Risk Atlas Nexus: A System for Managing AI Risks

Inge Vejsbjerg[1], **Rahul Nair**[1], **Elizabeth M. Daly**[1], **Dhaval Salwala**[1], **Seshu Tirupathi**[1]

[1]IBM Research

ingevejs@ie.ibm.com, rahul.nair@ie.ibm.com, elizabeth.daly@ie.ibm.com, dhaval.vinodbhai.salwala@ibm.com, seshutir@ie.ibm.com

Abstract

We present Risk Atlas Nexus, an open source system for governing AI risks. The system unifies several risk classification frameworks through a common ontology. Given an AI application use case (called an *intent*), the system estimates risks and associated mitigations that are linked to identified risks. The tool is designed to be incorporated in AI governance workflows where recommendations can be translated to business controls to cover risks arising from AI use in firms.

Code — https://github.com/IBM/ai-atlas-nexus/

Demonstrator —
https://huggingface.co/spaces/ibm/risk-atlas-nexus

Introduction

Generative AI systems come with a broad range of risks that can cause reputational harm or lead to legal or financial penalities in practice. Examples of these risks include data privacy, biases, or disclosure of confidential information. These risks are not merely technical (i.e. involving data and the AI model), but also contextual (i.e. involving the manner in which AI model is deployed). Several risk taxonomies that aim to structure thinking around risks have been proposed in the literature. However there can be differences between taxonomies that can complicate analysis. On these accounts, holistic AI governance in practice is a challenge.

System Overview

Risk Atlas Nexus aims to operationalise risk identification and move beyond taxonomies of risk definitions towards providing actionable evidence for stakeholders. A key part of our system is to create and maintain relationships between existing risk classification frameworks. This allows for effective translation between the taxonomies and recommended mitigations.

The system and demonstrator are available in open source and comes with nine AI risk taxonomies including the NIST AI RMF (Tabassi 2023), IBM Risk Atlas (IBM 2023), MIT AI Risk Repository (Slattery et al. 2025) and more. Additionally, the system can ingest arbitrary taxonomies specified by users. The main components of the system are shown in Figure 1.

Risk identification Given a use case or intent, Risk Atlas Nexus identifies and prioritizes risks from a chosen taxonomy. It additionally associates standard AI tasks from the use case description. Risk identification is interactive with the first assessments made using LLMs and reviewed by users. Machine annotations are determined using chain-of-thought prompts and a map of similar risks from other taxonomies are returned from the knowledge graph.

In addition to risk taxonomies, the library consists of several other governance assets such as datasets, benchmarks, compliance questionnaires, and risk controls. Given the assessed risks and taxonomy items, these assets can be linked to the user intent being analysed.

Ontology and knowledge graph The framework is underpinned by a standards-based ontology structure to describe risks posed by AI systems and models with one coherent schema. This ontology allows for multidimensional relationships between entities sourced from different taxonomies and vocabularies, such as AI risks and risk control mechanisms. The AI systems knowledge graph provides structured mappings to link between these heterogenous governance resources. We have developed an ontology discussed in more detail (Bagehorn et al. 2025). It is encoded in LinkML (Linked Data Modeling Language) (Moxon et al. 2021) and makes use of the existing AIRO (Golpayegani, Pandit, and Lewis 2022) and DPV (Pandit et al. 2024) standards when conceptualising risks. Cross-taxonomy mapping is done using the SSSOM (Matentzoglu et al. 2022) standard, and the mappings are lifted into the knowledge graph.

There are many entities available in the ontology, however for the purposes of this demonstration, we will examine a few key ones. The first of these is the entity *Risk* which is largely based on the AIRO/DPV conception of risk, defined as the state of uncertainty associated with an AI system, that has the potential to cause harms. Risk records in the knowledge graph include attributes and slots including related risks, actions, AI tasks, etc. These attributes of the risk are available to be queried via an API provided by the python library.

