An Enterprise Risk Management Framework to Design Pro-Ethical AI Solutions

Quintin McGrath[1] (✉) [0000-0003-1616-7668], Alan Hevner [2][0000-0003-4953-3900] and Gert-Jan de Vreede[3][0000-0001-6909-9836]

Muma College of Business, University of South Florida, Tampa, FL, USA
[1] qmcgrath@usf.edu, [2] ahevner@usf.edu, [3] gdevreede@usf.edu

Abstract. The wide use of Artificial Intelligence (AI) has immediate business benefits for an organization and its stakeholders through efficiency gains, greater repeatability, and new business models; but the potential exists for unintended ethical consequences. Enterprise Risk Management systems focus on maximizing business value by balancing opportunities and risks. AI system solutions (AISS) elevate risks because of their rapid and unexpected emergent behaviors, as well as their tight integration with the human user and the environment. They are complex socio-technical solutions. This research proposes a unique enhancement to Enterprise Risk Management frameworks focused on the management of ethical risks presented by advanced AISS in a dynamic, pro-ethical, and responsible way; for the benefit of all stakeholders.

Keywords: Design Science Research, Ethical AI, AI Principles, Enterprise Risk Management.

1 Introduction

The effective use of Artificial Intelligence (AI) technologies and systems generates immediate business benefits for an organization and its stakeholders through efficiency gains, greater repeatability, and new business models [1]. But the potential exists for unintended ethical consequences resulting in business and stakeholder harm, like those recorded in the various AI incident databases.[1] Some examples are the use of recruitment management systems used by many businesses to scan high numbers of candidate applications that would otherwise take many hours of work by human recruiters. However, viable candidates may be excluded because their resumes do not match certain criteria used by the algorithm [2]. Another example relates to the use of an AI system by a large online retailer to automatically manage and fire its gig-style workers.[2] While this use of AI is efficient for the business, the unintended consequences of workers

[1] For example, the "AI Incident Database" (https://incidentdatabase.ai/), the "AIAAIC - AI, Algorithmic and Automation Incidents & Controversies" repository (https://www.aiaaic.org/home), and the "AI Global" dataset (https://map.ai-global.org/).

[2] https://www.aiaaic.org/aiaaic-repository/ai-and-algorithmic-incidents-and-controversies/amazon-flex-algorithm-delivery-driver-firings

being terminated by an algorithm with little human intervention or chance for appeal has a negative impact on the business and its stakeholders.

Ethics is the lens through which the rightness and wrongness of decisions and business practices are evaluated. So, for long-term benefit, the development of AI System Solutions (AISS) should be guided by ethical thinking to ensure the greatest positive benefit and the least possible harm to the business and its stakeholders [3]. Deviating from what is generally considered right results in an ethical risk. This balancing act is common for business leaders who must assess opportunities and their associated risks using Enterprise Risk Management (ERM) frameworks [4, 5]. ERMs maximize business value by assessing the probability of the occurrence of a positive or negative event and managing its anticipated impacts. For the business leader, AISS are complex, adding new risks and opportunities. This is not only because of AI's complexity (e.g., self-learning potential, intelligent capabilities, and inscrutability) but also because of its emergent behaviors and tighter integration with human users in a socio-technical system. Because of this complexity and the dynamic, emergent behavior of AISS, we contend that a more nuanced ERM framework with a pro-ethical focus is sorely needed.

This research in progress paper proposes the dynamic management of AISS-related AI Ethical Risks to enhance an organization's risk posture. An iterative Design Science Research (DSR) approach will be used with the research objective "to build an enhanced Enterprise Risk Management (e-ERM) which has the goal of maximizing business and stakeholder benefit through effectively managing the AI Ethical Risks presented by advanced AISS."

2 Background

There is a growing body of research on the link between process design to differing desired outcomes [6, 7]. In addition, we draw from research on the power of information systems to support sustainable goals [8] and apply it to the achievement of ethical business practices. In particular, Shneiderman's 15 recommendations and levels of AI governance [9] are relevant to our thinking. Our research builds on this body of research to contribute to the nascent thinking on ERM frameworks for complex, cognitive AI solutions as briefly discussed here.

