Journal of Theoretical and Applied Information Technology
31st August 2024. Vol.102. No. 16
© Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A FRAMEWORK FOR ARTIFICIAL INTELLIGENCE RISK MANAGEMENT

DAVID LAU KEAT JIN[1], GANTHAN NARAYANA SAMY[2], FIZA ABDUL RAHIM[3], NURAZEAN MAAROP[4], MAHISWARAN SELVANANTHAN[5], MAZLAN ALI[6] & VALLIAPPAN RAMAN [7]

[1]Researcher, Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Malaysia

[2, 3, 4]Lecturer, Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Malaysia

[5, 6]Lecturer, Faculty of Social Sciences and Humanities, Universiti Teknologi Malaysia

[7]Lecturer, Department of Artificial Intelligence and Data Science, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India

[1]davidkeat@graduate.utm.my, [2]ganthan.kl@utm.my, [3]fiza.abdulrahim@utm.my, [4]nurazean.kl@utm.my, [5]mahiswaran@utm.my, [6]mazlanali.kl@utm.my, [7]valliappan@cit.edu.in

ABSTRACT

Artificial Intelligence (AI) affords tremendous benefits to multiple sectors and businesses as its capabilities extend to different domain of activities. Notwithstanding the benefits that it brings, there are also potential risks which cause concerns by its users and those impacted by its use. Effective risk management is thus essential for organizations planning to deploy AI in high-risk applications. This study introduced a framework developed using a knowledge graph that stores and manages information on risk management, the AI life cycle, and stakeholder involvement, adhering to established standards. The framework facilitated the retrieval and generation of insights that support decision-making related to risk management, as it can represent interrelationships between entities more effectively than relational databases or typographies. The insights that can be generated include distribution of risks according to AI life cycle phases, the countermeasure that could treat the greatest number of risks and the countermeasure that produced the greatest change in terms of impact and probability to the identified risk. In this study, Cypher language was used to develop the framework, while Python language was used to generate the insights from the framework. Future studies may consider the integration of the framework in an enhanced Enterprise Risk Management framework to enable real-time update of related information and response by the organization.

Keywords: *Artificial Intelligence, Risk Management, AI Life Cycle, Stakeholder*

1. INTRODUCTION

Artificial Intelligence (AI) systems are defined as systems that display intelligent behaviour by analyzing their environment and taking actions with some degree of autonomy toward achieving specific goals, often on par or exceed human intelligence [1]. The reasons to harness AI for businesses and governmental functions revolve around its emergent capabilities, such as prediction, classification, association, and optimization which increase the efficiency and quality of decision-making [2]. Determining the required capabilities is essential for selecting the appropriate algorithm.

As AI is data-dependent, inaccurate, biased or intentionally malicious data fed into the algorithmic model may produce inaccurate, biased, and erroneous output resulting in adverse or catastrophic consequences, depending on its actual application. For example, the Twitter chatbot launched by Microsoft was forced to shut down after other Twitter users trained it with racist information, which in turn produced racially offensive and insensitive statements [3]. This issue is exacerbated by the vulnerabilities and methods used to perform adversarial attacks on AI models, published online by the Open Worldwide Application Security Project (OWASP) [4] and MITRE [5]. The 'OWASP Top 10 for LLM applications' is a report outlining adversarial attack on LLMs, while the 'Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) Matrix' is a globally accessible knowledge base of adversary tactics and techniques against AI models.

Statistically, the number of incidents and controversies related to AI is increasing, as reported by AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC), as shown in Figure 1.

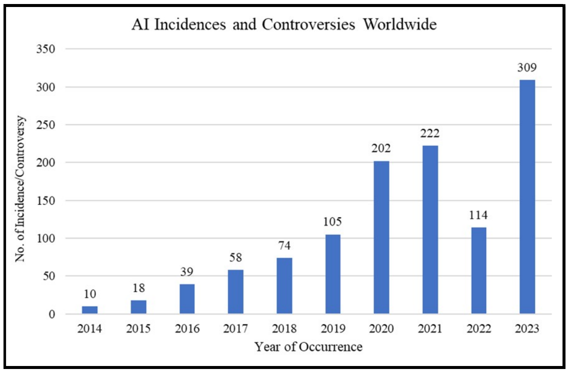


Figure 1: AI incidents and controversies from 2014 till 2023 [6]

Notwithstanding the known incidents and controversies related to AI, the proliferation of AI in multiple sectors is on an increasing trend. Due to its non-deterministic nature in generation of output, the results generated by AI systems cannot be completely assessed like traditional systems. Consequently, AI systems introduce a set of risks that current risk frameworks and approaches do not comprehensively address [7]. This situation prompted the ratification of the inaugural and comprehensive EU AI Act, which adopts a risk-based approach [8]. The legislation stipulates that performing risk management is a requirement for organizations using AI in high-risk applications. Such requirements are necessary, considering the different types and levels of risks involved in the application of AI.

1.1 Considerations for AI risk management

AI risk management aspects were addressed in established risk management standards. According to the ISO 31000 standard, risk is defined as uncertainty on objectives [9]. Risk should fulfil three characteristics [10]. Firstly, risk refers to a potential condition of existence that impacts an individual's welfare, either positively or negatively. Secondly, risk encompasses the uncertainty about the occurrence of such a condition in the future; therefore, events that are definite cannot be associated with risk. Thirdly, risk pertains to a potential state of existence, meaning a state that is impossible cannot be considered a risk. As shown in Figure 2, risk management comprises a series of activities grouped into processes that are executed both sequentially and simultaneously. Sequential processes include establishing the context, identifying risks, analyzing risks, evaluating risks, and treating risks. In contrast, processes implemented simultaneously include communication and consultation, as well as monitoring and review. Table 1 articulates the information required to perform these sequential processes in risk management [9].