A *Risk control* is a measure that maintains and/or modifies risk (and risk concepts), this can take the form of a manual

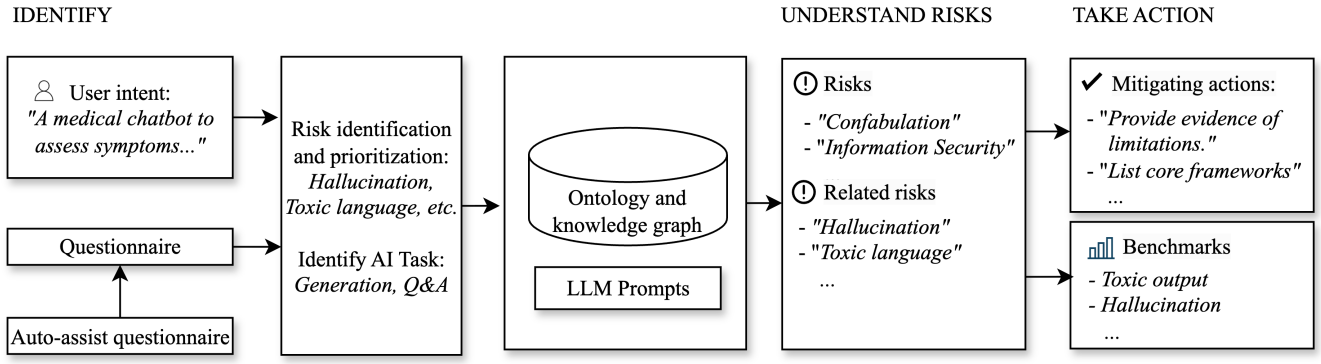

Figure 1: System flow from user intent to actionable risk mitigations

action, or a risk detector detecting risks, risk sources, consequences, and impacts.

A benchmark or *AI Evaluation* can be a metric, benchmark, card evaluation, a question or a combination of such entities.

Python library A Python library for risk management has been provided, which is geared at risk identification to mitigation and control. A notable feature of the framework is that is designed to be easily extensible for a customer or organisation's needs, facilitating a "bring your own risks" approach.

Online demonstration The demonstrator allows the user to interactively find out the potential risks from their generative AI use case. The user is invited to enter a text description of their use case and choose a judge-LLM to evaluate it. A drop-down containing AI risk taxonomies available in the Risk Atlas Nexus affords the user with the ability to switch between taxonomies. Each taxonomy has coverage of different sets of risks, which may be more applicable to their use case. A sample use case a user might enter could be:

"In a medical chatbot, create a triage system that assesses patients' symptoms and provides advice based on their medical history and current condition. The chatbot will identify potential medical issues, and offer recommendations to the patient or healthcare provider."

Visualising the results When the use case has been processed by the system, a list of potential risks is displayed to the user. The user views the potential risks, including "Data Privacy", "Information Integrity", and more. They see the risk "Confabulation", and would like to learn more. On selection of the risk, the user is shown the risk definition drawn from the source taxonomy, any mapped related risks from other taxonomies, related risk controls, and any related AI benchmarks. A small network diagram shows a simplified sub-graph for the selected risk, detailing the mitigations, evaluations and related risks available, as illustrated in Figure 2. The user can use this information to understand the risk "Confabulation", or any of the other listed potential AI risks that may be associated with their use case, and communicate mitigation strategies for this risk with their stakeholders. The user may also download the output of this interaction into a JSON file so that it can be used for discussion with stakeholders, or given to developers so that they can build further processes around the risks which have been uncovered.

