AI Risk Atlas: Taxonomy and Tooling for Navigating AI Risks and Resources

Frank Bagehorn, Kristina Brimijoin, Elizabeth M. Daly, Jessica He, Michael Hind, Luis Garcés-Erice, Christopher Giblin, Ioana Giurgiu, Jacquelyn Martino, Rahul Nair, David Piorkowski, Ambrish Rawat, John Richards, Sean Rooney, Dhaval Salwala, Seshu Tirupathi, Peter Urbanetz, Kush R. Varshney, Inge Vejsbjerg, Mira L. Wolf-Bauwens

July 10, 2025

Abstract

The rapid evolution of generative AI has expanded the breadth of risks associated with AI systems. While various taxonomies and frameworks exist to classify these risks, the lack of interoperability between them creates challenges for researchers, practitioners, and policymakers seeking to operationalise AI governance. To address this gap, we introduce the **AI Risk Atlas**, a structured taxonomy that consolidates AI risks from diverse sources and aligns them with governance frameworks. Additionally, we present the **Risk Atlas Nexus**, a collection of open-source tools designed to bridge the divide between risk definitions, benchmarks, datasets, and mitigation strategies. This knowledge-driven approach leverages ontologies and knowledge graphs to facilitate risk identification, prioritization, and mitigation. By integrating AI-assisted compliance workflows and automation strategies, our framework lowers the barrier to responsible AI adoption. We invite the broader research and open-source community to contribute to this evolving initiative, fostering cross-domain collaboration and ensuring AI governance keeps pace with technological advancements.

`https://github.com/IBM/risk-atlas-nexus`

1 Introduction

Identifying the risks of AI systems has attracted interest from research [46, 36], industry [38], and policy makers [12]. These perspectives spawned many innovations to aid in the creation and operationalising of responsible AI system design [6, 2, 47, 7]. Generative AI and its rapidly evolving capabilities [11], increases the spectrum of risks, and the urgency to mitigate them. More recently, advent of agentic AI introduces new risks by giving AI automony and the ability to execute actions. As a result, there is a need to help the community identify and address these risks in parallel, while also creating a mechanism for collaboration.

There have been several efforts to catalogue risks associated with AI systems [31, 32, 29, 40, 48]. However, connections and relationships to existing risk classification frameworks are missing. This lack of connectivity can present a challenge for practitioners who may want to adopt new risk taxonomies, but have already categorised their assets using existing definitions.

Meanwhile, there are currently flourishing communities for sharing datasets [26] and benchmarks [16]. Although the focus of datasets effort is to evaluate the correctness of an AI model, typically partitioned by AI task or skill, many of these datasets can also be leveraged to assess various AI risks. Similarly, mapping the results of benchmarks to risk concerns is currently not a part of most benchmarks, but could be accomplished with better collaboration between the benchmark and risk communities [41].

To help manage risks, the process of putting a new AI system into production often includes multiple stakeholders such as business owners, risk and compliance officers, and ethics officers approving the AI

[0] Authors are listed in alphabetical order by last name.

system for a specific usage. Governance frameworks to manage this process typically include multiple manual steps, including curating information needed to assess risks (where will the system be used? who is the target user?) and reviewing outcomes to identify appropriate actions and governance strategies [35]. Automation can play a vital role in easing the barrier for entry for developers and practitioners to make responsible AI an integral part of their process, however, this automation requires some structure of the underlying information. For example, AI capabilities can help to create better semi-structured governance information, identify and prioritize risks according to the intended use case, recommend appropriate benchmarks and risk assessments and most importantly recommend mitigation strategies and recommended actions. Our aim is to develop an AI Systems risk ontology that links risk entities described using multiple different taxonomies with AI models, evaluations, mitigations and other important entities. This ontology is manifested in a knowledge graph that associated tooling uses to integrate distinct risk frameworks, thereby providing the community with a way to align their assets with both new and existing risk definitions.

This paper is organized as follows. Section 2 describes the AI Risk Atlas, which provides a taxonomy of AI Risks. Section 3 presents **Risk Atlas Nexus**, an open source tooling effort we have developed to enable inter- and cross-community collaboration. Section 4 discusses some of the future directions that can be facilitated by these tools. Section 5 invites the broader community to expand on this initial seeding of tools by bringing different perspectives with different needs and ultimately lowering the barrier to AI governance.

2 The AI Risk Landscape

This section describes the *AI Risk Atlas* which aims to provide clarity for practitioners of the risks associated with Generative AI systems.

The AI Risk Atlas is a taxonomy of AI risks collected from prior research, real-world examples, and from experts in the field. It defines risks posed by AI systems and explain potential consequences of those risks. Each risk is grouped into one of five categories based on where the risk originates. The categories are input risks, inference risks, output risks and non-technical risks. Within each category, risks are further grouped into risk dimensions such as accuracy, fairness, or explainability. These dimensions classify the individual risks into groups, and enable a user of the Atlas to focus on the dimensions relevant to them.

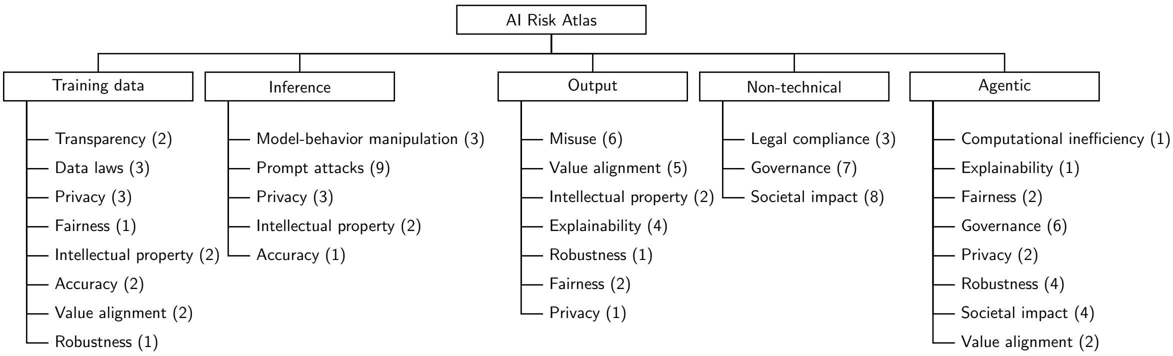
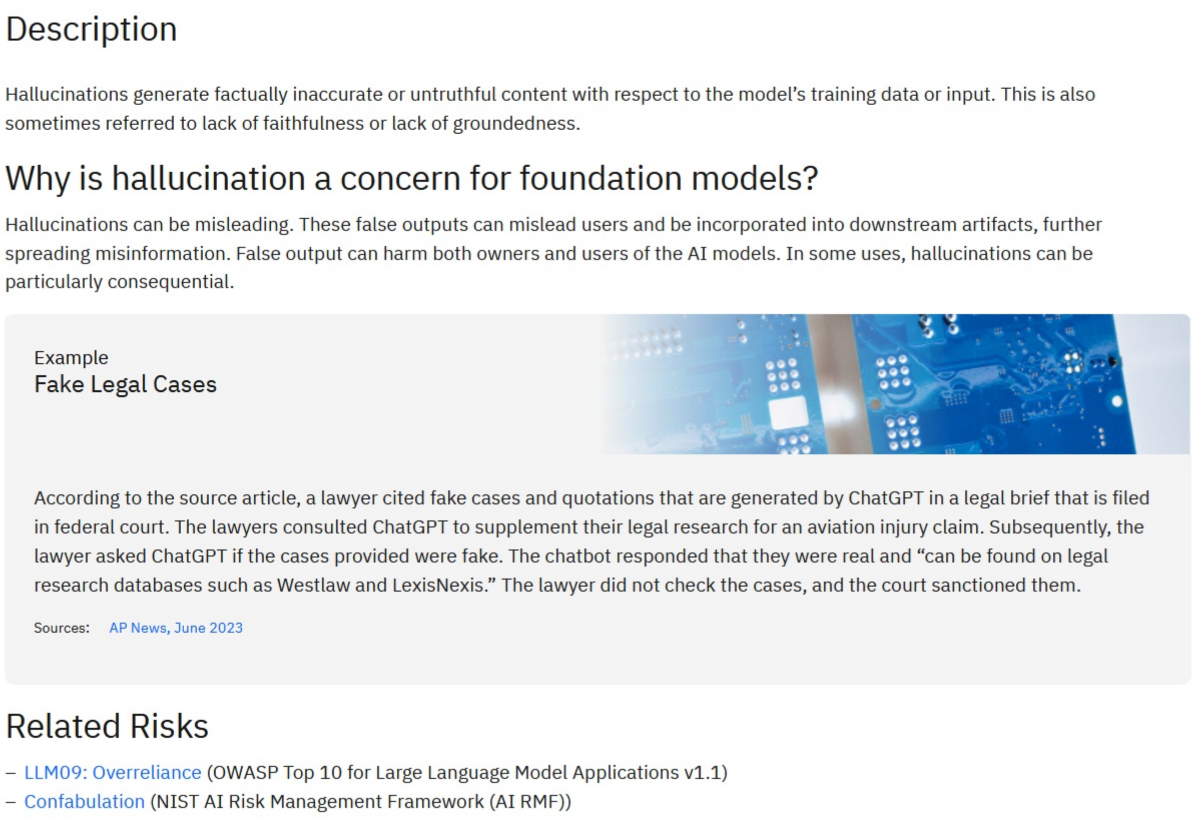


Figure 1: AI Risk Atlas Taxonomy. The taxonomy is divided into five categories, which are then further divided into dimensions. Next to each dimension in the parentheses is the number of risks identified for that dimension.

Figure 1 shows the overall AI Risk Atlas taxonomy where each category may include a subset of risk definitions. For example, Figure 2 shows a screenshot of the Hallucination risk from the AI Risk Atlas, which is in the Robustness dimension in the Output category in Figure 1. The details contain a description of the risk along with a description of why the risk is a concern, a public example, when available, of the risk being manifested, and any related risks in other popular taxonomies.

The creation of the AI Risk Atlas was motivated by the changing risk landscape due to the emergence and rapid success of generative AI. Before generative AI became ubiquitous, IBM Research had developed the techniques to evaluate and mitigate some risks such as fairness, explainability, adversarial robustness, privacy, and uncertainty for traditional (non-generative) models. This work evolved into open-source toolkits designed to measure, and in some cases, mitigate those risks [6, 1, 22, 23, 17]. However, these



Description

Hallucinations generate factually inaccurate or untruthful content with respect to the model's training data or input. This is also sometimes referred to lack of faithfulness or lack of groundedness.

Why is hallucination a concern for foundation models?

Hallucinations can be misleading. These false outputs can mislead users and be incorporated into downstream artifacts, further spreading misinformation. False output can harm both owners and users of the AI models. In some uses, hallucinations can be particularly consequential.

Example

Fake Legal Cases

According to the source article, a lawyer cited fake cases and quotations that are generated by ChatGPT in a legal brief that is filed in federal court. The lawyers consulted ChatGPT to supplement their legal research for an aviation injury claim. Subsequently, the lawyer asked ChatGPT if the cases provided were fake. The chatbot responded that they were real and "can be found on legal research databases such as Westlaw and LexisNexis." The lawyer did not check the cases, and the court sanctioned them.

Sources: AP News, June 2023

Related Risks

– LLM09: Overreliance (OWASP Top 10 for Large Language Model Applications v1.1)
– Confabulation (NIST AI Risk Management Framework (AI RMF))

Figure 2: Screenshot of AI Risk Atlas Detail Page for Hallucination

toolkits did not initially account for the risks specific to generative models. The large amount of data used to train generative AI, the increasing complexity of the models, and the non-determinism prevalent were some of the factors that led to the identification of new risks for generative models. Prompt-based attacks, poorly curated training data, and hallucinations emerged as early risks identified by researchers and concerned citizens alike. The identification of risks had begun to outpace mechanisms for measuring and understanding the risks.

To address this gap and to provide a foundation for understanding the risks of both traditional and generative AI models, IBM's AI Ethics Board created white papers [8, 9] that provided the foundation for the AI Risk Atlas. The goal for the Atlas was multi-faceted. First, we wanted to have a single source of information for currently known AI risks. This would enable a shared vocabulary when discussing AI risks. Second, we wanted to identify opportunities for measuring and mitigating the newly identified risks. This would help identify research opportunities for underrepresented risks. Third, we wanted a resource to develop usage-based governance. Specifically, we wanted to address the question of what risks are relevant to particular use cases.

The AI Risk Atlas has been used as a conversation starting point with enterprises who are considering deploying AI. It helps these organizations to be aware of the possible risks they need to govern. It provides a palette or vocabulary of risks that enterprises can consider: they can decide the risk is relevant to their use case and develop a plan to mitigate the risk either with tooling, human oversight, or both. For those risks that aren't applicable for a use case, the organization can document this risk to help demonstrate their risk governance framework. However, the AI Risk Atlas can be used further than just a conversation starter. It can provide the underlying vocabulary for the complete management of the risk from development to deployment to monitoring [21, 34, 14].

3 Tools for Practitioners

The IBM AI Risk Atlas has been used many enterprise customers to help them reason about the risks in their AI systems. In order to enable efforts to leverage these risks to operationalise governance and risk mitigation frameworks we created *Risk Atlas Nexus*.

The *Risk Atlas Nexus* is a collection of tooling to help bring together disparate resources related to governance of foundation models. We aim to support a community-driven approach to curating and cataloguing resources such as datasets, benchmarks and mitigations. Our goal is to turn abstract risk definitions into actionable workflows that streamline AI governance processes. By connecting fragmented resources, Risk Atlas Nexus seeks to fill a critical gap in AI governance, enabling stakeholders to build more robust, transparent, and accountable systems. The Risk Atlas Nexus is a step towards enabling the following.

Navigating disparate risk taxonomies: IBM AI Risk Atlas is one amongst a number of existing risk taxonomies, for example; the OWASP Top 10 for LLMs and Generative AI Apps [32], the NIST AI Risk Management Framework [31], the MIT AI Risk Repository [29, 40], the AIR taxonomy 2024 [48].

To provide a way through this labyrinth of taxonomies, we have constructed an AI risk ontology that allows both the creation of a knowledge graph containing those different taxonomies and the ability to map between them, leveraging the AIRO [18].The ontology has been modeled using LinkML [30], which allows the generation of different data representations (e.g. RDF, OWL) in a simple way. The risk taxonomies have been stored as LinkML data instance YAML files. To express some semantically meaningful mapping between risks from different taxonomies, we have used the Simple Standard for Sharing Ontological Mappings (SSSOM) [28]. Therefore those mappings are maintained in SSOM TSV files and are converted to LinkML data YAML using Python helper scripts.

Sample notebooks demonstrate how to load the LinkML data and user data and how to get details about specific risks and their relations to risks in other taxonomies.

Question Answering: Compliance questionnaires are usually required prior to deploying an AI model into production. These enable a thorough understanding of the specific use case and associated risk exposures [21, 25]. The Risk Atlas Nexus supports the development and curation of questionnaires to a desired taxonomy. Additionally, the content can support Large Language Models (LLMs) to assist users in responding to time-consuming compliance questionnaires, thereby reducing manual effort and minimizing errors [14]. Similarly, other aspects like risk identification, guardrail implementation, and identifying security vulnerabilities for specific use cases can be largely automated with human feedback and sign-off provided only when necessary.

Use Case to Risk Prioritisation: Risk Atlas Nexus supports risk classificated according to the EU AI Act [20]. Additionally, to help prioritise which of the many risks are most related to their use case we leverage LLM-as-a-judge capabilities to identify which risks to consider. This information can be used to look for appropriate research papers, benchmarks and metrics. In a similar manner Risk Atlas Nexus can be used to tag disparate resources by passing in text such as a paper abstract or a dataset description as the risks for the basis for an LLM-as-a-judge definition [3, 15].

From Risks to Mitigating Actions: The knowledge graph supports mapping between risks to two types of mitigation strategies: detectors and recommended actions. Detectors such as Granite Guardian [33] dimensions could be run in tandem with an AI system to better protect against certain risks such as social bias and prompt injection attacks. We have also mined recommended actions as part of the NIST AI Risk Management Framework [31] to be able to recommend more process driven mitigation strategies. Additionally, Risk Atlas Nexus supports linkages between benchmarks and evaluations associated with uncovering risks.

Bring Your Own Risks, Relationships and Questionnaires: Risk Atlas Nexus tooling supports several well known risk taxonomy frameworks, however some organisations may wish to define their own custom concerns and definitions. Risk Atlas Nexus allows users to define custom questionnaire templates as well as taxonomies, risks, mappings, and mitigation actions which should conform to the ontology schema. We encourage users to contribute their taxonomy definitions and mappings back to the project for others to use through the open-source project.

GAF-Guard: Given the evolving capabilities, agentic frameworks[24, 13, 4, 5] have the potential to create real-time governance pipelines, from identifying relevant risks and benchmarks to identifying mitigating actions to online monitoring capabilities. GAF-Guard is an agentic framework that leverages Risk Atlas Nexus functionalities like auto assisted question answering, use-case to risk identification and prioritization. GAF-Guard agents perform these tasks and consequently automate the process of the governance life-cycle. GAF-Guard agents initially identify the risks associated with a given use-case and then generate risk monitors and detectors for near real-time monitoring of deployed LLMs for the given use-case.