2.1 Constructs

The main constructs for this research work are risks and risk management, ethics and ethical risks, and the impact of AI on these constructs in the context of Enterprise Risk Management.

Risk and Risk Management. Risk Management is a process that considers risks and opportunities associated with an action and seeks to define the likelihood and impact of each. The aim is to define those actions that have the greatest likelihood of occurrence, achieving the greatest potential benefit with the least potential harm. Enterprise Risk Management (ERM) applies this concept to the enterprise as a whole to maximize

business value, which we contend, includes value to all stakeholders, by considering business opportunities and their associated risks. Steinberg et al. defines ERM as follows:

> Enterprise risk management is a process, effected by an entity's board of directors, management and other personnel, applied in strategy setting and across the enterprise, designed to identify potential events that may affect the entity, and manage risk to be within its risk appetite, to provide reasonable assurance regarding the achievement of entity objectives.[10]

Risk categories associated with ERM are organization-dependent, but broadly they are categorized into financial, operational, strategic and compliance risks [10, 11]. Others [e.g., 12] add reputational risk to the list.

Ethics and Ethical Risk. The field of ethics can be described as consisting of three areas, (1) meta-ethics, which considers universal moral truths and the nature of right, (2) normative ethical theories relating to the principles and standards that are used to judge right from wrong, and (3) applied ethics, which are codes of ethics that are applied to solving ethical dilemmas [13, 14]. Ethical risks are the unexpected negative consequences that arise from actions that are not aligned with an organization's codes of ethics. Francis states that "ethical risk is to be seen as a part of overall risk management" and that "managing ethical risk is an important aspect of managing risk in general" [15]. Thus, we hold that ethical risk is an additional ERM risk category.

The Impact of AI. There are many definitions of Artificial Intelligence and Ruehle [16] summarizes a number of them. As a working definition for this research, we define AI as the ability for an algorithm to perform tasks commonly associated with intelligent beings. Our focus is further on the cognitive abilities of AI which Sheth et al. describe as "ability to simulate human thought process in a computerized model" [17] resulting in 'cognitive services' which are the AI capabilities related to language, speech, vision, knowledge, decision support, and search [17, 18]. Wirtz, Weyerer, and Kehl [19] in defining an integrative framework for the governance of AI, identify six dimensions of AI Risks based on their literature review. These are "(1) technological, data, and analytical (2) informational and communicational, (3) economic, (4) social, (5) ethical, as well as (6) legal and regulatory AI risks." Our research focuses on the dimension of Ethical AI Risks.

As stated above, ethics is the lens through which the rightness and wrongness of business decisions and practices are evaluated. When the use of AI capabilities is considered in the light of ethical theory it is possible to derive a set of principles, called 'AI Ethical Principles,' which can guide the design, development, and use of AI by a business. This research adopts the AI Ethical Principles summarized by Floridi and Cowls [20] of autonomy, beneficence, non-maleficence, justice, and explainability to establish an organizational AI Ethical foundation. Built on these principles are practical guidance for the design, development, and use of AI system solutions that incorporate pro-ethical business practices into the organization.

AI Ethical Risks are the unexpected negative consequences of design, development, and operational actions relating to AISS, where these actions are not in line with an organization's agreed-to AI Ethical Principles or Practices. They therefore create risk for the organization that needs to be managed. We relate Wirtz et al's [19] Ethical AI Risks to our AI Ethical Principles, e.g., from their Table 1 "AI solutions may cause harm to humans" relates to the non-maleficence principle, and "AI cannot reflect human values (e.g., fairness and accountability)" relates to the justice principle. They go on to define Ethical AI Guidelines, which correspond to our AI Ethical Practices. These AI Ethical Risks are incorporated, along with the other Enterprise risks, into the proposed enhanced ERM framework described next.

3 The Enhanced ERM Framework

An Enterprise Risk Management framework identifies the negative effects of risk and manages opportunities and their positive potential [4]. While many ERM models discuss and include opportunities along with risks, the management of opportunities is often not addressed in execution frameworks [21, 22]. To make the most of the advanced AI capabilities, both the opportunities and the underlying risks need to be considered.

