Pattern-based AI Risk Assessment: A Taxonomy Expansion Use Case

Muhammad Ikhsan[1], Elmar Kiesling[1], Salma Mahmoud[2], Alexander Prock[1], Artem Revenko[2] and Fajar J. Ekaputra[1]

[1] *WU Wien, Austria*
[2] *Graphwise*

Abstract

As artificial intelligence (AI) is increasingly integrated into systems deployed in a wide range of application domains, the need to assess and mitigate the risks of these systems in diverse contexts has become a critical concern. Existing frameworks and methodologies for AI risk assessment support this process, but they often only provide general guidance disconnected from technical decisions. Furthermore, when an AI-based system is deployed in a new application context, they typically require a complete reassessment from scratch, which is a labour-intensive process that may miss potentially relevant risks. To tackle this challenge, this paper suggests a pattern-based approach to AI risk assessment that leverages semantic models of interlinked design and risk patterns to enable efficient and effective risk assessment across application contexts. We illustrate the effectiveness of our approach in a case study on a taxonomy expansion system in (i) a medical diagnosis application, and (ii) an e-commerce recommender application context and demonstrate how abstract risk patterns can support both architectural design decisions on the system level and structured risk assessments in a given application context. Our initial experiences suggest that the approach offers a promising and scalable method for assessing risks across application contexts.

Keywords

AI risk, risk assessment, risk patterns, design patterns

1. Introduction

Driven by vast expectations regarding potential business opportunities, Artificial Intelligence (AI) is increasingly integrated into a wide range of applications. At the same time, AI-related risk factors are increasingly becoming a major strategic concern for companies[1], driven by both the potentially severe consequences and increasing regulation – such as the European Union's AI Act – that affect both developers and users of AI system applications. Specifically, the AI Act mandates that developers of AI applications have to provide transparency and proper documentation whereas users are responsible for systematically managing risks associated with an AI application in their domain. This creates a gap between system design decisions and the inherent risks they entail on the one hand (i.e., the domain of the developer), and the consequences and impacts of these risks in a particular application context (i.e., the domain of the user).

We argue that a systematic, domain-specific risk assessment typically necessitates visibility into architectural choices, key design decisions and their respective risk implications, particularly as symbolic or subsymbolic AI methods are increasingly incorporated into complex systems composed of many components that introduce risks. This is, for instance, particularly relevant in the context of hybrid neuro-symbolic architectures where risks are difficult to trace and may compound. To tackle this challenge, we propose a structured, pattern-based risk assessment approach that bridges the gap

The Second Workshop on Knowledge Graphs and Neurosymbolic AI (KG-NeSy), co-located with SEMANTiCS'25: International Conference on Semantic Systems, September 3–5, 2025, Vienna, Austria

muhammad.ikhsan@wu.ac.at (M. Ikhsan); elmar.kiesling@wu.ac.at (E. Kiesling); salma.mahmoud@graphwise.ai (S. Mahmoud); alexander.prock@wu.ac.at (A. Prock); artem.revenko@graphwise.ai (A. Revenko); fajar.ekaputra@wu.ac.at (F. J. Ekaputra)
0009-0009-6718-9956 (M. Ikhsan); 0000-0002-7856-2113 (E. Kiesling); 0009-0008-4990-017X (S. Mahmoud); 0009-0006-0653-6481 (A. Prock); 0000-0001-6681-3328 (A. Revenko); 0000-0003-4569-2496 (F. J. Ekaputra)

[1] cf., e.g., a recent report https://autonomy.work/portfolio/their-capital-at-risk-the-rise-of-ai-as-a-threat-to-the-sp-500/

between – typically somewhat abstract – risk considerations in the AI system development process and the concrete risk assessment needs in a specific application context. Our approach is based on a semantic model of interconnected design and risk patterns that can be defined on an abstract level, incorporated as generic risks into system models, and instantiated in particular application contexts to provide an automatically generated framework for a concrete risk model. This reusable and modular approach provides a basis for the development of knowledge-based tool support in order to conduct thorough risk assessments efficiently and effectively.