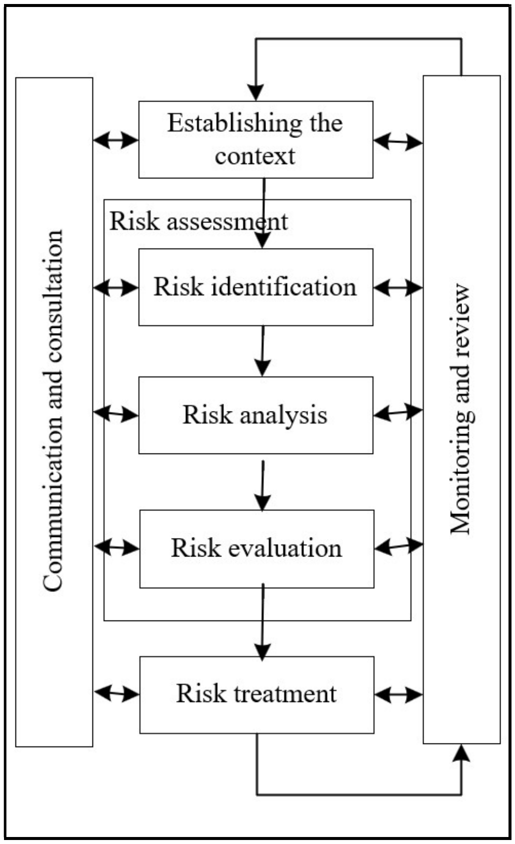


Figure 2: Risk management processes [7]

Table 1: Required information in sequential risk management processes [9]

Risk Management Process	Required Information
Context Establishment (CE)	The inherent risk varies when an AI system is utilized in different sectors and use cases. Hence, it is necessary to consider the sector in which the system operates as well as the applications it supports. For example, the inherent risk involved in providing credit rating estimations for financial institutions differs from that of predicting customer churn for a retail store.
Risk Identification (RI)	Risk can be traced to various sources which is context dependent. In addition, the use of different algorithms, which depends on the specific use cases, also generates various types of risks. In this study, the ethical, technological, data and analytical risks were considered.

Risk Management Process	Required Information
Risk Analysis and Evaluation (RAE)	Risk analysis involves determining the impact, consequence, and probability of each identified risk. On the other hand, risk evaluation pertains to the tasks of risk prioritization. The method used in risk evaluation may involve qualitative or quantitative approach. These two processes were combined in this study.
Risk Treatment (RT)	Although there are several options for risk treatment, risk mitigation was the countermeasure considered in this study as the reduction in impact or likelihood of a risk was the primary interest. The decision for risk avoidance, risk transfer or risk acceptance can be made upon deliberation of acceptable risk mitigation by the user.

Apart from risk management processes, the consideration of the AI life cycle in managing risk was emphasized by the NIST AI Risk Management Framework [7], ISO/IEC 22989 [11] and WEF Procurement Guidelines [12]. A comparative study of four online repositories that record actual incidents and issues related to AI concluded that the causes of AI incidents can be introduced within the system at many stages of the system lifecycle [13]. The study also asserted that AI failures tend to be context-specific, reinforcing the necessity of CE stage as highlighted in Table 1. The AI life cycle spans from its ideation to eventual retirement, encompassing risks and countermeasures that can be implemented at each phase. Hence, Table 2 provides a comparison of the life cycle phases as elucidated by the three frameworks. In this study, six phases chosen for further analysis: plan and design, data preparation, modelling, deployment, maintenance, and retirement, encompassed the scope highlighted by the three theoretical frameworks.

Table 2: Phases in AI life cycle considered

Phases Chosen for AI Life Cycle	[7]	[11]	[12]
Plan and Design	Plan and Design	Inception	Requirements gathering and analysis
Data Preparation		Design and Development	Design
Modelling	Build and Use Model		Implementation and coding
	Verify and Validate	Verification and Validation	Testing
Deployment	Deploy and Use	Deployment	Deployment
Maintenance	Operate and Monitor	Re-evaluate	Maintenance
Retirement	People and Planet	Retirement	

Lastly, the stakeholders responsible for executing various countermeasures must be considered. In fostering meaningful human control, contestability of output made by AI is essential [14] [15]. In this regard, the various control points where human intervention in the form of Human-Before-The-Loop (HBTL), Human-In-The-Loop (HITL) and Human-Over-The-Loop (HOTL) are pinpointed in Figure 3 [16]. This requirement ensures accountability for implementing controls at various stages of the AI life cycle [14, 17]. In fact, [15] defined the requisite controls at different points as Test, Evaluation, Verification and Validation (TEVV). In this study, the stakeholders considered were identified through a comparison with three other articles, as elucidated in Table 3.

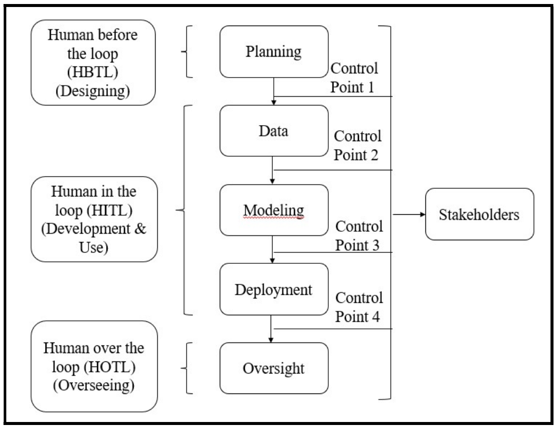

Human before the loop (HBTL) (Designing)
Human in the loop (HITL) (Development & Use)
Human over the loop (HOTL) (Overseeing)
Planning
Data
Modeling
Deployment
Oversight
Control Point 1
Control Point 2
Control Point 3
Control Point 4
Stakeholders

Figure 3: Control Points For Human Intervention In Applied AI [16]

Table 3: AI stakeholders considered

Stakeholders chosen for this study	[7]	[11]	[18]
Management team	C-suite executives		
Procurement team	Procurement experts		
Infrastructure	Modelers	AI	