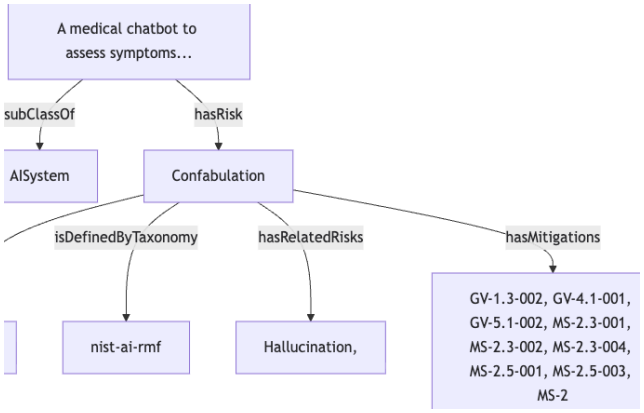

Figure 2: Network of NIST risk "Confabulation"

User feedback The online demonstrator application is deployed to Hugging Face and has had over 5,700 visits so far. Initial informal user feedback so far has been positive. Exploratory use of the library as part of an integration with one of our company's flagship products has provided feedback of strong customer interest and appetite for further integration of the library.

Discussion and Conclusion

We provide an interactive exploration of risks and mitigation with the Risk Atlas Nexus framework. The Risk Atlas Nexus framework and the underlying knowledge graph allows users to weigh the risks and opportunities of their approach, promoting transparency and responsible AI usage. We aim to grow the project to serve as a comprehensive toolkit for AI risk governance, combining taxonomies, automated tooling, and community resources to help organizations manage AI-related risks more effectively.

References

Bagehorn, F.; Brimijoin, K.; Daly, E. M.; He, J.; Hind, M.; Garces-Erice, L.; Giblin, C.; Giurgiu, I.; Martino, J.; Nair, R.; Piorkowski, D.; Rawat, A.; Richards, J.; Rooney, S.; Salwala, D.; Tirupathi, S.; Urbanetz, P.; Varshney, K. R.; Vejsbjerg, I.; and Wolf-Bauwens, M. L. 2025. AI Risk Atlas: Taxonomy and Tooling for Navigating AI Risks and Resources. arXiv:2503.05780.

Golpayegani, D.; Pandit, H.; and Lewis, D. 2022. AIRO: An Ontology for Representing AI Risks Based on the Proposed EU AI Act and ISO Risk Management Standards. In *International Conference on Semantic Systems*.

IBM. 2023. AI Risk Atlas. https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas.

Matentzoglu, N.; Balhoff, J. P.; Bello, S. M.; Bizon, C.; Brush, M.; Callahan, T. J.; Chute, C. G.; Duncan, W. D.; Evelo, C. T.; Gabriel, D.; Graybeal, J.; Gray, A.; Gyori, B. M.; Haendel, M.; Harmse, H.; Harris, N. L.; Harrow, I.; Hegde, H. B.; Hoyt, A. L.; Hoyt, C. T.; Jiao, D.; Jiménez-Ruiz, E.; Jupp, S.; Kim, H.; Koehler, S.; Liener, T.; Long, Q.; Malone, J.; McLaughlin, J. A.; McMurry, J. A.; Moxon, S.; Munoz-Torres, M. C.; Osumi-Sutherland, D.; Overton, J. A.; Peters, B.; Putman, T.; Queralt-Rosinach, N.; Shefchek, K.; Solbrig, H.; Thessen, A.; Tudorache, T.; Vasilevsky, N.; Wagner, A. H.; and Mungall, C. J. 2022. A Simple Standard for Sharing Ontological Mappings (SSSOM). *Database*, 2022. Baac035.

Moxon, S. A.; Solbrig, H.; Unni, D. R.; Jiao, D.; Bruskiewich, R. M.; Balhoff, J. P.; Vaidya, G.; Duncan, W. D.; Hegde, H.; Miller, M.; et al. 2021. The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics. *ICBO*, 3073: 148–151.

Pandit, H. J.; Esteves, B.; Krog, G. P.; Ryan, P.; Golpayegani, D.; and Flake, J. 2024. Data Privacy Vocabulary (DPV) – Version 2. arXiv:2404.13426.

Slattery, P.; Saeri, A. K.; Grundy, E. A. C.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; and Thompson, N. 2025. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. arXiv:2408.12622.

Tabassi, E. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD.