AI Risk: A Systems Perspective

Dr. Gnana Bharathy

University of Technology Sydney and Australian Research Data Commons (ARDC)

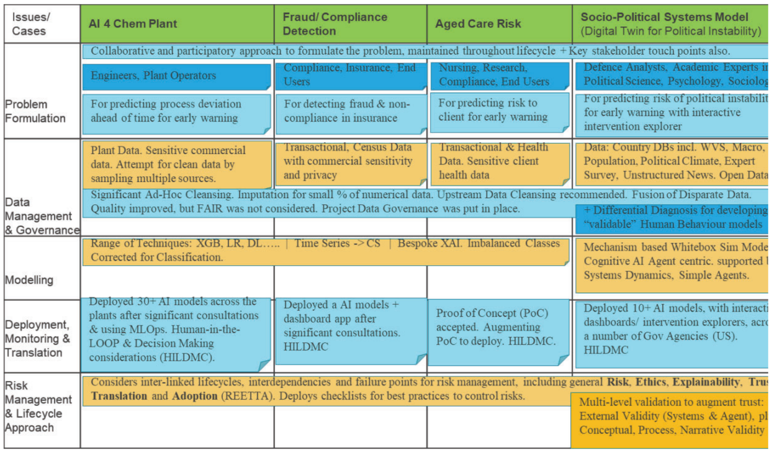
The widespread application of large-scale AI has undeniably brought about unprecedented convenience. However, it is essential to recognize that AI also presents a range of risks and challenges that must be addressed. To date, several countries have made progress in establishing principles and governance mechanisms for responsible AI. For instance, Australia has principles that encompass human and societal impact, human-centered values, reliability and safety, transparency and explainability, fairness, privacy, protection and security, contestability, and accountability. It is imperative to prioritize research on responsible AI, examining it from diverse perspectives and emphasizing the advancements made by various countries in implementing principles and governance mechanisms for responsible AI.

Systems thinking approaches, such as systems modelling (Guan et al., 2022), systems engineering (Devaney, 2016), and cybernetics (Morasso, 2023) offer valuable insights for managing AI risks. Guan et al. (2022) propose a system dynamics-based risk-factor model for ethical risks in AI decision-making, while Hohma et al. (2023) emphasize the importance of system characteristics such as balance, transparency, governance, and long-term orientation. The National Institute of Standards and Technology (NIST) in the US provides a taxonomy of characteristics to consider when dealing with AI risks (Golpayegani et al., 2022).

While these principles have been established as generic guidelines based on learning from other domains, there is a lack of studies applying these principles to actual AI development. In practice, managing AI risks from a systems thinking perspective involves a comprehensive approach, starting from problem formulation to a lifecycle approach, understanding and identifying interactions, hyper-stakeholder management, including participatory modelling, and anticipatory thinking.

In this study, we consider the risks and challenges associated with AI by examining four compelling case studies that the author had carried out in the past. These case studies shed light on the potential pitfalls and complexities that arise in the realm of AI, allowing us to gain a deeper understanding of the risks involved.

The first case involves the development of an AI



Issues/ Cases	AI 4 Chem Plant	Fraud/ Compliance Detection	Aged Care Risk	Socio-Political Systems Model (Digital Twin for Political Instability)
Problem Formulation	Collaborative and participatory approach to formulate the problem, maintained throughout lifecycle + Key stakeholder touch points also.			
	Engineers, Plant Operators	Compliance, Insurance, End Users	Nursing, Research, Compliance, End Users	Defence Analysts, Academic Experts i Political Science, Psychology, Sociolog
	For predicting process deviation ahead of time for early warning	For detecting fraud & non-compliance in insurance	For predicting risk to client for early warning	For predicting risk of political instabilit for early warning with interactive intervention explorer
Data Management & Governance	Plant Data. Sensitive commercial data. Attempt for clean data by sampling multiple sources.	Transactional, Census Data with commercial sensitivity and privacy	Transactional & Health Data. Sensitive client health data	Data: Country DBs incl. WVS, Macro, Population, Political Climate, Expert Survey, Unstructured News. Open Data
	Significant Ad-Hoc Cleansing. Imputation for small % of numerical data. Upstream Data Cleansing recommended. Fusion of Disparate Data. Quality improved, but FAIR was not considered. Project Data Governance was put in place.			+ Differential Diagnosis for developing "validable" Human Behaviour models
Modelling	Range of Techniques: XGB, LR, DL….. | Time Series -> CS | Bespoke XAI. Imbalanced Classes Corrected for Classification.			Mechanism based Whitebox Sim Mode Cognitive AI Agent centric. supported l Systems Dynamics, Simple Agents.
Deployment, Monitoring & Translation	Deployed 30+ AI models across the plants after significant consultations & using MLOps. Human-in-the-LOOP & Decision Making considerations (HILDMC).	Deployed a AI models + dashboard app after significant consultations. HILDMC	Proof of Concept (PoC) accepted. Augmenting PoC to deploy. HILDMC.	Deployed 10+ AI models, with interacti dashboards/ intervention explorers, acro a number of Gov Agencies (US). HILDMC
Risk Management & Lifecycle Approach	Considers inter-linked lifecycles, interdependencies and failure points for risk management, including general **Risk**, **Ethics**, **Explainability**, **Trus** **Translation** and **Adoption** (REETTA). Deploys checklists for best practices to control risks.			Multi-level validation to augment trust: External Validity (Systems & Agent), pl Conceptual, Process, Narrative Validity