Stakeholders chosen for this study	**[7]**	**[11]**	**[18]**
provider	Model experts	platform provider	
Software provider	Product managers	AI product or service provider	
Development team	AI designers	AI developer	Development Team
	System engineers	AI system integrator	
	System integrators		
	Software engineers		
Data provider	Data provider	Data provider	Data domain expert
	Data scientist		
Operation team	Domain experts	Data subject	
Impacted individuals	End users	AI user	Decision subject
	Impacted individuals	Other subjects	
	Impacted communities		
	Socio-cultural analysts		
Internal auditors	TEVV experts	AI evaluator	Auditing Team
External auditors	Impact assessors	AI auditor	
Legal advisors	Governance experts	Policy makers	
	Policy makers		
Compliance team	Compliance experts	Regulators	
	Socio-cultural analyst		
	Human factor experts		
	Standards organizations		
	Trade associations		
	Advocacy groups		
	Environmental groups		
	Civil society organizations		
Academia	Researchers		

1.2 Related studies

In the context of high-risk application, autonomous and intelligent systems embedded with AI are the use cases that garner the most attention. Studies were conducted in automotive and medical sector, which resulted in the formulation of a Socio-technical (SOTEC) framework [19, 20]. In this framework, the sources of risks were classified into five categories of structural, organizational, technological, epistemic, and cultural. According to the author, structural sources of risk were associated with the interdependencies and interactions between different technical and social structures. In addition, organizational sources of risk arise from social processes, organizing activities, human and contextual factors. Apart from that, technological sources of risk were due to capabilities, affordances, and constraints produced in and by material technologies. Also, epistemic sources of risk pertain to how knowledge and ignorance were constructed whereas cultural sources of risk were attributed to collective values, beliefs, norms, and practices. Similarly, Wirtz, et al. [21] categorized the risks of AI into six different groups which include technological, data, and analytical AI risks, informational and communicational AI risks, economic AI risks, social AI risks, ethical AI risks as well as legal and regulatory AI risks. However, both studies did not arrange the risks into processes that facilitate risk management. Notably, risk analysis and evaluation were omitted in both studies.

Further divergent in risk articulation was provided by Habbal, et al. [22]. In conceptualization of the Trust, Risk and Security Management (TRiSM) framework, the author classified the management of the threat vectors into AI trust, risk and security. Subjectively, AI trust and security can be considered two different types of risk that require examination in the implementation of risk management. While the author suggested improvements in phases of AI life cycle, these recommendations were insufficient as they exhibited the same shortcomings as the previous frameworks. A more comprehensive framework was provided by Golpayegani, et al. [23] in proposing the Health AI Risk Taxonomy (HART) which consists of risk source, risk, consequence, impact, areas of impact, AI technique, AI application, purpose, and stakeholder. However, relationship between all these categories to specific phases of AI life cycle and risk management process were still missing.

One of the studies that mapped the risks and countermeasures to the phases in AI life cycle distinctly was the work by Shahriar, et al. [24] in elaboration of privacy risks of AI. Arguably, the author also stated other risks in his exposition of the privacy risks with the inclusion of inaccuracy and non-transparency risks. However, this study did not recommend any solutions in the event of conflicting risks such as transparency and privacy or accuracy and privacy. Likewise, the study produced a list of risks and the associated countermeasures without sufficient reference to use cases and disregard any requirements for risk prioritization which is part of risk analysis and evaluation. Albeit rarely expounded in the literature, risk analysis and evaluation were discussed by Breier, et al. [25] and Moghadasi, et al. [26]. Nevertheless, both studies did not provide validation for the usage of the proposed methods in actual environment and may not be scalable to all AI use cases.

The requirements for AI life cycle considerations are more pronounced in the evaluation of security for AI models, whether it involves ML, DL or LLM and its application. This was evident from the studies related to model robustness, security, and adversarial attack [25, 27-30]. Collectively referred to as AI security, these studies dissected the mechanisms to perform the attacks that can affect the application in terms of confidentiality, integrity, and availability as well as generation of harmful response in the case of recommendation systems and chatbot-related applications. However, these studies only covered security-related risks and did not provide methods to prioritize the risks of different attacks. Prioritization of risks is crucial to justify the measures for risk treatment as resources are involved in the implementation of risk treatment. Table 4 illustrates the emphasis and coverage of previous studies in the context of risk management and AI life cycle.

Table 4: Consideration of AI life cycle phases and risk management processes from previous studies

No.		AI Life Cycle Phase				
	Article	Plan & Design	Data Preparation	Modelling	Deployment	Maintenance
1.	[19]	CE, RI, RT	-	-	-	-
2.	[20]	CE, RI, RT	-	-	-	-
3.	[21]	CE, RI	-	-	-	-
4.	[22]	CE, RI, RT	RI, RT	RI, RT	RI, RT	RI, RT
5.	[23]	CE,RI				
6.	[24]	RI, RT	RI, RT	RI, RT	RI, RT	RI, RT
7.	[25]	RAE	RAE	RAE	RAE	RAE
8.	[26]	RI, RA	RI, RAE	RI, RAE	RI, RAE	RI, RAE
9.	[27]	CE,RI	RI	RI	RI	RI
10.	[28]	CE	-	RI, RT	RI, RT	-
11.	[29]	CE	-	RI, RT	RI, RT	-
12.	[30]	CE,RI, RT	RI, RT	RI,RT	RI,RT	-

Note:
Retirement phase was omitted as none of the previous studies provided any input in that phase.