The first improvement proposed for the e-ERM relates to the nature of AISS being a socio-technical system consisting of the AI technologies, the humans using the technology, and the interaction between them. As a result, the e-ERM framework needs to account for both technical and social aspects. Second, because of the potential rapidly changing nature of AISS (e.g., through machine learning), along with the unpredictable changes in the social environment (e.g., the public pressures surrounding the use of AI), the e-ERM framework must be agile, dynamic, and responsive.

In conceptualizing the e-ERM (see Fig. 1), of the many ERM frameworks, the ISO 31000:2018 Risk Management Process [21] and the NIST conceptual AI Risk Management Framework [23] are considered the most applicable. The NIST model recognizes the need for integration of risk management into the AISS lifecycle (i.e., pre-design, design and development, test and evaluation, and deployment). The proposed e-ERM framework, like the NIST model, is grounded in the context of the pro-ethical AISS Development Lifecycle. It focuses first on identifying the risks and opportunities within the scope and context of the planned solution. Second, the e-ERM framework uses a structured tool to assess risks and opportunities. It then guides the actions in response to risks and opportunities. The actions are selected from a "Risk Reference Database" that links AI capabilities and applications to emerging ethical best practices, which guide the implementation of the AISS.

Because of the shifting nature of the AISS and its environment, the e-ERM framework needs to dynamically respond to changes. This "dynamic e-ERM engine" is unique and not explicitly incorporated in other frameworks. The underlying governance structures provide the organization and processes necessary for effective risk management. Each of these phases is described in more detail below.

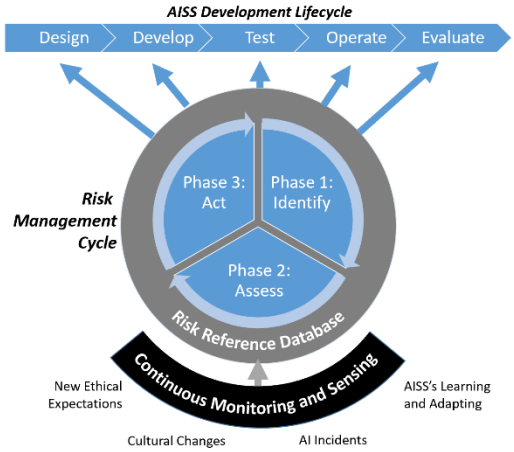

Fig. 1. The e-ERM Framework

3.1 Phase 1: Identify

The first phase defines identifies the business risks and opportunities related to the AISS and its environment. The focus is therefore on the socio-technical aspects of the AISS, the envelope within which it operates, and the AI Ethical Principles that will be used to guide the identification of the AI related ethical risks.

- The AISS as a Socio-Technical System. Asatiani et al. [24] point out that, because of the growth in AI's capabilities, focus has shifted from considering only the technical aspects of an AISS to incorporating the social component. The proposed e-ERM framework sees an AISS as a socio-technical system composed of (a) machine aspects (e.g., the algorithm and data), (b) human actors (e.g., the business goals and decisions associated with the system), and (c) the interaction between the machine, the human, and the associated controls that integrate the two.
- Envelopment. Because of the breadth of the risks and opportunities of AISS there is a need to establish operational boundaries. Nagbøl et al. [25] recommend a bounding approach called "envelopment," based on Robbins [26], for their AI risk assessment tool. In this way, defining what an AISS may and may not do. We use this same envelopment approach to constrain the socio-technical AISS and limit its scope to be within a defined set of the business' goals, system objectives, and stakeholders.
- AI Ethical Principles. This study uses a broad perspective on ethics, considering both not doing the wrong things but also encouraging right actions. When the use of AI is considered in the light of ethical theory, it is possible to derive a set of principles, called AI Ethical Principles which help to frame the risks in the e-ERM framework. The number of sources, analyses, and comparisons of these AI Ethical Principles is growing. Many, like Canca [27], apply the bioethical principles of autonomy, beneficence, non-maleficence, and justice to AI ethics. NIST provides a different perspective in its draft taxonomy of AI risks [28]. They distinguish between "categories" and "principles" with the categories relating to the technical design and socio-technical attributes, while their guiding (bioethical) principles contribute to AISS

trustworthiness. In a similar way we apply the bioethical principles as a foundation across the AISS lifecycle. We define human aspects as the ethical principles and risk of the human part of the socio-technical system, like business goals and fairness. The machine aspects are the ethical principles and risks implemented in the algorithms and data of an AISS. The interaction aspects focus on the risks relating to the AI Ethical Principles of accuracy and reliability of the solution, its transparency, and explainability.