Table **1**: Comparison across the Cases

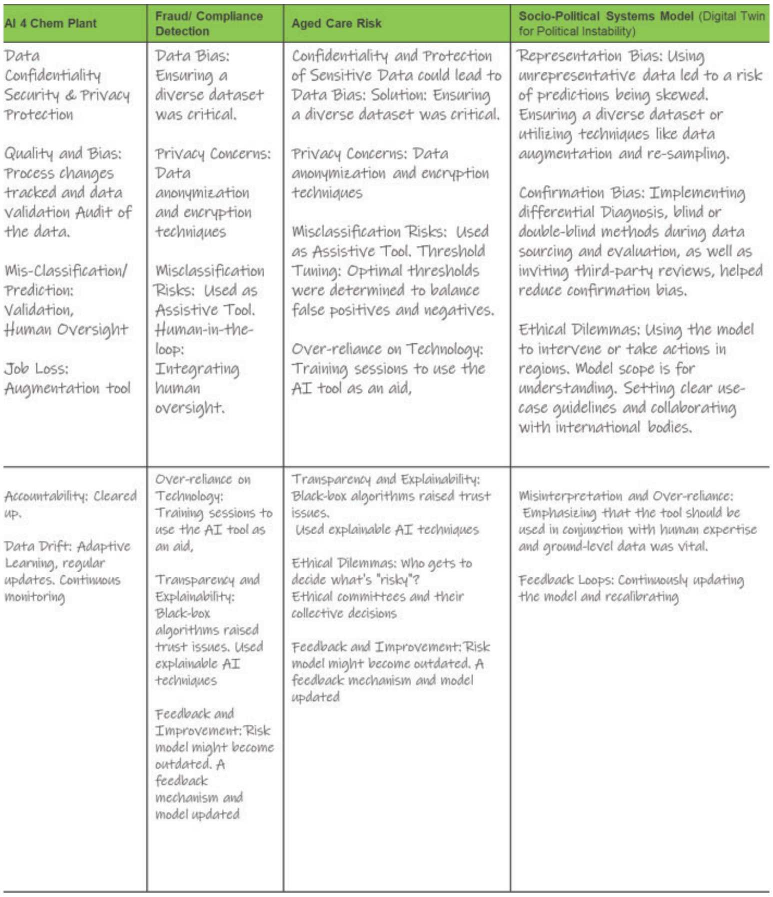
system for chemical process monitoring in various chemical plants. The system detects deviations in processes and provides advance warnings. The system has achieved an accuracy of around 88% in most cases, with a small percentage of cases having lower accuracy. The development and deployment of this system have been challenging.

In the second case, AI is used for insurance fraud detection. The system identifies non-compliance and provides recommendations, including nudging individuals towards regulatory actions. It analyzes data to identify anomalies and predict non-compliant behavior, providing risk scores and explanations for its recommendations. Currently, the system is only capable of initiating specific processes, such as communication and nudging, for low-risk cases.

The third case focuses on predicting the risk in elderly care. The AI system predicts the risk of harm for elderly individuals living alone or independently and provides recommendations to prevent harm. The performance of the model is satisfactory, but there are still privacy and security concerns that need to be addressed in subsequent iterations.

The fourth case involves a simulation model for predicting and exploring social and political conflicts. This model serves as a training system and provides broad recommendations for further research. It creates a virtual world where agents interact and make decisions based on socioeconomic and political conditions, shaping the trajectory of the virtual world.

The cases presented in Table 1 demonstrate the extensive range of applications for artificial intelligence (AI) and underscore the significance of collaboration with stakeholders during the development process. The acquisition and management of data are critical components of these AI systems, and it is essential to address concerns regarding data quality and privacy. It is emphasized that data correction and analysis are more important than the algorithmic aspect, highlighting the need for accurate and reliable data for AI models. To improve the quality of the data, measures such as data cleansing and