Abbreviation:
CE: Context Establishment
RI: Risk identification
RAE: Risk Analysis and Evaluation
RT: Risk Treatment

As evident from Table 4, most of the studies examined the risks of AI without examining the full gamut of risk management processes. There are interrelationships between the risk management processes which cannot be articulated in a typological structure. For example, in the context of using an LLM as a customer service chatbot, the organization may choose an open source LLM and set up the application using a Retrieval Augmented Generation (RAG) method [28]. In this case, some of the risks include harmful content and hallucinations where the corresponding mitigation techniques may include harmful content detection and defensive prompt design [29]. These set of risks and the associated mitigation strategies are different from the application of AI in a clinical setting used to detect malignant growth from radiologic images [31], for example. Moreover, there are instances where a mitigation strategy can reduce multiple risks concurrently and these relationships cannot be captured by the existing high-level structure [26]. Also, the visibility of information and interrelationships between dimensions of risk management processes, phases in AI life cycle and associated stakeholders are required for a risk-based acquisition framework because the specifications and responsibilities of different parties need to be stated unambiguously. This requirement was echoed by the inaugural European Union (EU) AI Act which specified the roles of “operators” which include providers, deployers, product manufacturers, authorized representatives, importers, distributors and downstream providers [8].

Recently, a survey of 277 curated respondents across 229 business organizations concluded that there is a gap between identified risks and solutions designed and implemented [32]. The same study also highlighted the need for an Enterprise Risk Management (ERM) framework in which available solutions are designed beforehand and ready for application. In line with this requirement, this study aims to address the following gaps:

i. How to incorporate risk management processes and life cycle phases in a framework?
ii. How to prioritize the risks and associated countermeasures in the context of an AI application?

From the perspective of an ERM framework, this study proposed a model for the construction of a risk assessment tool and the risk reference database as illustrated by [32] and reproduce in Figure 4. While [32] focused on ethical risks of AI solution (AIS), this study also addressed technological, data and analytical risks as articulated by [21] as it took the phases of AI life cycle into account.

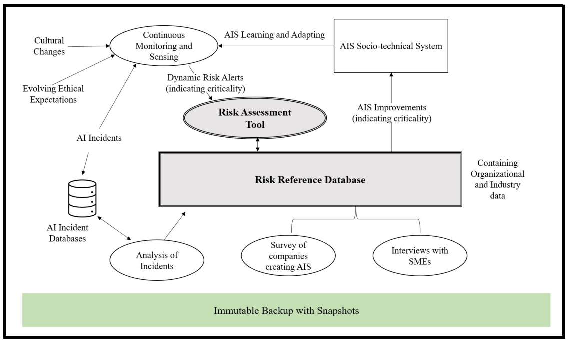

Cultural Changes
Continuous Monitoring and Sensing
AIS Learning and Adapting
AIS Socio-technical System
Evolving Ethical Expectations
Dynamic Risk Alerts (indicating criticality)
Risk Assessment Tool
AIS Improvements (indicating criticality)
AI Incidents
Risk Reference Database
Containing Organizational and Industry data
AI Incident Databases
Analysis of Incidents
Survey of companies creating AIS
Interviews with SMEs
Immutable Backup with Snapshots

Figure 4: Research scope in the enhanced ERM framework

2. METHODS

A graph comprises a collection of nodes and the connections that link them together [33]. Nodes within graph symbolize various entities, while the relationships between them represent the interactions that these entities have with their surrounding environment. This framework enables the representation of diverse scenarios, a concept referred to as ontology. Ontologies possess not only descriptive capabilities but also actionable functionalities. By being stored as nodes and relationships within the graph, one can formulate logical expressions (such as queries or patterns) that navigate through the data plane to the ontology and vice versa, thereby delivering inferential insights. The formation of a Knowledge Graph (KG) occurs when semantic similarities, like taxonomies, are incorporated to introduce additional layers of significance and exploit the underlying data more effectively. Previously, KG was used to optimize manufacturing process [34]. It was also extensively used in cybersecurity for generating alerts based on threat intelligence [35-38]. Furthermore, sectors that applies KG for risk management or assessment include: aviation [39], construction [40], supply chain [41] as well as oil and gas industry [42]. Notably, 'NIST Guide for Conducting Risk Assessment' highlighted that graph-based analysis is an effective way to account for the many-to-many relationships between: (i) threat sources and threat events, (ii) threat events and vulnerabilities and (iii) threat events and impacts/assets [43].

As an illustration of applicability, the use of AI in customer service chatbot for an organization is chosen as the use case, bearing in mind that different context entails different set of risks and countermeasures. The list of risks for this hypothetical chatbot application is shown in Table 5 [28, 29]. On the other hand, the list of controls is presented in Table 6 [28, 29]. It should be noted that while both lists may not be comprehensive for this use case, these were sufficient to demonstrate the effectiveness of the proposed approach in addressing the research questions.

Table 5: List of risks related to chatbot application [28, 29]

No.	Risk	Description	Life Cycle Phase
1.	Harmful content	Biased, toxic, or private information	Deployment
2.	Hallucination	Inaccurate information	Deployment
3.	Inappropriate content	Copyright violation and cyber attacks	Deployment
4.	Data leakage	Personally identifiable information or classified information is leaked	Deployment
5.	Software vulnerabilities	Vulnerabilities in the libraries used	Deployment
6.	Hardware overload	Insufficient capacity	Plan & Design
7.	Injection of factual errors	External tools compromised	Plan & Design
8.	Token limit	Occur when external models are used	Plan & Design
9.	Extraction attack	Building substitute models using black-box query access	Deployment
10.	Evasion attack	Leading shifts in	Modelling

No.	Risk	Description	Life Cycle Phase
		model predicts during model inference	
11.	Poisoning attack	Manipulating training data to cause model inference failure	Data preparation
12.	Overhead attack	Maximizing resource consumption to cause a denial of service	Deployment
13.	Inference attack	Using visible attribute data to infer hidden attribute data	Deployment
14.	Not-suitable-for-work (NSFW) prompts	Oriented towards race, religion, royalty, crime, politics, physical or mental harm	Deployment
15.	Adversarial prompts	Goal hijacking, one-step jailbreaks, prompt leaking, multi-step jailbreaks	Deployment
16.	Data drift	The LLM or the internal documents used as reference are outdated	Operation & Maintenance