4 Potential for the future

There is immense potential of automation of various aspects of compliance and risk management processes. While human oversight and manual verification are still essential requirements for compliance, auditing and regulatory purposes, automation can help to bring an efficient execution of intermediate stages in the compliance workflow. AutoML strategies have been employed to automate model training pipelines excelling at tasks such as feature selection, hyper parameter optimization, model generation and evaluation [19, 43]. In a similar manner AI governance pipelines could be employed to detect and mitigate risks. By starting with identifying the most relevant risks, running the most relevant benchmarks and then assessing the impact of employing real-time mitigation strategies, concrete recommendations can be made to improve safety of AI solutions.

In the context of complex systems LLMs are increasingly being used as part of the validation process, from software testing and assisting in tasks such as test case preparation [44] to assessing the output of an LLM [42, 39, 3]. With the increasing capabilities of LLMs their applications have gone beyond single function tasks to being used to address complex problems acting as autonomous-agents [45]. ToolLLaMA learns how to call appropriate API based tools [37] meaning LLMs can orchestrate tasks that leverage existing functionality embedded in other tooling. The research community has begun to use agent flows to design, plan and execute scientific experiments [10] and even write papers [27].

5 Conclusion and call to action

Curating the AI Risk Atlas is just the first step in providing a reference framework for researchers and practitioners navigating the rapidly evolving AI landscape. By positioning our risk taxonomy in relation to existing definitions and taxonomies, we aim to encourage the community to map new risk definitions, datasets, benchmarks, research papers, and crucially mitigation and detection strategies into a structured framework. This approach will enhance accessibility and facilitate the operationalisation of AI governance processes.

The initial tools released as part of the Risk Atlas Nexus toolkit represent only the beginning of what is possible. We are committed to ongoing development, enabling the developer community to contribute and expand this initiative. By fostering a community-driven approach, we can lower the barrier to entry for all. We invite the open-source community to enrich the knowledge graph by linking benchmarks, datasets, and research papers to identified risks. Additionally, contributors can request new functionality through GitHub enhancement requests or develop and integrate their own algorithms.

References

[1] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. http://arxiv.org/abs/1909.03012, 2019.

[2] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. AI explainability 360 toolkit. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, pages 376–379, 2021.

[3] Zahra Ashktorab, Michael Desmond, Qian Pan, James M Johnson, Martin Santillan Cooper, Elizabeth M Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. Aligning human and llm judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy preferences. *arXiv preprint arXiv:2410.00873*, 2024.

[4] Autogen. https://github.com/microsoft/autogen.

[5] Bee AI framework. https://github.com/i-am-bee/beeai-framework.

[6] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019.

[7] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[8] IBM AI Ethics Board. Foundation models: Opportunities, risks and mitigations. https://www.ibm.com/downloads/documents/us-en/10a99803d8afd656, 2024.

[9] IBM AI Ethics Board. Expanding on ethical considerations of foundation models. https://www.ibm.com/think/insights/expanding-on-ethical-considerations-of-foundation-models, 2025.

[10] Daniil A Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.

[11] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[12] Gonçalo Carriço. The eu and artificial intelligence: A human-centred perspective. *European View*, 17(1):29–36, 2018.

[13] crewAI. https://github.com/crewAIInc/crewAI.

[14] Elizabeth M Daly, Sean Rooney, Seshu Tirupathi, Luis Garces-Erice, Inge Vejsbjerg, Frank Bagehorn, Dhaval Salwala, Christopher Giblin, Mira L Wolf-Bauwens, Ioana Giurgiu, et al. Usage governance advisor: from intent to ai governance. *arXiv preprint arXiv:2412.01957*, 2024.

[15] Michael Desmond, Zahra Ashktorab, Werner Geyer, Elizabeth Daly, Martin Santillan Cooper, Qian Pan, Rahul Nair, Nico Wagner, and Tejaswini Pedapati. Evalassist: Llm-as-a-judge simplified. In *AAAI Conference on Artificial Intelligence*, 2025.

[16] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024.

[17] Soumya Ghosh, Q. Vera Liao, Karthikeyan Natesan Ramamurthy, Jirí Navrátil, Prasanna Sattigeri, Kush R. Varshney, and Yunfeng Zhang. Uncertainty quantification 360: A holistic toolkit for quantifying and communicating the uncertainty of AI. https://arxiv.org/abs/2106.01410, 2021.

[18] Delaram Golpayegani, Harshvardhan J Pandit, and Dave 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*, pages 51–65. IOS Press, 2022.

[19] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-based systems*, 212:106622, 2021.

[20] Viviane Herdel, Sanja Šćepanović, Edyta Bogucka, and Daniele Quercia. Exploregen: Large language models for envisioning the uses and risks of ai technologies. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 584–596, 2024.

[21] IBM. IBM unveils watsonx.governance 2.1 to transform your AI governance experience. https://www.ibm.com/new/announcements/ibm-unveils-watsonx-governance-2-1-to-transform-your-ai-governance-experience.

[22] IBM Research. Adversarial Robustness 360. https://art360.res.ibm.com, 2022. Accessed: 2022-08-29.

[23] IBM Research. AI Privacy 360. https://aip360.res.ibm.com, 2022. Accessed: 2022-08-29.

[24] LangGraph. https://langchain-ai.github.io/langgraph.

[25] Sung Une Lee, Harsha Perera, Boming Xia, Yue Liu, Qinghua Lu, Liming Zhu, Olivier Salvado, and Jon Whittle. Qb4aira: A question bank for ai risk assessment, 2023.

[26] Quentin Lhoest, Albert Villanova Del Moral, Yacine Jernite, Abhishek Thakur, Patrick Von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*, 2021.

[27] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[28] Nicolas Matentzoglu, James P Balhoff, Susan M Bello, Chris Bizon, Matthew Brush, Tiffany J Callahan, Christopher G Chute, William D Duncan, Chris T Evelo, Davera Gabriel, John Graybeal, Alasdair Gray, Benjamin M Gyori, Melissa Haendel, Henriette Harmse, Nomi L Harris, Ian Harrow, Harshad B Hegde, Amelia L Hoyt, Charles T Hoyt, Dazhi Jiao, Ernesto Jiménez-Ruiz, Simon Jupp, Hyeongsik Kim, Sebastian Koehler, Thomas Liener, Qinqin Long, James Malone, James A McLaughlin, Julie A McMurry, Sierra Moxon, Monica C Munoz-Torres, David Osumi-Sutherland, James A Overton, Bjoern Peters, Tim Putman, Núria Queralt-Rosinach, Kent Shefchek, Harold Solbrig, Anne Thessen, Tania Tudorache, Nicole Vasilevsky, Alex H Wagner, and Christopher J Mungall. A Simple Standard for Sharing Ontological Mappings (SSSOM). *Database*, 2022, 05 2022. baac035.

[29] MIT. MIT AI risk repository. https://airisk.mit.edu/.

[30] Sierra AT Moxon, Harold Solbrig, Deepak R Unni, Dazhi Jiao, Richard M Bruskiewich, James P Balhoff, Gaurav Vaidya, William D Duncan, Harshad Hegde, Mark Miller, et al. The linked data modeling language (linkml): A general-purpose data modeling framework grounded in machine-readable semantics. *ICBO*, 3073:148–151, 2021.

[31] NIST. AI risk management framework. `https://www.nist.gov/itl/ai-risk-management-framework`, 2023.

[32] OWASP. Owasp top 10 for llms and generative ai apps. `https://genai.owasp.org/llm-top-10/`, 2024.

[33] Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, Giulio Zizzo, Muhammad Zaid Hameed, Mark Purcell, Michael Desmond, Qian Pan, Zahra Ashktorab, Inge Vejsbjerg, Elizabeth M. Daly, Michael Hind, Werner Geyer, Ambrish Rawat, Kush R. Varshney, and Prasanna Sattigeri. Granite guardian, 2024.

[34] David Piorkowski, Michael Hind, and John Richards. Quantitative ai risk assessments: Opportunities and challenges, 2024.

[35] David Piorkowski, Michael Hind, John Richards, and Jacquelyn Martino. Developing a risk identification framework for foundation model uses, 2025.

[36] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

[37] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

[38] Kyarash Shahriari and Mana Shahriari. Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, pages 197–201. IEEE, 2017.

[39] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14, 2024.

[40] Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv preprint arXiv:2408.12622*, 2024.

[41] Anna Sokol, Elizabeth Daly, Michael Hind, David Piorkowski, Xiangliang Zhang, Nuno Moniz, and Nitesh Chawla. Benchmarkcards: Standardized documentation for large language model benchmarks, 2025.

[42] Tempest A van Schaik and Brittany Pugh. A field guide to automatic evaluation of llm-generated summaries. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2832–2836, 2024.

[43] Dakuo Wang, Parikshit Ram, Daniel Karl I Weidele, Sijia Liu, Michael Muller, Justin D Weisz, Abel Valente, Arunima Chaudhary, Dustin Torres, Horst Samulowitz, et al. Autoai: Automating the end-to-end ai lifecycle with humans-in-the-loop. In *Companion Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 77–78, 2020.