We illustrate the risk assessment approach in the context of a system for automated taxonomy construction. Taxonomies are important tools that aid in knowledge management and information retrieval. Automating the task of building or expanding the taxonomies from corpora could improve efficiency by saving time and effort spent to extract all the terms within the corpus, linking the terms to broader entities and placing them correctly within a taxonomy. However, this process can introduce risks whose consequences and impact are heavily dependent on the application context. We explore the manifestation of these risks in (i) the medical fields, where errors in the system could lead to misdiagnosis and patient harm, and (ii) in e-commerce, where the impact of these errors affects the business and the profit.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work in AI system modeling and risk assessment, Section 3 introduces our pattern-based risk assessment approach, Section 4 demonstrates its use for two distinct use cases of a taxonomy expansion system, Section 5 concludes the paper with an outlook on future research.

2. Related Work

This section provides an overview of (i) AI systems modeling approaches and representations and (ii) approaches that address the increasing need for systematic AI risk governance, including governance frameworks, standard, guidelines, and taxonomies.

2.1. AI Systems Modeling

Modeling and representing AI systems is particularly important in hybrid/neurosymbolic AI, where symbolic and sub-symbolic components form complex architectures. Early work, including the boxology framework [1] and its extension [2], introduced visual notations for NeSy-AI design patterns, later adapted for LLM-based systems [3]. To formalize such boxology notation, Mossakowski [4] proposed a symbolic approach using the DOL meta-language, while Ellis et al. [5] introduced EASY-AI with semantic axioms and the SNOOP-AI tool [6]. Building on our earlier system-centric AI system representation [7], we recently developed Boxology Extended Annotation Model (BEAM) [8]. BEAM extends boxology with additional system and annotation elements, enabling structured representation of risks and mitigation strategies to support AI engineering processes.

2.2. AI Risk Management

Risk management frameworks such as the NIST AI RMF[2] [9] aim to support organizations in identifying, assessing, and managing risks associated with AI systems. **Standards** such as ISO/IEC 42001:2023 [10] and ISO/IEC 23894:2023 [11] define structured approaches for AI risk management; whereas the former covers establishing AI Management Systems within organizations, the latter provides more specific guidance on how organizations can manage risks related to AI. In a broader context, *tools* such as the *Assessment list for trustworthy artificial intelligence (ALTAI)*[3] or the Canadian Treasury Board's *Algorithmic Impact Assessment (AIA)*[4] use questionnaires to support organizations in

[2] https://doi.org/10.6028/NIST.AI.100-1

[3] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[4] (https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

assessing ethical considerations, risks and/or the fulfillment of requirements. Finally, there are several commercial tools that aim to support AI governance, transparency, and accountability – including IBM's *AI Factsheets*[5], which aims to track metadata across the model development life-cycle, and Google's *Model Cards* [12], which aim to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited.

Guidelines for Trustworthy AI include, for instance, the *Ethics Guidelines for Trustworthy AI*[6] developed by the EU High-Level Expert Group. There are also guidelines on the policy level, such as the *OECD AI principles*[7], aiming to guide countries in crafting policies to tackle AI risks.

AI Risk Taxonomies address the need for collecting and organizing AI risks; they include commercial and non-commercial initiatives such as the IBM AI Risk Atlas[8], MITRE Atlas[9] (from a security perspective), as well as academic initiatives such as the MIT AI Risk Repository [10] [13], a meta-repository that captures risks extracted from 65 existing frameworks and classifications of AI risks. Furthermore, more narrow taxonomies such as OWASP for LLMs [11] address particular sub-areas of AI. Finally **Risk Ontologies and Vocabularies** are most closely related to the idea of semantic risk modeling in this paper in that they define reusable concepts in a semantic model. Specifically, the AI Risk Ontology (AIRO) [12] [14] and the Vocabulary of AI Risks (VAIR) [13] [15] fall into this category. We reuse both in our pattern-based risk assessment framework introduced in section Section 3.

To conclude, there is a wealth of related work that can provide a basis for pattern-based risk assessment, but as of yet there is no structured approach for reusable semantic risk modeling based on interlinked design- and risk-patterns.

3. Pattern-based risk assessment

This section describes our pattern-based risk assessment method. As a guiding structure for our method, we developed a set of competency questions (CQs) that reflect key dimensions relevant to AI risk. These questions were derived through a synthesis of existing AI risk taxonomies and frameworks, combined with insights from modeling real-world AI systems across domains.

[CQ1] What risks are generally associated with given components or activities?

[CQ2] What are abstract consequences independent of system usage or application context?

[CQ3] What are specific risks in a given application context and how do they relate to technical design choices?