Table 6: List of countermeasures for identified risks [28, 29]

No.	Risk Treatment	Targeted Risk	Applicable Life Cycle Phase
1.	Detection	Harmful content Hallucinations Poisoning attack	Deployment
2.	Intervention	Harmful Content	Deployment
3.	Watermarking	Extraction attack	Plan & Design
4.	Control-flow integrity	Software vulnerabilities	Plan & Design
5.	Monitoring of utilization	Hardware overload	Operation & Maintenance
6.	Hardware error correction	Hardware overload	Deployment
7.	Differential privacy	Inference attack	Data preparation
8.	Adversarial training	Inference attack	Modelling
9.	Data minimization	Data leakage	Data preparation
10.	Data minimization	Inappropriate content	Data preparation
11.	Data anonymization	Data leakage	Data preparation
12.	Incorporation of guardrails	– Inference attack – Evasion attack – NSFW prompts – Adversarial prompts – Data leakage	Plan & Design
13.	Exploiting external knowledge	Hallucination	Plan & Design
14.	Learning from human feedback	Hallucination	Deployment
15.	Reranking strategy	Hallucination	Plan & Design
16.	Use local models	Token limit	Plan & Design
17.	Traffic monitoring	Overhead attack	Operation & Maintenance
18.	Cleaning training data	Poisoning attack	Data preparation
19.	Multi-agent interaction	Hallucination	Deployment
20.	Improving decoding strategies	Hallucination	Modelling
21.	Safety pre-prompt	Adversarial prompts	Plan & Design
22.	Safety pre-prompt	Injection of factual errors	Plan & Design
23.	Changing input format	Extraction attack	Plan & Design
24.	Adjusting the order of pre-defined prompt	Adversarial prompts	Plan & Design
25.	Keyword matching	Adversarial prompts	Plan & Design
26.	Content classifier	Adversarial prompts	Plan & Design
27.	Evaluation metrics	Data drift	Deployment

In this study, Neo4j Desktop was used for the generation of KG. Firstly, the framework was developed using Cypher query language and stored in the KG. Then, Python programming language was leveraged to construct the logic and generate the results based on the information stored in KG. The complete hardware and software requirements for implementation of this approach are given in Table 7.

Table 7: Specification of hardware and software used

Requirement	Specification
Hardware	
▪ Processor	➢ 13th Gen Intel(R) Core(TM) i7-13700Hx, 2100 Mhz, 16 Cores, 24 Logical Processors
▪ RAM	➢ 16 GB
▪ System Type	➢ 64 -bit operating system, x64-based processsor
▪ Operating System	➢ Windows 11 Pro
Software	
▪ Code Editor	➢ Jupyterlab 3.6.3

Requirement	Specification
▪ Programming Language ▪ Knowledge Graph	➢ Python 3.11.5, Cypher Query Language ➢ Neo4j Desktop 1.5.9

The basic framework consisted of nodes and relationships that incorporated the information required for risk management and AI life cycle. The data population of KG was implemented by importing the required data from an excel file where the column headers denote the properties in KG. Table 8 provides details for the nodes and relationships.

Table 8: Basic structure of nodes and relationships for the knowledge graph

No.	Element	Label	Property	Description
1.	Node	Context	Sector	The industry; example: public, finance, automotive, healthcare
2.			Model	The AI model; example: ML, DL, LLM
3.			Application	The use case example: chatbot
4.		Risk	Name	Type of risk; example: respond with harmful content
5.			LC_Phase	Life cycle phase; example: deployment
6.		Treatment	Name	Type of controls; example: harmful content detection
7.			LC_Phase	Life cycle phase; example: deployment
8.		Stakeholder	Name	Stakeholders responsible for implementation of controls, example: development team
9.	Relationship	AFFECTS	Impact	Impact of risk to the context, example: high
10.			Probability	Probability of risk occurring, example: low
11.		MODIFIES	I_Effect	The effect on consequence of a risk when the risk treatment is applied, example: nil
12.			P_Effect	The effect on probability of a risk when the risk treatment is applied, example: high
13.		RESPONSIBLE_FOR	-	Specifying the relationship between Stakeholder and Treatment

3. RESULTS AND DISCUSSION

The source code for this study is available in http://www.github.com/renaissance2005/AI-Risk-Management. Risk management processes together with the phases in AI life cycle were embedded in the developed KG. Table 9 gives the sample of Cypher commands to create the nodes and relationships in the KG. In total, 42 nodes and 65 relationships were created. This addressed the first research question. The developed KG is illustrated in Figure 5.

Table 9: Sample Cypher commands for KG creation

Element	Label	Cypher Command
Node	Context	MERGE (chatbot: Context {Sector:'Judiciary', Model:'LLM', Application:'Legal Advice Chatbot')
	Risk	MERGE (harm: Risk {Name:'Harmful content', LC_Phase:'Deployment'})
	Treatment	MERGE (detection: Treatment {Name:'Detection', LC_Phase:'Deployment'})
	Stakeholder	MERGE (development:Stakeholder {Name:"Development team"})
Relationship	AFFECTS	MERGE (harm)-[:AFFECTS {Impact:'moderate', Probability:'moderate'}]->(chatbot))
	MODIFIES	MERGE (detection)-[:MODIFIES {I_effect:'moderate', P_effect:'moderate'}]->(harm)
	RESPONSIBLE_FOR	MERGE (development)-[:RESPONSIBLE_FOR]->(intervention)