AI 4 Chem Plant	Fraud/ Compliance Detection	Aged Care Risk	Socio-Political Systems Model (Digital Twin for Political Instability)
Data Confidentiality Security & Privacy Protection Quality and Bias: Process changes tracked and data validation Audit of the data. Mis-Classification/ Prediction: Validation, Human Oversight Job Loss: Augmentation tool	Data Bias: Ensuring a diverse dataset was critical. Privacy Concerns: Data anonymization and encryption techniques Misclassification Risks: Used as Assistive Tool. Human-in-the-loop: Integrating human oversight.	Confidentiality and Protection of Sensitive Data could lead to Data Bias: Solution: Ensuring a diverse dataset was critical. Privacy Concerns: Data anonymization and encryption techniques Misclassification Risks: Used as Assistive Tool. Threshold Tuning: Optimal thresholds were determined to balance false positives and negatives. Over-reliance on Technology: Training sessions to use the AI tool as an aid,	Representation Bias: Using unrepresentative data led to a risk of predictions being skewed. Ensuring a diverse dataset or utilizing techniques like data augmentation and re-sampling. Confirmation Bias: Implementing differential Diagnosis, blind or double-blind methods during data sourcing and evaluation, as well as inviting third-party reviews, helped reduce confirmation bias. Ethical Dilemmas: Using the model to intervene or take actions in regions. Model scope is for understanding. Setting clear use-case guidelines and collaborating with international bodies.
Accountability: Cleared up. Data Drift: Adaptive Learning, regular updates. Continuous monitoring	Over-reliance on Technology: Training sessions to use the AI tool as an aid, Transparency and Explainability: Black-box algorithms raised trust issues. Used explainable AI techniques Feedback and Improvement: Risk model might become outdated. A feedback mechanism and model updated	Transparency and Explainability: Black-box algorithms raised trust issues. Used explainable AI techniques Ethical Dilemmas: Who gets to decide what's "risky"? Ethical committees and their collective decisions Feedback and Improvement: Risk model might become outdated. A feedback mechanism and model updated	Misinterpretation and Over-reliance: Emphasizing that the tool should be used in conjunction with human expertise and ground-level data was vital. Feedback Loops: Continuously updating the model and recalibrating

Table 2: Risks and Ethics in Different Contexts

preprocessing are necessary. The deployment of AI models involves human involvement and decision-making, indicating that these systems are not intended to replace human judgment but rather to assist and support decision-making processes. Human expertise is required to interpret the results and make informed decisions based on the AI recommendations. Managing risks is a crucial consideration in the AI systems lifecycle, including addressing concerns related to data privacy, bias, and potential ethical implications. To protect data privacy, measures such as anonymization and encryption techniques can be employed, and ensuring a diverse dataset can help mitigate bias in AI models.

Drilling down a little deeper, one can also look at the key issues in AI Risk. As one can see, data sensitivity is common for all cases, but is most important for Fraud Detection and Aged Care cases, owing to sensitive nature of the data handled. In the case of Chemical Plant case, it takes the form of business confidentiality. In all cases, however, appropriate security, privacy and data protection policies need to be implemented. However, in the case of Aged Care, this takes an interesting relationship with bias. For example, data privacy issues might compel minority aged care clients to decline data sharing. While protecting the individual's privacy, this decision would decrease the number of data points available for model training in the case of minority groups, resulting in a biased model.

In summary, the effective management of AI risk through a systems thinking approach requires the establishment of a comprehensive governance process (Felländer et al. 2022) prior to any activity, the dedication of sufficient resources towards problem formulation, the engagement and collaboration of all relevant stakeholders, the identification and modeling of potential risk factors, the adoption of proactive anticipatory thinking, consideration for deployment at the onset, and the integration of responsible AI principles.

It is particularly important to consider the responsible AI principle, which involves providing explanations to enhance trust and adoption, understanding the cyclic nature of subsystems, interactions and dependencies, paths to value, and points of risk and failure. The iterative nature of the development process is an advantage in this regard. Additionally, effective collaboration, data management, human oversight, and risk mitigation are critical factors in the AI systems lifecycle and are fundamental to the successful application of AI technologies across diverse domains.

References

1.Devaney, K. (2016). An integral approach to risk management. Incose International Symposium, 26(1), 892-908. https://doi.org/10.1002/j.2334-5837.2016.00200.x

2.Guan, H., Liye, D., & Zhao, A. (2022). Ethical risk factors and mechanisms in artificial intelligence decision making. Behavioral Sciences, 12(9), 343. https://doi.org/10.3390/bs12090343

3.Felländer, A., Rebane, J., Larsson, S., Wiggberg, M., & Heintz, F. (2022). Achieving a data-driven risk assessment methodology for ethical ai. Digital Society, 1(2). https://doi.org/10.1007/s44206-022-00016-0

4.Morasso, P. (2023). The quest for cognition in purposive action: from cybernetics to quantum computing. Journal of Integrative Neuroscience, 22(2), 39. https://doi.org/10.31083/j.jin2202039

5.Golpayegani, D., Pandit, H., & Lewis, D. (2022). Airo: an ontology for representing ai risks based on the proposed eu ai act and iso risk management standards.. https://doi.org/10.3233/ssw220008