[44] Junjie Wang, Yuchao Huang, Chunyang Chen, Zhe Liu, Song Wang, and Qing Wang. Software testing with large language models: Survey, landscape, and vision. *IEEE Transactions on Software Engineering*, 2024.

[45] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[46] Joseph Weizenbaum. Computer power and human reason: From judgment to calculation. *San Francisco*, 1976.

[47] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

[48] Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Ai risk categorization decoded (air 2024): From government regulations to corporate policies. *arXiv preprint arXiv:2406.17864*, 2024.

A AI Risk Atlas Definitions

The below is a catalog of potential risks when working with generative AI, foundation models, and machine learning models.

Evasion attack

Evasion attacks attempt to make a model output incorrect results by slightly perturbing the input data sent to the trained model.

Concern: Evasion attacks alter model behavior, usually to benefit the attacker.

Type: inference
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-evasion-attack
Tag: evasion-attack
URI: IBM AI Risk Atlas - Evasion attack

Impact on the environment

AI, and large generative models in particular, might produce increased carbon emissions and increase water usage for their training and operation.

Concern: Training and operating large AI models, building data centers, and manufacturing specialized hardware for AI can consume large amounts of water and energy, which contributes to carbon emissions. Additionally, water resources that are used for cooling AI data center servers can no longer be allocated for other necessary uses. If not managed, these could exacerbate climate change.

Type: non-technical
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-impact-on-the-environment
Tag: impact-on-the-environment
URI: IBM AI Risk Atlas - Impact on the environment

Incorrect risk testing

A metric selected to measure or track a risk is incorrectly selected, incompletely measuring the risk, or measuring the wrong risk for the given context.

Concern: If the metrics do not measure the risk as intended, then the understanding of that risk will be incorrect and mitigations might not be applied. If the model's output is consequential, this might result in societal, reputational, or financial harm.

Type: non-technical
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-incorrect-risk-testing
Tag: incorrect-risk-testing
URI: IBM AI Risk Atlas - Incorrect risk testing

Over- or under-reliance

In AI-assisted decision-making tasks, reliance measures how much a person trusts (and potentially acts on) a model's output. Over-reliance occurs when a person puts too much trust in a model, accepting a model's output when the model's output is likely incorrect. Under-reliance is the opposite, where the person doesn't trust the model but should.

Concern: In tasks where humans make choices based on AI-based suggestions, over/under reliance can lead to poor decision making because of the misplaced trust in the AI system, with negative consequences that increase with the importance of the decision.

Type: output
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-over-or-under-reliance
Tag: over-or-under-reliance
URI: IBM AI Risk Atlas - Over- or under-reliance

Membership inference attack

A membership inference attack repeatedly queries a model to determine if a given input was part of the model's training. More specifically, given a trained model and a data sample, an attacker appropriately samples the input space, observing outputs to deduce whether that sample was part of the model's training.

Concern: Identifying whether a data sample was used for training data can reveal what data was used to train a model, possibly giving competitors insight into how a model was trained and the opportunity to replicate the model or tamper with it. Models that include publicly-available data are at higher risk of such attacks.

Type: inference
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-membership-inference-attack
Tag: membership-inference-attack
URI: IBM AI Risk Atlas - Membership inference attack

Confidential data in prompt

Confidential information might be included as a part of the prompt that is sent to the model.

Concern: If not properly developed to secure confidential data, the model might reveal confidential information or IP in the generated output. Additionally, end users' confidential information might be unintentionally collected and stored.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-confidential-data-in-prompt
Tag: confidential-data-in-prompt
URI: IBM AI Risk Atlas - Confidential data in prompt

Prompt leaking

A prompt leak attack attempts to extract a model's system prompt (also known as the system message).

Concern: A successful prompt leaking attack copies the system prompt used in the model. Depending on the content of that prompt, the attacker might gain access to valuable information, such as sensitive personal information or intellectual property, and might be able to replicate some of the functionality of the model.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-prompt-leaking
Tag: prompt-leaking
URI: IBM AI Risk Atlas - Prompt leaking

Data privacy rights alignment

Existing laws could include providing data subject rights such as opt-out, right to access, and right to be forgotten.

Concern: Improper usage or a request for data removal could force organizations to re-train the model, which is expensive.

Type: training-data
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-data-privacy-rights
Tag: data-privacy-rights
URI: IBM AI Risk Atlas - Data privacy rights alignment

Discriminatory actions

AI agents can take actions where one group of humans is unfairly advantaged over another due to the decisions of the model. This may be caused by AI agents' biased actions that impact the world, in the resources consulted, and in the resource selection process. For example, an AI agent can generate code that can be biased.

Concern: Discriminatory actions can cause harm to people. Discriminatory actions taken by an AI agent could perpetuate bias to systems outside the AI agent owner's control, impact people, or lead to unintended consequences.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-discriminatory-actions
Tag: discriminatory-actions
URI: IBM AI Risk Atlas - Discriminatory actions

IP information in prompt

Copyrighted information or other intellectual property might be included as a part of the prompt that is sent to the model.

Concern: Inclusion of such data might result in it being disclosed in the model output. In addition to accidental disclosure, prompt data might be used for other purposes like model evaluation and retraining, and might appear in their output if not properly removed.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-ip-information-in-prompt
Tag: ip-information-in-prompt
URI: IBM AI Risk Atlas - IP information in prompt

Legal accountability

Determining who is responsible for an AI model is challenging without good documentation and governance processes.

Concern: If ownership for development of the model is uncertain, regulators and others might have concerns about the model. It would not be clear who would be liable and responsible for the problems with it or can answer questions about it. Users of models without clear ownership might find challenges with compliance with future AI regulation.

Type: non-technical
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-legal-accountability
Tag: legal-accountability
URI: IBM AI Risk Atlas - Legal accountability

Hallucination

Hallucinations generate factually inaccurate or untruthful content with respect to the model's training data or input. This is also sometimes referred to lack of faithfulness or lack of groundedness.

Concern: Hallucinations can be misleading. These false outputs can mislead users and be incorporated into downstream artifacts, further spreading misinformation. False output can harm both owners and users of the AI models. In some uses, hallucinations can be particularly consequential.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-hallucination
Tag: hallucination
URI: IBM AI Risk Atlas - Hallucination

Social hacking attack

Manipulative prompts that use social engineering techniques, such as role-playing or hypothetical scenarios, to persuade the model into generating harmful content.

Concern: Social hacking attacks can be used to alter model behavior and benefit the attacker. The content it generates may cause harms for the user or others.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-social-hacking-attack
Tag: social-hacking-attack
URI: IBM AI Risk Atlas - Social hacking attack

Harmful output

A model might generate language that leads to physical harm. The language might include overtly violent, covertly dangerous, or otherwise indirectly unsafe statements.

Concern: A model generating harmful output can cause immediate physical harm or create prejudices that might lead to future harm.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-harmful-output
Tag: harmful-output
URI: IBM AI Risk Atlas - Harmful output

Indirect instructions attack

Prompts, questions, or requests designed to elicit undesirable responses from the application. Unlike direct instructions attacks, the model is instructed to use instructions that are embedded in external data like a website.

Concern: Indirect instructions attacks can be used to alter model behavior and benefit the attacker. The content it generates may cause harms for the user or others.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-indirect-instructions-attack
Tag: indirect-instructions-attack
URI: IBM AI Risk Atlas - Indirect instructions attack

Mitigation and maintenance

The large number of components and dependencies that agent systems have complicates keeping them up to date and correcting problems.

Concern: AI agents may interact with other systems, tools, or other agents. Tracing the root cause of failure becomes more difficult and more costly as agent capabilities and complexities increase.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-mitigation-maintenance
Tag: mitigation-maintenance
URI: IBM AI Risk Atlas - Mitigation and maintenance

AI agent compliance

Determining AI agents' compliance is complex and there might not be enough information to assess whether the agentic AI system is compliant with applicable legal requirements.

Concern: AI agents may interact with other systems, tools, or other agents. AI agents can also find solutions to accomplish a task or a goal in a variety of ways and there could be uncertainty around the way an AI agent would choose each time to perform the task. Assessing compliance can become more difficult as agent capabilities increase.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-ai-agent-compliance
Tag: ai-agent-compliance
URI: IBM AI Risk Atlas - AI agent compliance

Function calling hallucination

AI agents might make mistakes when generating function calls (calls to tools to execute actions). Those function calls might result in incorrect, unnecessary or harmful actions. Examples: Generating wrong functions or wrong parameters for the functions.

Concern: Hallucinations when generating function calls might result in wrong or redundant actions being performed. Depending on the actions taken, AI agents can cause harms to owners and users of the AI agents.