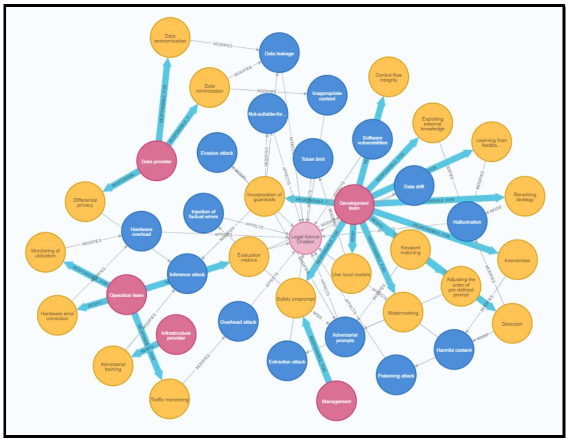
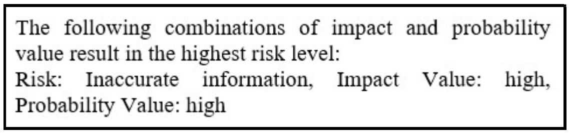


Figure 5: Knowledge graph constructed

In actual use case, the risk manager would enter the values for impact and probability for relationship between risk and context, as well as the corresponding values after the treatment is applied. To prioritize the risks recorded in KG, the values of 'Impact' and 'Probability' for all the 'AFFECT' relationship were considered. The values were matched in accordance with the matrix in Table 10 to determine the risk level [43]. The output from the code that enumerated the risk level is given in Figure 4.

Table 10. Determination of risk level from the matrix of impact versus probability [43]

Probability	Impact				
	Very Low	Low	Moderate	High	Very High
Very High	Very Low	Low	Moderate	High	Very High
High	Very Low	Low	Moderate	High	Very High
Moderate	Very Low	Low	Moderate	Moderate	High
Low	Very Low	Low	Low	Low	Moderate
Very Low	Very Low	Very Low	Very Low	Low	Low



The following combinations of impact and probability value result in the highest risk level:
Risk: Inaccurate information, Impact Value: high, Probability Value: high

Figure 4: Result for risk prioritization

To determine the phase in the AI life cycle that generated the most risks for this use case, the information in the LC_Phase of each node labelled as 'Risk' will be considered. In this regard, a pie chart was generated to illustrate the percentage of risks generated by each phase as shown in Figure 6.

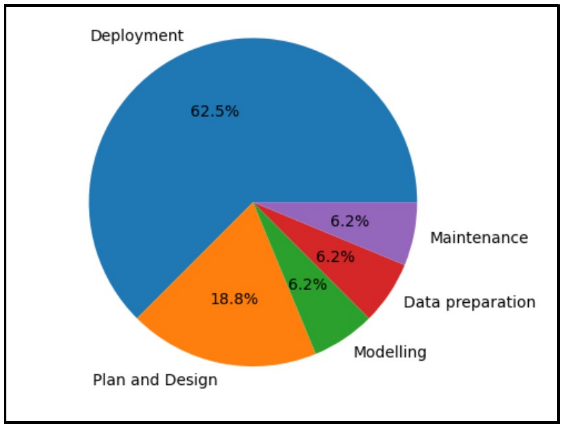


Figure 6: Distribution of risks according to the phases of AI life cycle

Next, to determine the risk treatment that modifies the greatest number of different risks, the number of 'MODIFIES' relationship will be considered. Note that the values for the properties of 'I_effect' and 'P_effect' for this relationship were arbitrarily entered as an example. To enhance clarity, a histogram output was generated as shown in Figure 7 to illustrate the number of treated risks for all the available treatments.

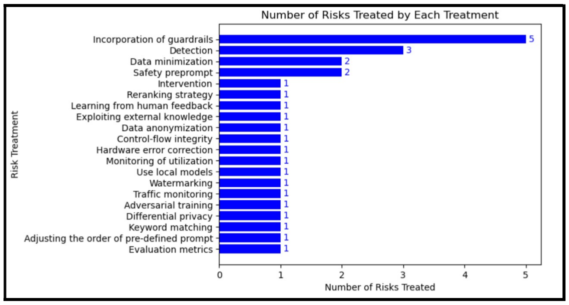
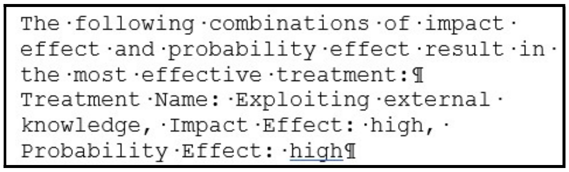


Figure 7: The number of risks treated by each treatment

Lastly, to find out the treatment that has the potential to make the most changes to the impact and probability of any risks, the values for I_effect and P_effect of the 'MODIFIES' relationship is considered. The code listed out all the combinations of the two values for each treatment and highlighted the most effective treatment according to Figure 8.



```
The·following·combinations·of·impact·
effect·and·probability·effect·result·in·
the·most·effective·treatment:¶
Treatment·Name:·Exploiting·external·
knowledge,·Impact·Effect:·high,·
Probability·Effect:·high¶
```

Figure 8: The most effective treatment

Hence, the results addressed the second research question. Note that the KG must be constructed first with the required information entered according to reliable sources such as literature or survey reports.

The KG framework facilitated storage of the required information for risk management in AI, considering the phases of AI life cycle and stakeholders. In addition, the fulfilment of the research questions provided a foundation for the management and practitioners to make informed decisions with regards to risk management. In fact, further questions could be asked based on the information stored in KG such as which stakeholders were involved in each phase of AI life cycle, but the presented results were sufficient to illustrate its utility.

4. RESEARCH CONTRIBUTION AND FUTURE DIRECTION

Notably, relational database management system is not designed for recursive path analysis or enumeration of multi-level relationships [33]. Furthermore, dynamic changes are not practical when the database is integrated with input from external sources, such as the example depicted in Figure 4. Recently, Graph Query Language, a non-proprietary language for KG similar to Cypher was ratified as a standard, underscoring the rising importance of KG in storing information that facilitate further analysis [44]. Moreover, this framework can be further extended to store other useful information such as the cost of treatment. For organizations that deploy more than one application, a new 'Context' node can be created with associated 'Risk' and 'Treatment' linked like the KG in this study. Hence, this framework is scalable for other use cases and further nodes can be added as risk and treatment are known from ongoing research and development.