3.2 Phase 2: Assess

As part of the assessment phase of the e-ERM framework, a risk assessment tool is planned. This tool identifies, analyzes, and evaluates risks and opportunities to aid decision-making relating to the socio-technical AISS. An example of such a tool is Nagbøl et al.'s [25] Artificial Intelligence Risk Assessment (AIRA) tool. The AIRA consists of three modules that target specific groups to allow for an efficient risk assessment. The first module considers business needs for the AISS. The second focuses on the technical details of the system. The final module is a synthesis of the outputs of the previous modules balancing the AISS's business and technical aspects.

The outputs from our assessment phase are a clearly defined AISS system, with its context, AI capabilities, and objectives, along with a prioritized list of risks and opportunities that should be acted upon in the next phase. Most important here are the ethical risks associated with the three elements of the socio-technical AISS. This assessment phase provides a static or point-in-time assessment of the planned AISS and is based on known risks. Because of the changing nature of the AISS and its environment, continual reassessment will be needed (see below.)

3.3 Phase 3: Act

With the output from the previous phase, we can enter the next phase where the appropriate action for each of the risks can be determined, e.g., avoid, mitigate, share, transfer, or accept the risks [4]. To enable focused action, we will create a "Risk Reference Database" (RRD). This is based on the analysis of publicly available AI incident databases combined with input on what are considered best practices from subject matter experts. This RRD uses the output from the assessment phase, including AI capabilities, the AISS context and its goals, and the prioritized ethical risks to identify which AI Ethical Principles to focus on and then identifies which emerging AI ethical best practice strategies are most relevant. These strategies guide the designers and developers to implement and improve the AISS.

3.4 Dynamic e-ERM Engine

To be effective, the e-ERM framework must be dynamic and adaptable. This is because of the changing nature of the AISS and its operating environment. Effective monitoring of the operation of the AISS is necessary, enabling identification of emerging risks in both direct and indirect environments. Manual and AI agent-based tracking is needed

to register changes in system behaviors (e.g., becoming biased, privacy exposures, and lower reliability). From a direct environment perspective, active operational monitoring will be needed to identify any changes in outcomes (e.g., failures due to unforeseen inputs, exploitation of vulnerabilities, and activities of adversaries). Also, sensing of the users' sentiments will be monitored. In terms of the broader environment, new AI incidents will be analyzed for new insights into the evolving ethical expectations of the AISS users. The evolution of cultural perspectives on ethics will also be monitored.

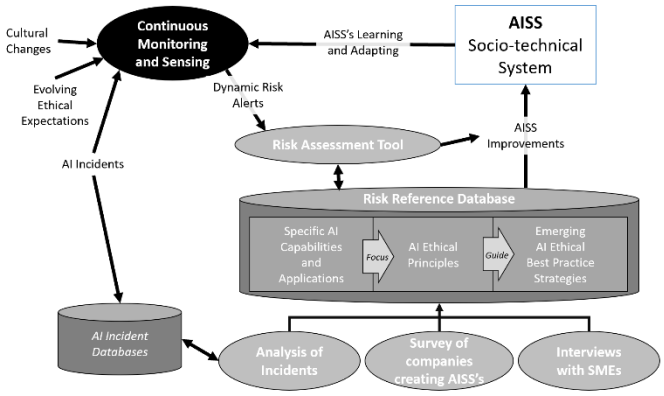

Fig. 2. The Dynamic e-ERM Engine

This continuous monitoring and sensing will trigger dynamic risk alerts to initiate a review and, if necessary, an in-depth assessment phase. Based on the assessment, an action phase could be executed to update the RRD with the latest context, associated risks, related AI Ethical Principles, and emerging best practices. It could also initiate changes to redesign and improve elements of the AISS application. These together form the Dynamic e-ERM engine.