[CQ4] What are the possible impacts of risks on specific stakeholders?

[CQ5] Which strategies can be employed to mitigate the identified risks?

These questions address recurring modeling needs related to linking system design patterns with risks, consequences, stakeholder impacts, and mitigation strategies. Although the questions make certain assumptions, the structure has proven effective in enabling reusable and structured modeling.

[5] https://www.ibm.com/docs/en/software-hub/5.1.x?topic=services-ai-factsheets
[6] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[7] https://www.oecd.org/en/topics/sub-issues/ai-principles.html
[8] https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas
[9] https://atlas.mitre.org/
[10] https://airisk.mit.edu
[11] https://genai.owasp.org/llm-top-10-2023-24/
[12] https://delaramglp.github.io/airo/
[13] https://delaramglp.github.io/vair/

Method overview Figure 1 provides an overview of the resulting approach, which enables a modular, efficient, and thorough risk assessment by (i) assembling components from a *design pattern and abstract risk pattern library* into a larger *AI system model*, which based on the risk patterns linked to the components creates a *generic risk model* and (ii) given models of both the system and an *application context model*, deriving application-specific risks from the association of system design patterns to risk patterns. The result of this process is an *application-specific risk model* in the form of a knowledge graph. Elements of the risk model, i.e., *risks*, *consequences*, *impacts* and *mitigations*, are associated with the system or specific system components.

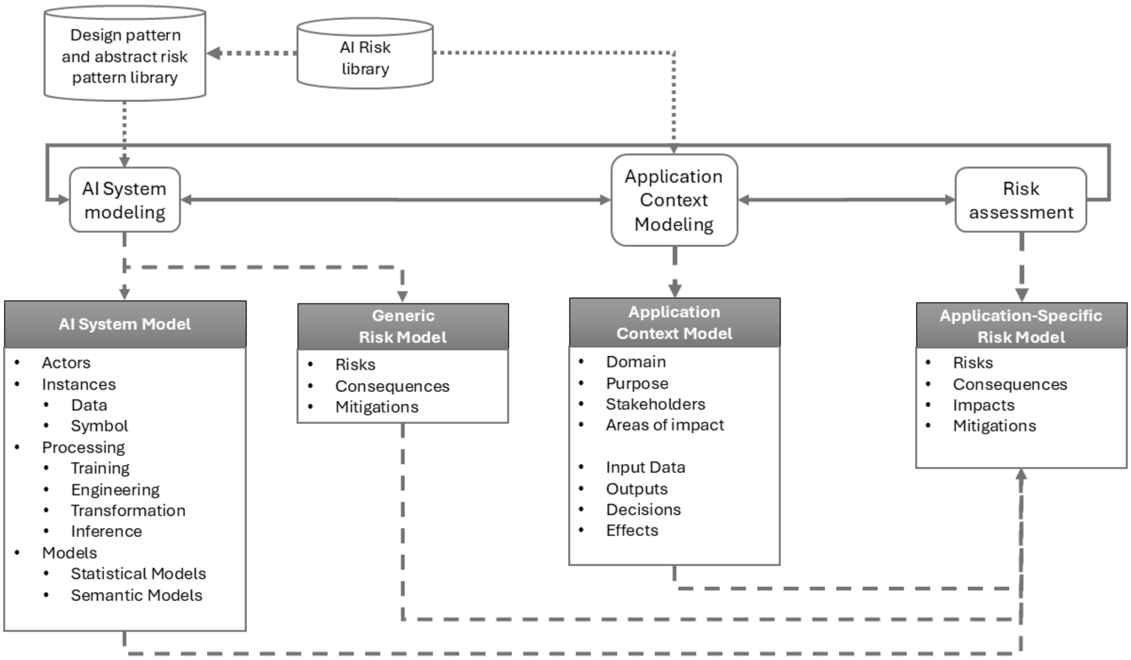

Figure 1: Pattern-based Risk Assessment Methodology - Overview

Conceptualization At the core of our method is a library of AI system design patterns with associated abstract risk patterns. These risk patterns are grounded in established AI risk taxonomies and frameworks, notably VAIR [15] and AIRO [14]. Figure 2 provides an example from the pattern library concerned with the usage of Large Language Models (LLMs) and the inherent associated risk pattern 'hallucination'.