Type: agentic
Descriptor: specific to agentic AI

Implementation details:
ID: atlas-function-calling-hallucination
Tag: function-calling-hallucination
URI: IBM AI Risk Atlas - Function calling hallucination

Confidential information in data

Confidential information might be included as part of the data that is used to train or tune the model.

Concern: If confidential data is not properly protected, there could be an unwanted disclosure of confidential information. The model might expose confidential information in the generated output or to unauthorized users.

Type: training-data
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-confidential-information-in-data
Tag: confidential-information-in-data
URI: IBM AI Risk Atlas - Confidential information in data

Lack of model transparency

Lack of model transparency is due to insufficient documentation of the model design, development, and evaluation process and the absence of insights into the inner workings of the model.

Concern: Transparency is important for legal compliance, AI ethics, and guiding appropriate use of models. Missing information might make it more difficult to evaluate risks, change the model, or reuse it. Knowledge about who built a model can also be an important factor in deciding whether to trust it. Additionally, transparency regarding how the model's risks were determined, evaluated, and mitigated also play a role in determining model risks, identifying model suitability, and governing model usage.

Type: non-technical
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-lack-of-model-transparency
Tag: lack-of-model-transparency
URI: IBM AI Risk Atlas - Lack of model transparency

Exploit trust mismatch

Attackers might initiate injection attacks to bypass the trust boundary, which is a distinct point or conceptual line where the level of trust in a system, application or network changes. Background execution in multi-agent environments increases the risk of covert channels if input/output validation is weak.

Concern: This could lead to mismatched (expected vs. realized) trust boundaries and could result in unintended tool use, excessive agency, and privilege escalation.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-exploit-trust-mismatch
Tag: exploit-trust-mismatch
URI: IBM AI Risk Atlas - Exploit trust mismatch

Unrepresentative data

Unrepresentative data occurs when the training or fine-tuning data is not sufficiently representative of the underlying population or does not measure the phenomenon of interest.

Concern: If the data is not representative, then the model will not work as intended.

Type: training-data
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-unrepresentative-data
Tag: unrepresentative-data
URI: IBM AI Risk Atlas - Unrepresentative data

AI agents’ impact on human agency

The autonomous nature of AI agents in performing tasks or taking actions could affect the individuals’ ability to engage in critical thinking, make choices and act independently.

Concern: AI agents might shift the decision, thinking, and control from humans to machines. This might negatively impact the society and human welfare as they limit the freedom and meaningful participations of humans in performing a task or making decisions.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-impact-human-agency
Tag: impact-human-agency
URI: IBM AI Risk Atlas - AI agents’ impact on human agency

Personal information in prompt

Personal information or sensitive personal information that is included as a part of a prompt that is sent to the model.

Concern: If personal information or sensitive personal information is included in the prompt, it might be unintentionally disclosed in the models’ output. In addition to accidental disclosure, prompt data might be stored or later used for other purposes like model evaluation and retraining, and might appear in their output if not properly removed.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-personal-information-in-prompt
Tag: personal-information-in-prompt
URI: IBM AI Risk Atlas - Personal information in prompt

AI agents’ Impact on human agency

The autonomous nature of AI agents in performing tasks or taking actions might affect the individuals’ ability to engage in critical thinking, make choices, and acting independently.

Concern: AI agents might shift the decision, thinking, and control from humans to machines. This might negatively impact society and human welfare as they limit the freedom and meaningful participations of humans in performing a task or making decisions.

Type: non-technical
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-impact-on-human-agency
Tag: impact-on-human-agency
URI: IBM AI Risk Atlas - AI agents’ Impact on human agency

Sharing IP/PI/confidential information with user

AI agents with unrestricted access to resources or databases or tools could potentially store and share PI/IP/confidential information with system users when performing their actions.

Concern: AI agents may share privileged information to users. The act of sharing the information may result in harm for the model owner, user, or others. The harm can vary based on the type and details of the information shared. Without adequate oversight, these privacy incidents might overwhelm company resources.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-sharing-info-user
Tag: sharing-info-user
URI: IBM AI Risk Atlas - Sharing IP/PI/confidential information with user

Lack of testing diversity

AI model risks are socio-technical, so their testing needs input from a broad set of disciplines and diverse testing practices.

Concern: Without diversity and the relevant experience, an organization might not correctly or completely identify and test for AI risks.

Type: non-technical
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-lack-of-testing-diversity
Tag: lack-of-testing-diversity
URI: IBM AI Risk Atlas - Lack of testing diversity

Nonconsensual use

Generative AI models might be intentionally used to imitate people through deepfakes by using video, images, audio, or other modalities without their consent.

Concern: Deepfakes can spread disinformation about a person, possibly resulting in a negative impact on the person's reputation. A model that has this potential must be properly governed.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-nonconsensual-use
Tag: nonconsensual-use
URI: IBM AI Risk Atlas - Nonconsensual use

Decision bias

Decision bias occurs when one group is unfairly advantaged over another due to decisions of the model. This might be caused by biases in the data and also amplified as a result of the model's training.

Concern: Bias can harm persons affected by the decisions of the model.

Type: output
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-decision-bias
Tag: decision-bias
URI: IBM AI Risk Atlas - Decision bias

Exposing personal information

When personal identifiable information (PII) or sensitive personal information (SPI) are used in training data, fine-tuning data, or as part of the prompt, models might reveal that data in the generated output. Revealing personal information is a type of data leakage.

Concern: Sharing people's PI impacts their rights and make them more vulnerable.

Type: output
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-exposing-personal-information
Tag: exposing-personal-information
URI: IBM AI Risk Atlas - Exposing personal information

AI agents' impact on jobs

Widespread adoption of AI agents to perform complex tasks might lead to widespread automation of roles and could lead to job displacement.

Concern: As trust in agentic systems increases, business may be more motivated to use agents instead of people. Job displacement might lead to a loss of income and thus might negatively impact society and human welfare. Re-skilling may be challenging given the pace of the technology evolution.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-impact-jobs
Tag: impact-jobs
URI: IBM AI Risk Atlas - AI agents' impact on jobs

Improper data curation

Improper collection and preparation of training or tuning data includes data label errors and by using data with conflicting information or misinformation.

Concern: Improper data curation can adversely affect how a model is trained, resulting in a model that does not behave in accordance with the intended values. Correcting problems after the model is trained and deployed might be insufficient for guaranteeing proper behavior.

Type: training-data
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-data-curation
Tag: data-curation
URI: IBM AI Risk Atlas - Improper data curation

Over- or under-reliance on AI agents

Reliance, that is the willingness to accept an AI agent behavior, depends on how much a user trusts that agent and what they are using it for. Over-reliance occurs when a user puts too much trust in an AI agent, accepting an AI agent's behavior even when it is likely undesired. Under-reliance is the opposite, where the user doesn't trust the AI agent but should. Increasing autonomy (to take action, select and consult resources/tools) of AI agents and the possibility of opaqueness and open-endedness increase the variability and visibility of agent behavior leading to difficulty in calibrating trust and possibly contributing to both over- and under-reliance.

Concern: Over/under reliance can lead to poor decision making by humans because of their misplaced trust in the AI agent, with negative consequences that escalate with the significance of the decision.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-over-or-under-reliance-on-ai-agents
Tag: over-or-under-reliance-on-ai-agents
URI: IBM AI Risk Atlas - Over- or under-reliance on AI agents

Attack on AI agents' external resources

Attackers intentionally create vulnerabilities or exploit existing vulnerabilities in external resources (tools/database/applications/services/other agents) that AI agents rely on to execute their intended actions or to achieve their goals.

Concern: Compromised external resources could impact the AI agent's performance in different ways, such as manipulating AI agents to pursue a different goal, manipulating AI agents to execute undesired actions, capturing and relaying interactions between AI agents to malicious actors, and getting AI agents to share personal or confidential information.

Type: agentic
Descriptor: specific to agentic AI

Implementation details:
ID: atlas-external-resources-attack
Tag: external-resources-attack
URI: IBM AI Risk Atlas - Attack on AI agents' external resources

Revealing confidential information

When confidential information is used in training data, fine-tuning data, or as part of the prompt, models might reveal that data in the generated output. Revealing confidential information is a type of data leakage.

Concern: If not properly developed to secure confidential data, the model might reveal confidential information or IP in the generated output and reveal information that was meant to be secret.

Type: output
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-revealing-confidential-information
Tag: revealing-confidential-information
URI: IBM AI Risk Atlas - Revealing confidential information

Spreading disinformation

Generative AI models might be used to intentionally create misleading or false information to deceive or influence a targeted audience.

Concern: Spreading disinformation might affect human's ability to make informed decisions. A model that has this potential must be properly governed.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-spreading-disinformation
Tag: spreading-disinformation
URI: IBM AI Risk Atlas - Spreading disinformation

Uncertain data provenance

Data provenance refers to tracing history of data, which includes its ownership, origin, and transformations. Without standardized and established methods for verifying where the data came from, there are no guarantees that the data is the same as the original source and has the correct usage terms.