4 Discussion and Further Work

The e-ERM framework is based on our experiences with AISS and the extant literature, but it needs to be validated and tested in practice. Because of the iterative nature of this research, a Design Science Research approach will be used. We envelope our project guided by a pilot study which found that decision support and machine learning are the AI capabilities receiving the greatest attention, with Management and the Human Resources (HR) / Talent department showing the most interest in leveraging AI. Further, an initial analysis of the available AI incident databases indicates that the acquisition (hiring) and termination (firing) steps of the HR process are prone to ethical failures, many related to AI decision support and machine learning. For instance, racial filtering of candidates via facial recognition or biased recruitment algorithms have become issues for many companies, as have automated performance assessments of staff by AI.

This research will therefore focus on organizations developing AISS for these parts of their HR processes.

To drive the dynamic nature of the e-ERM, the feasibility of three approaches to continuous monitoring and sensing will be tested. The first is a manually triggered model to link the latest AI incidents to AI Ethical Principles and to store these in the RRD. The second is applying AI machine learning to the publicly available information on incidents to regularly enrich the RRD. Finally, the use of a Generative Adversarial Networks (GAN) approach to actively track the ethical risk profile of the AISS as compared to the emerging best practices in the RRD will be evaluated. We propose the following stages for our future research directions:

- **Problem Space/Diagnosis Stage**: We have completed the first iteration of the DSR process through the creation of the e-ERM framework artifact. We will validate the various parts of the model, its relevance, and potential usefulness using expert interviews and focus groups of subject matter experts. We will use the results of the focus group to refine the model.
- **Solution Space/Design Stage:** We will then move to the solution space and enter the design stage via the following paths: (1) conduct interviews with around 20 IT Leaders to analyze successful strategies to link an organization's AI Ethical Principles to risk mitigating practices; (2) in order to incorporate the end user and indirect stakeholder's perspective, analyze the various existing AI incident databases to establish a relationship between incidents and the relevant AI Ethical Principles and the related risk impacts; and (3) iterate the artifact design with the subject matter experts. Evaluating the dynamic monitoring approaches and their design will also be completed.
- **Pilot Intervention/Implementation Stage:** The implementation/pilot stage will include deploying the elements of the e-ERM framework and validating its benefits for the risk management. This will be done in 2-3 organizations currently implementing AISS Applications for their HR processes, and will include risk managers, HR professionals, as well as the designers and developers of the AISS. As we have access to organizations in both the USA and India, we have an opportunity to study cultural differences that might impact the design and implementation of AISS across multiple cultures for HR.

5 Contributions and Limitations

This research adds to the current discussion on nascent ERM frameworks for complex, cognitive AI solutions using the acquisition and termination steps of the HR process as a basis. Unique contributions relate to the use of AI capabilities (machine learning and GANs) to support a dynamic e-ERM framework and its underlying RRD. The continuous monitoring and sensing of the changes within the AISS and its direct and indirect environments, as well as tracking AI incidents to trigger risk reviews is another contribution. The use of an RRD to link AI incidents to AI Ethical Principles and thereby provide risk-based, agile improvements to the AISS is a further contribution. We will support an ability to continuously analyze the AI incident databases to understand the causes of the incidents. We can then link these incidents to emerging best practices

elicited from successful organizations, allowing for effective continuous improvement of the AISS and proactively addressing potential ethical risks. Further contributions include the synthesis of the work by NIST, ISO, and others and defining the AI principles supporting responsible, pro-ethical AISS. Our envelopment approach supports an important set of HR applications for study.

In terms of limitations, as this is early work, there is much that still needs to be tested and validated. Also, future research will investigate an approach for the inclusion of input from those indirectly impacted by the e-ERM and the resulting AISS.