To apply our risk assessment method for a specific system, the first step is to create a model of the AI system based on the Boxology Extended Annotation Model (BEAM) [8] notation, which provides both a visual notation and an ontology for the representation of AI systems. BEAM includes elements to represent system components, i.e. processes, models, inputs, outputs and actors, as well as the system workflow.

A generic (i.e. application-independent) risk model for the AI system is derived by identifying the design patterns that occur in the AI system model. For example, if the AI system model includes components that match the LLM prompting pattern depicted in Figure 2, the associated risk pattern concerned with hallucination will be included in the generic risk model. The generic risk model includes abstract risks, risk sources, consequences and mitigations. Elements of the risk model are connected to elements of the system model, i.e., a risk can be traced to the system component it originates from.

To perform a risk assessment of an AI system in a specific application, the context must be modeled first. The resulting application context model includes the application domain, the purpose of the system, the stakeholders of the system and areas of impact. Furthermore, it includes concrete inputs and outputs as well as decisions affected and expected effects. The concepts of the application context

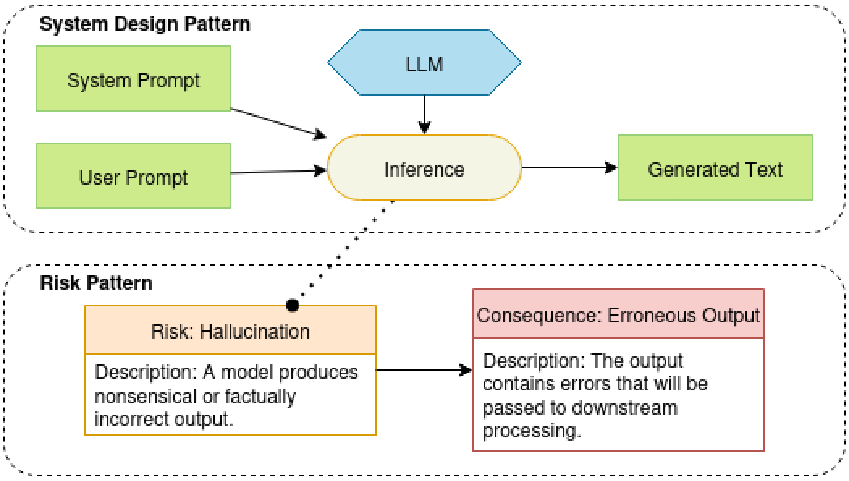
System Design Pattern
System Prompt
User Prompt
LLM
Inference
Generated Text
Risk Pattern
Risk: Hallucination
Description: A model produces nonsensical or factually incorrect output.
Consequence: Erroneous Output
Description: The output contains errors that will be passed to downstream processing.

Figure 2: In this example from the design pattern and abstract risk pattern library, an LLM is prompted to generate text. The risk of hallucination is associated to the LLM, with the consequence of erroneous output caused by the risk.

model are reused from VAIR [15], where applicable. Table 1 provides an example application context model.

In addition to specifying the details of the context model, connections between the data model and the system model are created, including connections between abstract input and output elements from the system model and application-specific inputs and outputs in the context model.

The final step is the application-specific risk assessment, in which the AI system model, the generic risk model and the application context models are combined to derive an application-specific risk model. During this process, generic risks, consequences and risk controls are extended and refined based on the application context. Furthermore, consequences are concretized and their impacts modeled and connected to the affected stakeholders.

4. Use Case: Taxonomy Assistant

The use case to demonstrate our methodology is *Corpus Analysis system*, a tool developed at Graphwise that automates the building and expansion of taxonomies. It facilitates knowledge management and information retrieval by processing large corpora, making it valuable across a variety of domains. The primary goal of this automation is to improve efficiency, but this introduces risks that require a thorough assessment to enable the selection and implementation of effective mitigation mechanisms.

The Corpus Analysis system incorporates a complex workflow with parallel processing, optional human-in-the-loop steps, and calls to external web services. To scope our analysis, we focus on a single critical step: invocation of an LLM. The LLM's task is to analyze a new term within its context and link it to an appropriate broader entity in an existing knowledge model. In this context, a 'broader entity' is defined as one that may serve as a 'hypernym' indicating a 'kind of' relationship (e.g., 'vehicle' is a hypernym of 'car'); a 'holonym' signifying a whole entity of which another word represents a part (e.g., 'car' is a holonym of 'engine'), or another form of conceptual inclusion, depending on the structure of the target hierarchy.