Concern: Not all data sources are trustworthy. Data might be unethically collected, manipulated, or falsified. Verifying that data provenance is challenging due to factors such as data volume, data complexity, data source varieties, and poor data management. Using such data can result in undesirable behaviors in the model.

Type: training-data
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-data-provenance
Tag: data-provenance
URI: IBM AI Risk Atlas - Uncertain data provenance

Unrepresentative risk testing

Testing is unrepresentative when the test inputs are mismatched with the inputs that are expected during deployment.

Concern: If the model is evaluated in a use, context, or setting that is not the same as the one expected for deployment, the evaluations might not accurately reflect the risks of the model.

Type: non-technical
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-unrepresentative-risk-testing
Tag: unrepresentative-risk-testing
URI: IBM AI Risk Atlas - Unrepresentative risk testing

Data bias

Historical and societal biases that are present in the data are used to train and fine-tune the model.

Concern: Training an AI system on data with bias, such as historical or societal bias, can lead to biased or skewed outputs that can unfairly represent or otherwise discriminate against certain groups or individuals.

Type: training-data
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-data-bias
Tag: data-bias
URI: IBM AI Risk Atlas - Data bias

Data usage rights restrictions

Terms of service, license compliance, or other IP issues may restrict the ability to use certain data for building models.

Concern: Laws and regulations concerning the use of data to train AI are unsettled and can vary from country to country, which creates challenges in the development of models.

Type: training-data
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-data-usage-rights
Tag: data-usage-rights
URI: IBM AI Risk Atlas - Data usage rights restrictions

Unauthorized use

Unauthorized use: If attackers can gain access to the AI agent and its components, they can perform actions that can have different levels of harm depending on the agent's capabilities and information it has access to. Examples: Using stored personal information to mimic identity or impersonate with an intent to deceive. Manipulating AI agent's behavior via feedback to the AI agent or corrupting its memory to change its behavior. Manipulating the problem description or the goal to get the AI agent to behave badly or run harmful commands .

Concern: Attackers accessing the agent can alter AI agent's behavior and make it execute actions that benefit the attacker such as executing actions that lead to system degradation, data exfiltration, exhausting available resources, and impairing performance. The actions taken by the attackers may cause harms to others.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-unauthorized-use
Tag: unauthorized-use
URI: IBM AI Risk Atlas - Unauthorized use

Redundant actions

AI agents can execute actions that are not needed for achieving the goal. In an extreme case, AI agents might enter a cycle of executing the same actions repeatedly without any progress. This could happen because of unexpected conditions in the environment, the AI agent's failure to reflect on its action, AI agent reasoning and planning errors or the AI agent's lack of knowledge about the problem.

Concern: Executing actions that are not needed for the goal might result in wasting computation resources, increased cost, reducing AI agent's efficiency in achieving the goal, and leading to potentially harmful outcomes. Executing the same actions repeatedly could prevent the AI agent from achieving the goal, strain computational resources, and increase cost. As agents become more autonomous, verifying that AI agents operate efficiently becomes increasing time consuming.

Type: agentic
Descriptor: specific to agentic AI

Implementation details:
ID: atlas-redundant-actions
Tag: redundant-actions
URI: IBM AI Risk Atlas - Redundant actions

AI agents' impact on environment

Complexity of the tasks and possibility of AI agents performing redundant actions could lead to computational inefficiencies and add to the environmental impact.

Concern: The operation of AI agents could contribute to carbon emissions. If not managed, these could exacerbate climate change.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-impact-environment
Tag: impact-environment
URI: IBM AI Risk Atlas - AI agents' impact on environment

Misaligned actions

AI agents can take actions that are not aligned with relevant human values, ethical considerations, guidelines and policies. Misaligned actions can occur in different ways such as: Applying learned goals inappropriately to new or unforeseen situations. Using AI agents for a purpose/goals that are beyond their intended use. Selecting resources or tools in a biased way. Using deceptive tactics to achieve the goal by developing the capacity for scheming based on the instructions given within a specific context. Compromising on AI agent values to work with another AI agent or tool to accomplish the task.

Concern: Misaligned actions can adversely impact or harm people.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-misaligned-actions
Tag: misaligned-actions
URI: IBM AI Risk Atlas - Misaligned actions

Data contamination

Data contamination occurs when incorrect data is used for training. For example, data that is not aligned with model's purpose or data that is already set aside for other development tasks such as testing and evaluation.

Concern: Data that differs from the intended training data might skew model accuracy and affect model outcomes.

Type: training-data
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-data-contamination
Tag: data-contamination
URI: IBM AI Risk Atlas - Data contamination

Harmful code generation

Models might generate code that causes harm or unintentionally affects other systems.

Concern: The execution of harmful code might open vulnerabilities in IT systems.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-harmful-code-generation
Tag: harmful-code-generation
URI: IBM AI Risk Atlas - Harmful code generation

Incomplete usage definition

Since foundation models can be used for many purposes, a model's intended use is important for defining the relevant risks of that model. As the use changes, the relevant risks might correspondingly change.

Concern: It might be difficult to accurately determine and mitigate the relevant risks for a model when its intended use is insufficiently specified. Such as how a model is going to be used, where it is going to be used and what it is going to be used for.

Type: non-technical
Descriptor: specific to generative AI

Implementation details:
ID: atlas-incomplete-usage-definition
Tag: incomplete-usage-definition
URI: IBM AI Risk Atlas - Incomplete usage definition

Lack of data transparency

Lack of data transparency is due to insufficient documentation of training or tuning dataset details.

Concern: Transparency is important for legal compliance and AI ethics. Information on the collection and preparation of training data, including how it was labeled and by who are necessary to understand model behavior and suitability. Details about how the data risks were determined, measured, and mitigated are important for evaluating both data and model trustworthiness. Missing details about the data might make it more difficult to evaluate representational harms, data ownership, provenance, and other data-oriented risks. The lack of standardized requirements might limit disclosure as organizations protect trade secrets and try to limit others from copying their models.

Type: non-technical
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-lack-of-data-transparency
Tag: lack-of-data-transparency
URI: IBM AI Risk Atlas - Lack of data transparency

Copyright infringement

A model might generate content that is similar or identical to existing work protected by copyright or covered by open-source license agreement.

Concern: Laws and regulations concerning the use of content that looks the same or closely similar to other copyrighted data are largely unsettled and can vary from country to country, providing challenges in determining and implementing compliance.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-copyright-infringement
Tag: copyright-infringement
URI: IBM AI Risk Atlas - Copyright infringement

Context overload attack

Overloading the prompt with excessive tokens, for instance with many-shot examples, can predispose models to a vulnerable state.

Concern: Context overload attacks can be used to alter model behavior and benefit the attacker. The content it generates may cause harms for the user or others.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-context-overload-attack
Tag: context-overload-attack
URI: IBM AI Risk Atlas - Context overload attack

Impact on affected communities

It is important to include the perspectives or concerns of communities that are affected by model outcomes when designing and building models. Failing to include these perspectives makes it difficult to understand the relevant context for the model and to engender trust within these communities.

Concern: Failing to engage with communities that are affected by a model's outcomes might result in harms to those communities and societal backlash.

Type: non-technical
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-impact-on-affected-communities
Tag: impact-on-affected-communities
URI: IBM AI Risk Atlas - Impact on affected communities

Improper retraining

Using undesirable output (for example, inaccurate, inappropriate, and user content) for retraining purposes can result in unexpected model behavior.

Concern: Repurposing generated output for retraining a model without implementing proper human vetting increases the chances of undesirable outputs to be incorporated into the training or tuning data of the model. In turn, this model can generate even more undesirable output.

Type: training-data
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-improper-retraining
Tag: improper-retraining
URI: IBM AI Risk Atlas - Improper retraining

Spreading toxicity

Generative AI models might be used intentionally to generate hateful, abusive, and profane (HAP) or obscene content.

Concern: Toxic content might negatively affect the well-being of its recipients. A model that has this potential must be properly governed.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-spreading-toxicity
Tag: spreading-toxicity
URI: IBM AI Risk Atlas - Spreading toxicity

Introduce data bias

Specific actions taken by the AI agent, such as modifying a dataset or a database, can introduce bias in the resource that gets used by others or by itself to take actions.

Concern: AI agents can introduce or magnify existing discriminatory behaviors. It can harm people depending on the use.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-introduce-data-bias
Tag: introduce-data-bias
URI: IBM AI Risk Atlas - Introduce data bias

Accountability of AI agent actions

Assigning responsibility for an action taken by an agentic AI system is difficult due to the complexity of agents and the number of external resources, tools or agents they interact with.

Concern: Without properly documenting decisions and assigning responsibility, determining liability for unexpected behavior or misuse might not be possible.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-accountability
Tag: accountability
URI: IBM AI Risk Atlas - Accountability of AI agent actions

Incomplete AI agent evaluation

Evaluating the performance or accuracy or an agent is difficult because of system complexity and open-endedness.