References

1. Wamba-Taguimdje, S.-L., Fosso Wamba, S., Kala Kamdjoug, J.R., Tchatchouang Wanko, C.E.: Influence of artificial intelligence (AI) on firm performance. Business Process Management Journal 26, 1893-1924 (2020)
2. Fuller, J.B., Raman, M., Sage-Gavin, E., Hines, K.: Hidden Workers: Untapped Talent. Harvard Business School Project on Managing the Future of Work and Accenture (2021)
3. Rushton, K.: Business ethics: a sustainable approach. Business Ethics: A European Review 11, 137-139 (2002)
4. Hillson, D.: Extending the risk process to manage opportunities. International Journal of Project Management 20, 235-240 (2002)
5. Nishimura, A.: Comprehensive opportunity and lost opportunity control model and enterprise risk management. International Journal of Business and Management 10, 73 (2015)
6. Figl, K., Recker, J.: Exploring cognitive style and task-specific preferences for process representations. Requirements Engineering 21, 63-85 (2016)
7. Mendling, J., Recker, J., Reijers, H.A., Leopold, H.: An Empirical Review of the Connection Between Model Viewer Characteristics and the Comprehension of Conceptual Process Models. Information Systems Frontiers 21, 1111-1135 (2019)
8. Seidel, S., Recker, J., von Brocke, J.: Sensemaking and Sustainable Practicing: Functional Affordances of Information Systems in Green TransformUnknown article. MIS Quarterly 37, 1257-1299 (2013)
9. Shneiderman, B.: Human-Centered AI. Oxford University Press (2022)
10. Steinberg, R.M., Everson, M.E.A., Martens, F.J., Nottingham, L.E.: Enterprise Risk Management - Integrated Framework: Executive Summary. COSO (2004)
11. Cormican, K.: Integrated Enterprise Risk Management: From Process to Best Practice. Modern Economy 05, 401-413 (2014)
12. Nocco, B.W., Stulz, R.M.: Enterprise Risk Management: Theory and Practice. Journal of Applied Corporate Finance 18, 8-20 (2006)
13. Brown University, https://www.brown.edu/academics/science-and-technology-studies/framework-making-ethical-decisions
14. Ewest, T.: The Challenges Within Ethical Leadership Theories. Prosocial Leadership, pp. 23-42. Palgrave Macmillan US, New York (2018)

15. Francis, R.D.: Ethical Risk Management. In: Farazmand, A. (ed.) Global Encyclopedia of Public Administration, Public Policy, and Governance, pp. 1-5. Springer International Publishing (2016)
16. Ruehle, C.R.: Investigating Market and Regulatory Forces Shaping Artificial Intelligence Adoptions. Muma Business Review 4, 177-192 (2020)
17. Sheth, A., Yip, H.Y., Iyengar, A., Tepper, P.: Cognitive Services and Intelligent Chatbots: Current Perspectives and Special Issue Introduction. IEEE Internet Computing 23, 6-12 (2019)
18. Marshall, T.E., Lambert, S.L.: Cloud-Based Intelligent Accounting Applications: Accounting Task Automation Using IBM Watson Cognitive Computing. Journal of Emerging Technologies in Accounting 15, 199-215 (2018)
19. Wirtz, B.W., Weyerer, J.C., Kehl, I.: Governance of artificial intelligence: A risk and guideline-based integrative framework. Government Information Quarterly 101685 (2022)
20. Floridi, L., Cowls, J.: A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review 1, (2019)
21. ISO: Risk Management - ISO 31000 Brochure. ISO, Geneva (2018)
22. Everson, M.E.A., Chesley, D.L., Martens, F.J., Bagin, M., Katz, H., Sylvis, K.T., Perraglia, S.J., Zelnik, K.C., Grimshaw, M.: Enterprise Risk Management: Integrating with Strategy and Performance - Executive Summary. COSO (2017)
23. NIST: AI Risk Management Framework Concept Paper. (2021)
24. Asatiani, A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T., Salovaara, A.: Sociotechnical Envelopment of Artificial Intelligence. Journal of the Association for Information Systems 22, 8 (2021)
25. Nagbøl, P.R., Müller, O., Krancher, O.: Designing a Risk Assessment Tool for Artificial Intelligence Systems. In: Chandra Kruse, L., Seidel, S., Hausvik, G.I. (eds.) The Next Wave of Sociotechnical Design. DESRIST., pp. 328-339. Springer (2021)
26. Robbins, S.: AI and the path to envelopment. AI & SOCIETY 35, 391-400 (2019)
27. Canca, C.: Operationalizing AI ethics principles. Communications of the ACM 63, 18-21 (2020)
28. NIST: NIST Draft - Taxonomy of AI Risk. NIST (2021)