To investigate how the risks inherent in this task manifest, we apply our assessment methodology in two distinct use cases and demonstrate how general risks tied to the system's components can be identified and then specialized according to each application's unique context. Consequently, this highlights how the severity and nature of impacts and consequences are contingent on the specific domain of deployment.

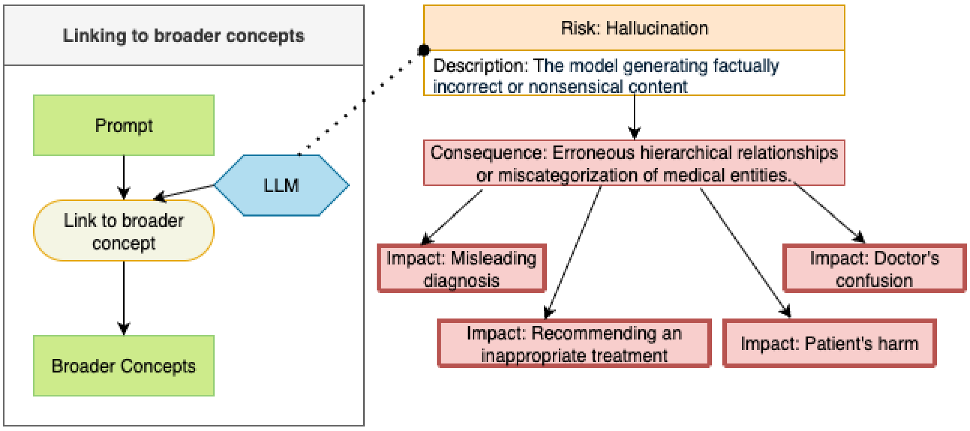
Linking to broader concepts
Prompt
LLM
Link to broader concept
Broader Concepts
Risk: Hallucination
Description: The model generating factually incorrect or nonsensical content
Consequence: Erroneous hierarchical relationships or miscategorization of medical entities.
Impact: Misleading diagnosis
Impact: Doctor's confusion
Impact: Recommending an inappropriate treatment
Impact: Patient's harm

Figure 3: Risk Assessment on one component of the system for AI-Assisted Orthopedic Diagnosis and Treatment Planning use case

4.1. AI-Assisted Orthopedic Diagnosis and Treatment Planning

In the medical field, having a taxonomy for musculoskeletal anatomy aids in medical analysis of X-ray images, guiding appropriate treatment selection. In this application context, an AI system that incorporates the automatically generated taxonomy can act as a critical link between image analysis and treatment recommendation, helping, e.g., in identifying a proper treatment for a broken bone.

Domain	Healthcare
Purpose	AssessingHealthRisk (VAIR [15])
Stakeholders	Hospital, Doctor, Patient
Area of Impact	Physical health
Input Data	Patient records, X-ray images
Output	Assessment result and clinical suggestions
Decision	Clinical referral or treatment decision based on system output
Effects	Support medical decision-making and care planning

Table 1
Example of an application context model for a medical AI system

The final system would process a patient's X-ray images (e.g. knee, wrist, spine) as *input*, perform *Image Analysis* using a machine learning-based AI component to analyze the X-ray and identify anomalies, potentially broken bones, fractures, or structural damage. It localizes the damage to specific anatomical regions using an automatically generated taxonomy of musculoskeletal anatomy, which is continuously updated and will be the focus of our illustrative excerpt.

An AI-based *treatment recommendation* component will then query the taxonomy based on the AI's analysis of the X-ray and the bone/injury identified in order to suggest potential diagnoses and appropriate treatment protocols. The whole process involves *human oversight* by an orthopedic surgeon or radiologist, who reviews the AI's findings, proposed diagnosis, and treatment recommendations, making the final decision.