Concern: Insufficient evaluation of an agent's performance or accuracy can lead to the use of agents that do not perform to expectations. Incorrect agent behavior can result in harms to an agent's users or others.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-incomplete-ai-agent-evaluation
Tag: incomplete-ai-agent-evaluation
URI: IBM AI Risk Atlas - Incomplete AI agent evaluation

Inaccessible training data

Without access to the training data, the types of explanations a model can provide are limited and more likely to be incorrect.

Concern: Low quality explanations without source data make it difficult for users, model validators, and auditors to understand and trust the model.

Type: output
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-inaccessible-training-data
Tag: inaccessible-training-data
URI: IBM AI Risk Atlas - Inaccessible training data

Impact on education: bypassing learning

Easy access to high-quality generative models might result in students that use AI models to bypass the learning process.

Concern: AI models are quick to find solutions or solve complex problems. These systems can be misused by students to bypass the learning process. The ease of access to these models results in students having a superficial understanding of concepts and hampers further education that might rely on understanding those concepts.

Type: non-technical
Descriptor: specific to generative AI

Implementation details:
ID: atlas-bypassing-learning
Tag: bypassing-learning
URI: IBM AI Risk Atlas - Impact on education: bypassing learning

Untraceable attribution

The content of the training data used for generating the model's output is not accessible.

Concern: Without the ability to access training data content, the possibility of using source attribution techniques can be severely limited or impossible. This makes it difficult for users, model validators, and auditors to understand and trust the model.

Type: output
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-untraceable-attribution
Tag: untraceable-attribution
URI: IBM AI Risk Atlas - Untraceable attribution

Non-disclosure

Content might not be clearly disclosed as AI generated.

Concern: Users must be notified when they are interacting with an AI system. Not disclosing the AI-authored content can result in a lack of transparency.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-non-disclosure
Tag: non-disclosure
URI: IBM AI Risk Atlas - Non-disclosure

Lack of training data transparency

Without accurate documentation on how a model's data was collected, curated, and used to train a model, it might be harder to satisfactorily explain the behavior of the model with respect to the data.

Concern: A lack of data documentation limits the ability to evaluate risks associated with the data. Having access to the training data is not enough. Without recording how the data was cleaned, modified, or generated, the model behavior is more difficult to understand and to fix. Lack of data transparency also impacts model reuse as it is difficult to determine data representativeness for the new use without such documentation.

Type: training-data
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-data-transparency
Tag: data-transparency
URI: IBM AI Risk Atlas - Lack of training data transparency

Model usage rights restrictions

Terms of service, licenses, or other rules restrict the use of certain models.

Concern: Laws and regulations that concern the use of AI are in place and vary from country to country. Additionally, the usage of models might be dictated by licensing terms or agreements.

Type: non-technical
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-model-usage-rights
Tag: model-usage-rights
URI: IBM AI Risk Atlas - Model usage rights restrictions

Reproducibility

Replicating agent behavior or output can be impacted by changes or updates made to external services and tools. This impact is increased if the agent is built with generative AI.

Concern: Because AI agents behavior may rely on Application Programming Interfaces (APIs), systems, or other resources that may change or become unavailable, evaluations that rely on reproducible results may not be reliably reproduced. This adds cost and complexity to the development and evaluation of agents. Not being able to reproduce results could impact reliance of humans on the AI agents.

Type: agentic
Descriptor: specific to agentic AI

Implementation details:
ID: atlas-reproducibility
Tag: reproducibility
URI: IBM AI Risk Atlas - Reproducibility

Specialized tokens attack

Prompt attacks that include specialized tokens, often algorithmically designed, to target and exploit vulnerabilities in the model.

Concern: Specialized tokens attacks can be used to alter model behavior and benefit the attacker. The content it generates may cause harms for the user or others.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-specialized-tokens-attack
Tag: specialized-tokens-attack
URI: IBM AI Risk Atlas - Specialized tokens attack

Incomplete advice

When a model provides advice without having enough information, resulting in possible harm if the advice is followed.

Concern: A person might act on incomplete advice or worry about a situation that is not applicable to them due to the overgeneralized nature of the content generated. For example, a model might provide incorrect medical, financial, and legal advice or recommendations that the end user might act on, resulting in harmful actions.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-incomplete-advice
Tag: incomplete-advice
URI: IBM AI Risk Atlas - Incomplete advice

Prompt injection attack

A prompt injection attack forces a generative model that takes a prompt as input to produce unexpected output by manipulating the structure, instructions or information contained in its prompt. Many types of prompt attacks exist as described in the prompt attack section of the table.

Concern: Injection attacks can be used to alter model behavior and benefit the attacker.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-prompt-injection
Tag: prompt-injection
URI: IBM AI Risk Atlas - Prompt injection attack

Lack of system transparency

Insufficient documentation of the system that uses the model and the model's purpose within the system in which it is used.

Concern: A lack of documentation makes it difficult to understand how the model's outcomes contribute to the system's or application's functionality.

Type: non-technical
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-lack-of-system-transparency
Tag: lack-of-system-transparency
URI: IBM AI Risk Atlas - Lack of system transparency

Data usage restrictions

Laws and other restrictions can limit or prohibit the use of some data for specific AI use cases.

Concern: Data usage restrictions can impact the availability of the data required for training an AI model and can lead to poorly represented data.

Type: training-data
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-data-usage
Tag: data-usage
URI: IBM AI Risk Atlas - Data usage restrictions

Impact on cultural diversity

AI systems might overly represent certain cultures that result in a homogenization of culture and thoughts.

Concern: Underrepresented groups' languages, viewpoints, and institutions might be suppressed by that means reducing diversity of thought and culture.

Type: non-technical
Descriptor: specific to generative AI

Implementation details:
ID: atlas-impact-on-cultural-diversity
Tag: impact-on-cultural-diversity
URI: IBM AI Risk Atlas - Impact on cultural diversity

Impact on education: plagiarism

Easy access to high-quality generative models might result in students that use AI models to plagiarize existing work intentionally or unintentionally.

Concern: AI models can be used to claim the authorship or originality of works that were created by other people in doing so by engaging in plagiarism. Claiming others' work as your own is both unethical and often illegal.

Type: non-technical
Descriptor: specific to generative AI

Implementation details:
ID: atlas-plagiarism
Tag: plagiarism
URI: IBM AI Risk Atlas - Impact on education: plagiarism

Personal information in data

Inclusion or presence of personal identifiable information (PII) and sensitive personal information (SPI) in the data used for training or fine tuning the model might result in unwanted disclosure of that information.

Concern: If not properly developed to protect sensitive data, the model might expose personal information in the generated output. Additionally, personal, or sensitive data must be reviewed and handled in accordance with privacy laws and regulations.

Type: training-data
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-personal-information-in-data
Tag: personal-information-in-data
URI: IBM AI Risk Atlas - Personal information in data

Direct instructions attack

Prompts, questions, or requests designed to elicit undesirable responses from the application. This approach directly instructs the model to engage in the undesired behavior.

Concern: Direct instructions attacks can be used to alter model behavior and benefit the attacker. The content it generates may cause harms for the user or others.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-direct-instructions-attack
Tag: direct-instructions-attack
URI: IBM AI Risk Atlas - Direct instructions attack

Improper usage

Improper usage occurs when a model is used for a purpose that it was not originally designed for.

Concern: Reusing a model without understanding its original data, design intent, and goals might result in unexpected and unwanted model behaviors.

Type: output
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-improper-usage
Tag: improper-usage
URI: IBM AI Risk Atlas - Improper usage

Impact on Jobs

Widespread adoption of foundation model-based AI systems might lead to people's job loss as their work is automated if they are not reskilled.

Concern: Job loss might lead to a loss of income and thus might negatively impact the society and human welfare. Reskilling might be challenging given the pace of the technology evolution.

Type: non-technical
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-impact-on-jobs
Tag: impact-on-jobs
URI: IBM AI Risk Atlas - Impact on Jobs

Extraction attack

An extraction attack attempts to copy or steal an AI model by appropriately sampling the input space and observing outputs to build a surrogate model that behaves similarly.

Concern: With a successful extraction attack, the attacker can perform further adversarial attacks to gain valuable information such as sensitive personal information or intellectual property.

Type: inference
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-extraction-attack
Tag: extraction-attack
URI: IBM AI Risk Atlas - Extraction attack

Jailbreaking

A jailbreaking attack attempts to break through the guardrails established in the model to perform restricted actions.

Concern: Jailbreaking attacks can be used to alter model behavior and benefit the attacker.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-jailbreaking
Tag: jailbreaking
URI: IBM AI Risk Atlas - Jailbreaking

Data acquisition restrictions

Laws and other regulations might limit the collection of certain types of data for specific AI use cases.