This study developed a framework that represents the interrelationships between dimensions of risk management processes, phases in AI life cycle and associated stakeholders that could facilitate decision making by an organization. The insights that can be generated include distribution of risks according to AI life cycle phases, the countermeasure that could treat the greatest number of risks and the countermeasure that produce the greatest change in terms of impact and probability to the identified risk. Such insights are useful for acquisition of AI where risk management has been specified as requirements by certain government regulations [8, 45].

Furthermore, the responsibilities of various parties can be specified based on the risk-based approach of this framework. Depending on severity of the potential risk identified, terms and conditions can be specified such that phenotypical information related to AI systems such as the system input and output associated with the solution, logging of end-user behaviour, design of the user interface, training and testing data sets, or model characteristics can be included in contractual agreement [13]. As AI involve various forms of data, technology, tools and services, organizations are bound to perform acquisition to leverage on AI for their businesses or operations. Similar to the service model in cloud computing, AI as a Service (AIaaS) can also be segregated into three layers based on the scope provided by the service providers [46]. Figure 9 illustrates the AIaaS stack [46].

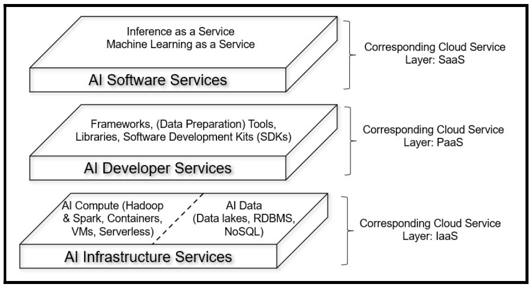

Inference as a Service
Machine Learning as a Service
AI Software Services
Corresponding Cloud Service Layer: SaaS
Frameworks, (Data Preparation) Tools, Libraries, Software Development Kits (SDKs)
AI Developer Services
Corresponding Cloud Service Layer: PaaS
AI Compute (Hadoop & Spark, Containers, VMs, Serverless)
AI Data (Data lakes, RDBMS, NoSQL)
AI Infrastructure Services
Corresponding Cloud Service Layer: IaaS

Figure 9: AI as a Service Stack [46]

Future research may integrate the framework in this study with the other components of enhanced ERM architecture as depicted in Figure 4 and validate its usage with organizations that adopt AI in its core business. As a risk assessment tool in the enhanced ERM architecture, the information on risk criticality and effectiveness of the corresponding treatment would require input regarding the property of impact and/or probability of both nodes. As a risk reference database, new nodes for 'Risk' and 'Treatment' can be created as information is made available from the sources stated in Figure 4.

5. CONCLUSION

As diffusion of AI in multiple sectors gains traction, more risks and the associated treatments will continue to be highlighted in various studies. However, an approach that could clearly map the interrelationships between the processes pertaining to risk management as well as the associations to phases in the AI life cycle are required. Furthermore, scalability of the framework is crucial to keep up with the rapid progress in AI. The use of KG for this scenario fulfils this important and urgent need for AI adoption.

The approach proposed in this study should be empirically validated in organizations that are using AI and intends to manage risks in accordance with existing standards. Moving forward, the integration of KG with language model and dashboard can provide enhanced functionalities related to natural language queries and visibility for organizations in managing AI-related risks.

REFERENCES:

[1] *IEEE guide for terms and concepts in intelligent process automation*, I. C. A. Group, 2017.

[2] D. De Silva and D. Alahakoon, "An artificial intelligence life cycle: From conception to production," *Patterns,* vol. 3, no. 6, 2022.

[3] E. Hunt. "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter." The Guardian. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter (accessed Aug 23, 2023.

[4] OWASP. "OWASP Top 10 for Large Language Model Applications." https://owasp.org/www-project-top-10-for-large-language-model-applications/ (accessed December 30, 2023.

[5] MITRE. "ATLAS Matrix." https://atlas.mitre.org/matrices/ATLAS (accessed 4 April, 2024.

[6] AIAAIC. "AI Index Report." https://spectrum.ieee.org/state-of-ai-2023 (accessed September 18, 2023.

[7] *NISTIR 8332-Artifical Intelligence Risk Management Framework*, NIST, 2023.

[8] E. Parliament. "EU AI Act: first regulation on artificial intelligence." https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed March 2, 2024.

[9] *ISO 31000:2018 Risk management — Guidelines*, ISO, 2018.

[10] E. A. Rosa, "Metatheoretical foundations for post-normal risk," *Journal of risk research,* vol. 1, no. 1, pp. 15-44, 1998.

[11] *ISO/IEC 22989-Information technology — Artificial intelligence — Vocabulary*, ISO/IEC, 2022.

[12] WEF. "Guidelines for AI Procurement." https://www3.weforum.org/docs/WEF_Guidelines_for_AI_Procurement.pdf (accessed 31 May 2023.

[13] V. Turri and R. Dzombak, "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*, 2023, pp. 576-583.

[14] F. Santoni de Sio and G. Mecacci, "Four responsibility gaps with artificial intelligence: Why they matter and how to address them," *Philosophy & Technology,* vol. 34, pp. 1057-1084, 2021.

[15] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," 2023.

[16] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: a review," *ACM Computing Surveys (CSUR),* vol. 55, no. 2, pp. 1-38, 2022.

[17] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective," *Artificial Intelligence,* vol. 279, p. 103201, 2020.

[18] M. Yurrita, D. Murray-Rust, A. Balayn, and A. Bozzon, "Towards a multi-stakeholder value-based assessment framework for algorithmic systems," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 535-563.

[19] C. Macrae, "Managing risk and resilience in autonomous and intelligent systems: Exploring safety in the development, deployment, and use of artificial intelligence in healthcare," *Risk Analysis,* 2024 Jan 2024, doi: 10.1111/risa.14273.

