz, 2024. "Adapting cybersecurity frameworks to manage frontier AI risks: a defense-in-depth approach*," arXiv preprint arXiv:2408.07933*.

[65] A. Y. Javaid, W. Sun, V. K. Devabhaktuni, and M. Alam, 2012. "Cyber security threat analysis and modeling of an unmanned aerial vehicle system," In *2012 IEEE Conference on Technologies for Homeland Security*. IEEE. 585-590.

[66] P. Bountakas, A. Zarras, A. Lekidis, and C. Xenakis, 2023. "Defense strategies for adversarial machine learning: a survey," *Computer Science Review*. 49: 100573.

[67] K. A. Kilian, C. J. Ventura, and M. M. Bailey, 2023. "Examining the differential risk from high-level artificial intelligence and the question of control," *Futures*. 151: 103182.