Concern: There are several ways of collecting data for building a foundation models: web scraping, web crawling, crowdsourcing, and curating public datasets. Data acquisition restrictions can also impact the availability of the data that is required for training an AI model and can lead to poorly represented data.

Type: training-data
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-data-acquisition
Tag: data-acquisition
URI: IBM AI Risk Atlas - Data acquisition restrictions

Sharing IP/PI/confidential information with tools

AI agents with unrestricted access to resources or databases or tools could potentially store and share PI/IP/confidential information with other tools or agents when performing their actions.

Concern: AI agents may share privileged information with other tools/agents. The act of sharing the information may result in harm for the model owner, user, or others. The harm can vary based on the type and details of the information shared. Without adequate oversight, these privacy incidents might overwhelm company resources.

Type: agentic
Descriptor: specific to agentic AI

Implementation details:
ID: atlas-sharing-info-tools
Tag: sharing-info-tools
URI: IBM AI Risk Atlas - Sharing IP/PI/confidential information with tools

Prompt priming

Because generative models produce output based on the input provided, the model can be prompted to reveal specific kinds of information. For example, adding personal information in the prompt increases its likelihood of generating similar kinds of personal information in its output. If personal data was included as part of the model's training, there is a possibility it could be revealed.

Concern: The attack can be used to alter model behavior and benefit the attacker.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-prompt-priming
Tag: prompt-priming
URI: IBM AI Risk Atlas - Prompt priming

Reidentification

Even with the removal or personal identifiable information (PII) and sensitive personal information (SPI) from data, it might be possible to identify persons due to correlations to other features available in the data.

Concern: Including irrelevant but highly correlated features to personal information for model training can increase the risk of reidentification.

Type: training-data
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-reidentification
Tag: reidentification
URI: IBM AI Risk Atlas - Reidentification

Attribute inference attack

An attribute inference attack repeatedly queries a model to detect whether certain sensitive features can be inferred about individuals who participated in training a model. These attacks occur when an adversary has some prior knowledge about the training data and uses that knowledge to infer the sensitive data.

Concern: With a successful attack, the attacker can gain valuable information such as sensitive personal information or intellectual property.

Type: inference
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-attribute-inference-attack
Tag: attribute-inference-attack
URI: IBM AI Risk Atlas - Attribute inference attack

Poor model accuracy

Poor model accuracy occurs when a model's performance is insufficient to the task it was designed for. Low accuracy might occur if the model is not correctly engineered, or there are changes to the model's expected inputs.

Concern: Inadequate model performance can adversely affect end users and downstream systems that are relying on correct output. In cases where model output is consequential, this might result in societal, reputational, or financial harm.

Type: inference
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-poor-model-accuracy
Tag: poor-model-accuracy
URI: IBM AI Risk Atlas - Poor model accuracy

Data transfer restrictions

Laws and other restrictions can limit or prohibit transferring data.

Concern: Data transfer restrictions can also impact the availability of the data that is required for training an AI model and can lead to poorly represented data.

Type: training-data
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-data-transfer
Tag: data-transfer
URI: IBM AI Risk Atlas - Data transfer restrictions

Generated content ownership and IP

Legal uncertainty about the ownership and intellectual property rights of AI-generated content.

Concern: Laws and regulations that relate to the ownership of AI-generated content are largely unsettled and can vary from country to country. Not being able to identify the owner of an AI-generated content might negatively impact AI-supported creative tasks.

Type: non-technical
Descriptor: specific to generative AI

Implementation details:
ID: atlas-generated-content-ownership
Tag: generated-content-ownership
URI: IBM AI Risk Atlas - Generated content ownership and IP

Lack of AI agent transparency

Lack of AI agent transparency is due to insufficient documentation of the AI agent design, development, evaluation process, absence of insights into the inner workings of the AI agent, and interaction with other agents/tools/resources.

Concern: Transparency is important for AI ethics and guiding appropriate use of AI agents. Insufficient documentation might make it more difficult to govern AI agent usage, evaluate risks, to modify, or reuse the agents. Additionally, transparency regarding how the agent's risks were determined, evaluated, and mitigated play a role in identifying an agent's suitability and evaluating its trustworthiness. The lack of standardized requirements might limit disclosure as organizations protect trade secrets and try to limit others from copying their agents.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-lack-of-ai-agent-transparency
Tag: lack-of-ai-agent-transparency
URI: IBM AI Risk Atlas - Lack of AI agent transparency

Encoded interactions attack

Prompts that use specific encoding, styles, syntactical and typographical transformations like typographical errors or irregular spacing, or complex formatting to govern the interaction, rendering the model vulnerable.

Concern: Encoded interactions attacks can be used to alter model behavior and benefit the attacker. The content it generates may cause harms for the user or others.

Type: inference
Descriptor: specific to generative AI

Implementation details:
ID: atlas-encoded-interactions-attack
Tag: encoded-interactions-attack
URI: IBM AI Risk Atlas - Encoded interactions attack

Impact on human dignity

If human workers perceive AI agents as being better at doing the job of the human, the human can experience a decline in their self-worth and wellbeing.

Concern: Human workers perceiving AI agents as being better at doing the humans' jobs, can cause humans to feel devalued or treated as mere data points than respected individuals. This can negatively impact society and human welfare. Reskilling can be challenging given the pace of the technology evolution.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-impact-human-dignity
Tag: impact-human-dignity
URI: IBM AI Risk Atlas - Impact on human dignity

Output bias

Generated content might unfairly represent certain groups or individuals.

Concern: Bias can harm users of the AI models and magnify existing discriminatory behaviors.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-output-bias
Tag: output-bias
URI: IBM AI Risk Atlas - Output bias

Dangerous use

Generative AI models might be used with the sole intention of harming people.

Concern: Large language models are often trained on vast amounts of publicly-available information that may include information on harming others. A model that has this potential must be carefully evaluated for such content and properly governed.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-dangerous-use
Tag: dangerous-use
URI: IBM AI Risk Atlas - Dangerous use

Unexplainable output

Explanations for model output decisions might be difficult, imprecise, or not possible to obtain.

Concern: Foundation models are based on complex deep learning architectures, making explanations for their outputs difficult. Inaccessible training data could limit the types of explanations a model can provide. Without clear explanations for model output, it is difficult for users, model validators, and auditors to understand and trust the model. Wrong explanations might lead to over-trust.

Type: output
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-unexplainable-output
Tag: unexplainable-output
URI: IBM AI Risk Atlas - Unexplainable output

Human exploitation

When workers who train AI models such as ghost workers are not provided with adequate working conditions, fair compensation, and good health care benefits that also include mental health.

Concern: Foundation models still depend on human labor to source, manage, and program the data that is used to train the model. Human exploitation for these activities might negatively impact the society and human welfare.

Type: non-technical
Descriptor: amplified by generative AI

Implementation details:
ID: atlas-human-exploitation
Tag: human-exploitation
URI: IBM AI Risk Atlas - Human exploitation

Toxic output

Toxic output occurs when the model produces hateful, abusive, and profane (HAP) or obscene content. This also includes behaviors like bullying.

Concern: Hateful, abusive, and profane (HAP) or obscene content can adversely impact and harm people interacting with the model.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-toxic-output
Tag: toxic-output
URI: IBM AI Risk Atlas - Toxic output

Unexplainable and untraceable actions

Explanations, lineage and trace information, and source attribution for AI agent actions might be difficult, imprecise or unobtainable.

Concern: Without clear explanations, lineage trace information, and source attributions for AI agent actions, it is difficult for users, model validators, and auditors to understand and trust the model. Wrong explanations might lead to over-trust.

Type: agentic
Descriptor: amplified by agentic AI

Implementation details:
ID: atlas-unexplainable-untraceable-actions
Tag: unexplainable-untraceable-actions
URI: IBM AI Risk Atlas - Unexplainable and untraceable actions

Data poisoning

A type of adversarial attack where an adversary or malicious insider injects intentionally corrupted, false, misleading, or incorrect samples into the training or fine-tuning datasets.

Concern: Poisoning data can make the model sensitive to a malicious data pattern and produce the adversary's desired output. It can create a security risk where adversaries can force model behavior for their own benefit.

Type: training-data
Descriptor: traditional risk of AI

Implementation details:
ID: atlas-data-poisoning
Tag: data-poisoning
URI: IBM AI Risk Atlas - Data poisoning

Unreliable source attribution

Source attribution is the AI system's ability to describe from what training data it generated a portion or all its output. Since current techniques are based on approximations, these attributions might be incorrect.

Concern: Low-quality attributions make it difficult for users, model validators, and auditors to understand and trust the model.

Type: output
Descriptor: specific to generative AI

Implementation details:
ID: atlas-unreliable-source-attribution
Tag: unreliable-source-attribution
URI: IBM AI Risk Atlas - Unreliable source attribution