[20] C. Macrae, "Learning from the failure of autonomous and intelligent systems: accidents, safety and sociotechnical sources of risk. SSRN," ed, 2021.

[21] B. W. Wirtz, J. C. Weyerer, and I. Kehl, "Governance of artificial intelligence: A risk and guideline-based integrative framework," *Gov. Inf. Q.,* vol. 39, no. 4, p. 101685, 2022/10/01/ 2022, doi: https://doi.org/10.1016/j.giq.2022.101685.

[22] A. Habbal, M. K. Ali, and M. A. Abuzaraida, "Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions," *Expert Systems with Applications,* vol. 240, 2024 Apr 2024, Art no. 122442, doi: 10.1016/j.eswa.2023.122442.

[23] D. Golpayegani, J. Hovsha, L. W. Rossmaier, R. Saniei, and J. Mišić, "Towards a Taxonomy of AI Risks in the Health

Domain," in *2022 Fourth International Conference on Transdisciplinary AI (TransAI)*, 2022: IEEE, pp. 1-8.

[24] S. Shahriar, S. Allana, S. M. Hazratifard, and R. Dara, "A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle," *IEEE Access,* vol. 11, pp. 61829-61854, 2023, doi: 10.1109/ACCESS.2023.3287195.

[25] J. Breier, A. Baldwin, H. Balinsky, and Y. Liu, "Risk Management Framework for Machine Learning Security," *arXiv preprint arXiv:2012.04884,* 2020.

[26] N. Moghadasi *et al.*, "Risk Analysis of Artificial Intelligence in Medicine with a Multilayer Concept of System Order," *Systems,* Article vol. 12, no. 2, 2024, Art no. 47, doi: 10.3390/systems12020047.

[27] H. Jing, W. Wei, C. Zhou, and X. He, "An Artificial Intelligence Security Framework," in *Journal of Physics: Conference Series*, 2021, vol. 1948, no. 1: IOP Publishing, p. 012004.

[28] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven Failure Points When Engineering a Retrieval Augmented Generation System," *arXiv preprint arXiv:2401.05856,* 2024.

[29] T. Cui *et al.*, "Risk taxonomy, mitigation, and assessment benchmarks of large language model systems," *arXiv preprint arXiv:2401.05778,* 2024.

[30] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," National Institute of Standards and Technology, 2024.

[31] L. Vearrier, A. R. Derse, J. B. Basford, G. L. Larkin, and J. C. Moskop, "Artificial Intelligence in Emergency Medicine: Benefits, Risks, and Recommendations," *Journal of Emergency Medicine,* vol. 62, no. 4, pp. 492-499, 2022 2022, doi: 10.1016/j.jemermed.2022.01.001.

[32] Q. McGrath, A. R. Hevner, and G.-J. de Vreede, "Managing Ethical Risks of Artificial Intelligence in Business Applications," *Authorea Preprints,* 2024.

[33] I. Robinson, J. Webber, and E. Eifrem, *Graph databases: new opportunities for connected data*. " O'Reilly Media, Inc.", 2015.

[34] M. Jawad, C. Dhawale, A. A. B. Ramli, and H. Mahdin, "Adoption of knowledge-graph best development practices for scalable and optimized manufacturing processes," *MethodsX,* p. 102124, 2023.

[35] N. AfzaliSeresht, Y. Miao, Q. Liu, A. Teshome, and W. Ye, "Investigating cyber alerts with graph-based analytics and narrative visualization," in *2020 24th International Conference Information Visualisation (IV)*, 2020: IEEE, pp. 521-529.

[36] L. F. Sikos, "Cybersecurity knowledge graphs," *Knowledge and Information Systems,* vol. 65, no. 9, pp. 3511-3531, 2023.

[37] P. Najafi, A. Mühle, W. Pünter, F. Cheng, and C. Meinel, "MalRank: a measure of maliciousness in SIEM-based knowledge graphs," in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019, pp. 417-429.

[38] A. Pingle, A. Piplai, S. Mittal, A. Joshi, J. Holt, and R. Zak, "Relext: Relation extraction using deep learning approaches for cybersecurity knowledge graph improvement," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 879-886.

[39] Q. Zhao, Q. Li, and J. Wen, "Construction and application research of knowledge graph in aviation risk field," in *MATEC Web of Conferences*, 2018, vol. 151: EDP Sciences, p. 05003.

[40] M. A. Isah and B.-S. Kim, "Development of Knowledge Graph Based on Risk Register to Support Risk Management of Construction Projects," *KSCE Journal of Civil Engineering,* vol. 27, no. 7, pp. 2733-2744, 2023.

[41] E. E. Kosasih, F. Margaroli, S. Gelli, A. Aziz, N. Wildgoose, and A. Brintrup, "Towards knowledge graph reasoning for supply chain risk management using graph neural networks," *International Journal of Production Research,* pp. 1-17, 2022.

[42] J. Canon, T. Broussard, A. Johnson, W. Singletary, and L. Colmenares-Diaz, "A Knowledge-Based Artificial Intelligence Approach to Risk Management," in *SPE Annual Technical Conference and Exhibition?*, 2022: SPE, p. D022S089R001.

[43] R. S. Ross, "Guide for conducting risk assessments," 2012.

[44] ISO/IEC, "ISO/IEC 39075:2024 Information technology Database languages GQL," 2024. [Online]. Available: https://www.iso.org/standard/76120.html.

[45] J. R. Biden, "Executive order on the safe, secure, and trustworthy development and use of artificial intelligence," 2023.

[46] S. Lins, K. D. Pandl, H. Teigeler, S. Thiebes, C. Bayer, and A. Sunyaev, "Artificial intelligence as a service: classification and research directions," *Business & Information Systems Engineering,* vol. 63, pp. 441-456, 2021.