One of the key risks associated with the creation of the taxonomy used in this application is the selection of broader concepts using LLMs, which comes with the potential risk of *hallucination* (cf. Figure 3 for an excerpt of the application-specific risk model). For instance, a 'lunate bone' might be mistakenly categorized under 'forearm bones' rather than 'carpal bones'. This miscategorization introduces erroneous hierarchical relationships within the taxonomy, directly impacting the system's users (i.e., Doctors) by potentially causing confusion and leading to misleading diagnostic pathways. Furthermore, since the AI system utilizes this taxonomy for treatment planning, an incorrect classification could recommend a forearm treatment for an injured lunate bone instead of a suitable treatment for carpal bones. The potential consequences for the patient are severe, including an increased risk of

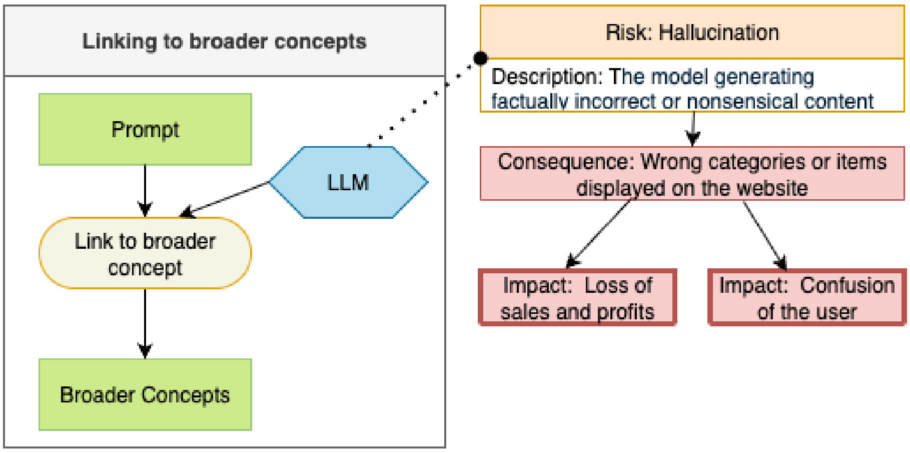
Linking to broader concepts
Prompt
LLM
Link to broader concept
Broader Concepts
Risk: Hallucination
Description: The model generating factually incorrect or nonsensical content
Consequence: Wrong categories or items displayed on the website
Impact: Loss of sales and profits
Impact: Confusion of the user

Figure 4: Risk Assessment on one component of the system for Targeted Recommendation in an E-commerce use case

complications, prolonged recovery, permanent damage, or unnecessary invasive procedures.

4.2. Targeted Recommendation in an E-commerce system

For the Targeted Recommendation in an E-commerce system, the use case is as follows. The machine learning-based AI system includes a taxonomy, which is expanded using LLM, of items sold by the website that are clustered by categories, subcategories, and attributes (size, color, etc.), and by using the user's preferences, filters, and interactions, the system could output the recommendations. The system's input would be the user's interactions with the items (search / view / buy), then the taxonomy would be used to categorize the items sold by the website in a hierarchical structure and display it on the website, then the AI system would use the taxonomy to recommend items to the user that they might want to check out to buy. However, when we assess one of the risks associated with the LLM component, *hallucination*, we find that this risk can result in a misleading categorization of items, as shown in Figure 4. The consequences of incorrect output generated by the LLM is having incorrect categories or items displayed on the website. Additionally, users' behavior would not reflect their true interests, leading to wrong recommendations. This would significantly impact the business and its users. For the business, there would be loss of sales and profit when the website is not easy to navigate, and incorrect items are displayed to the user making it difficult for the users to purchase what they need, and for the users, confusion would occur due to having the incorrect categories or items on the website or receiving irrelevant recommendations. Eventually, the user may abandon the website.

By showcasing two different use cases where our method can be used, we demonstrated how a systematic risk assessment method can be useful for use case owners. The use case owner would start by identifying the different components in the system and modeling them, then attach risks to each component from a pre-existing risk catalog, finally according to the specifications of the use case the product owner could determine the consequences and impacts of these risks on the stakeholders and the system, enabling the relevant teams to implement risk mitigation strategies.

5. Conclusions and Future Work

This paper introduced a pattern-based approach for AI risk assessment, which systematically identifies, describes, and graphically models potential harms. We demonstrated its efficacy through two distinct use cases, analyzing the application-specific risks of employing AI-generated taxonomies in different domains. Our findings confirm that the underlying semantic model provides an effective and efficient foundation for modular, thorough, and context-specific risk analysis. We contend that, with appropriate tooling, this method can facilitate the widespread adoption of risk-driven development – an approach that has become indispensable for navigating the increasing complexity and regulatory landscapes of modern AI systems.

In our future work, we aim to thoroughly evaluate which relevant risks can be adequately organized into reusable patterns and what limitations to reuse apply. Based on the findings, we will continue to develop and publish the abstract pattern libraries, incorporating knowledge from existing risk taxonomies and catalogs and generalizing them from a range of use cases in a large-scale national flagship project[14] on responsible AI. Furthermore, we aim to investigate opportunities to leverage semantics to reasoning about risks, implications and chains of causality, and risk propagation. This will provide a basis for the development of tools to enable the modular and risk-aware design of responsible AI systems.

Acknowledgments

This research was funded in part by the Austrian Science Fund (FWF) 10.55776/COE12 and the Austrian Research Promotion Agency (FFG) FAIR-AI project (Grant Nr. FO999904624).

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly, Writefull and Google Gemini in order to conduct Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] F. Van Harmelen, A. Ten Teije, A boxology of design patterns for hybrid learning and reasoning systems, Journal of Web Engineering 18 (2019) 97–123. URL: https://doi.org/10.13052/jwe1540-9589.18133.

[2] M. Van Bekkum, M. De Boer, F. Van Harmelen, A. Meyer-Vitali, A. T. Teije, Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases, Applied Intelligence 51 (2021-09) 6528–6546. URL: https://link.springer.com/10.1007/s10489-021-02394-3. doi:10.1007/s10489-021-02394-3.

[3] M. De Boer, Q. Smit, M. van Bekkum, A. Meyer-Vitali, T. Schmid, Modular design patterns for generative neuro-symbolic systems, in: Joint Proceedings of the ESWC 2024 Workshops and Tutorials co-located with 21th European Semantic Web Conference (ESWC 2024), 2024.

[4] T. Mossakowski, Modular design patterns for neural-symbolic integration: refinement and combination, in: Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 2nd International Joint Conference on Learning & Reasoning (IJCLR), 2022. URL: https://arxiv.org/abs/2206.04724.

[5] A. Ellis, B. Dave, H. Salehi, S. Ganapathy, C. Shimizu, Easy-ai: semantic and composable glyphs for representing ai systems, in: HHAI 2024: Hybrid Human AI Systems for the Social Good, IOS Press, 2024, pp. 105–113.

[6] A. Ellis, B. Dave, H. Salehi, S. Ganapathy, C. Shimizu, Implementing snoop-ai in comodide, in: NAECON 2024-IEEE National Aerospace and Electronics Conference, IEEE, 2024, pp. 101–104.

[7] F. J. Ekaputra, M. Llugiqi, M. Sabou, A. Ekelhart, H. Paulheim, A. Breit, A. Revenko, L. Waltersdorfer, K. E. Farfar, S. Auer, Describing and organizing semantic web and machine learning systems in the SWeMLS-KG, in: Proceedings of the 20th International Conference, ESWC 2023, Hersonissos, Crete, Greece, May 28–June 1, 2023, volume 13870 LNCS, 2023, pp. 372–389. URL: https://doi.org/10.1007/978-3-031-33455-9_22.

[14] https://fair-ai.at

[8] F. J. Ekaputra, A. Prock, E. Kiesling, Towards supporting ai system engineering with an extended boxology notation, in: The 2nd International Workshop on Knowledge Graphs for Responsible AI (KG-STAR) Co-located with the Extended Semantic Web Conference (ESWC 2025), CEUR-WS, 2025.

[9] National Institute of Standards and Technology, Artificial intelligence risk management framework (ai rmf 1.0) (2023). URL: https://doi.org/10.6028/NIST.AI.100-1.

[10] Iso/iec 42001:2023 information technology — artificial intelligence — management system, 2023. URL: https://www.iso.org/standard/42001.

[11] Iso/iec 23894:2023 information technology — artificial intelligence — guidance on risk management, 2023. URL: https://www.iso.org/standard/77304.html.

[12] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 220–229. URL: https://doi.org/10.1145/3287560.3287596. doi:10.1145/3287560.3287596.

[13] P. Slattery, A. K. Saeri, E. A. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, N. Thompson, The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence, arXiv preprint arXiv:2408.12622 (2024).

[14] D. Golpayegani, H. J. Pandit, D. Lewis, 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, 2022, pp. 51–65.

[15] D. Golpayegani, H. J. Pandit, D. Lewis, To be high-risk, or not to be—semantic specifications and implications of the ai act's high-risk ai applications and harmonised standards, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 905–915.