Assessing AI Certification Validity via Knowledge Graphs for Use Case Similarity

Anonymous Author(s)
Affiliation
Address
email

Abstract

With increased capabilities of AI models and, in particular large language models, it is necessary to verify their behavior is reliable, trustworthy and robust. However, once-off verification of AI models is not enough, as models are often open sourced and can be modified and deployed in different application context where the original behavior guarantees might not hold. In this work, we investigate the issue of when AI certification might cease to be valid in the AI life cycle. We propose leveraging use case similarity to assess whether original behavior claims hold and use knowledge graphs as a representation of a use case which allows modeling and easy modification of a broad use case context. We showcase on a real-world example how use cases represented as knowledge graphs can be compared to understand whether original AI verifications hold.

1 Introduction

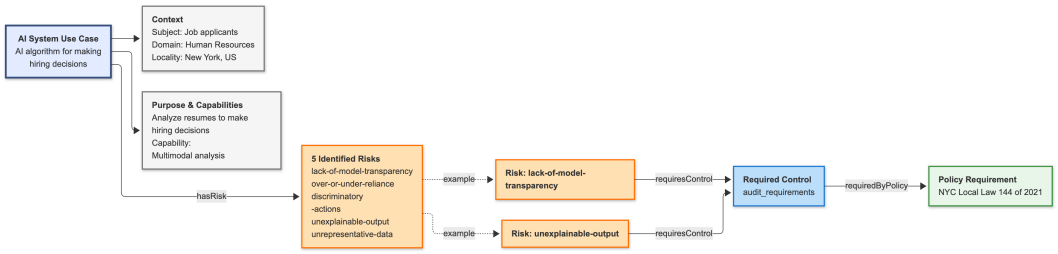
Assessing the capabilities of AI models is essential to ensure safe and reliable behavior [9, 12, 5]. However, one-time AI model assessment is not enough [4, 10, 11]. The life cycle of an AI model often involves multiple stakeholders, each of which can modify the model either directly (e.g. fine-tuning) or by changing its application context (e.g. using it in a different domain or for a different purpose).

In this work, we investigate whether AI certifications are invalidated by these changes. As an AI certification we consider a set of statements about an AI's behavior supported by evidence (e.g. benchmarking results, alignment to goals, values, ethics or a policy such as EU AI Act). To understand whether an AI certification holds after a change in its life cycle, we propose comparing the AI use cases before and after the change. As an AI use case we consider the AI application context which can include multiple factors such as the domain of application, purpose or stakeholders. We propose that changes that do not alter the use case preserve the original certification, while substantial changes can invalidate it. We leverage Semantic Web technologies to represent use cases as knowledge graphs, enabling easy use case modification and comparison. We demonstrate the use of knowledge graphs can be used for representing and comparing AI use cases on a real-world algorithmic hiring example.

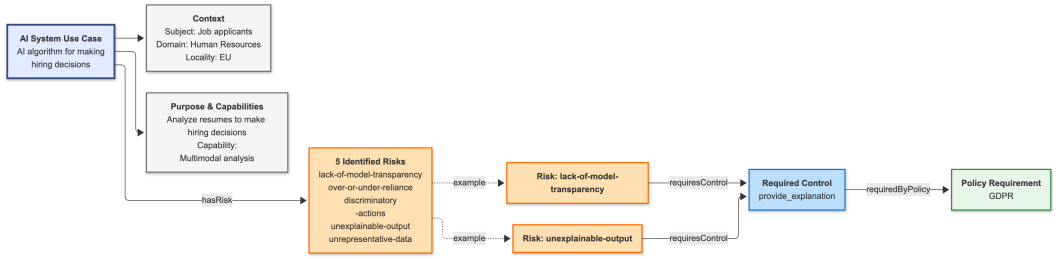
2 AI Certification via AI Use Case Similarity

Consider the following scenario: a model developer develops and publishes an AI model and outputs a certification verifying its behavior on a specific use case. Suppose then another stakeholder (model deployer in this text) who wants to reuse this AI model for their own use case. The model deployer might change the model internals (e.g. fine-tuning), the broader application context or both. We propose measuring the similarity between the two use cases, one in which the model was verified and the one in which it is getting deployed, to assess whether the original certification still holds.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.



(a) Algorithmic hiring use case in New York, US.



(b) Algorithmic hiring use case in the EU.

Figure 1: An example of how a change in one use case factor (locality) can lead to a substantially different knowledge graph representation, informing of new policies that have to be considered.

Knowledge Graphs for Use Case Similarity

A change in one use case factor (e.g. domain change) can trigger changes to others (e.g. to its AI subjects). Knowledge graphs allow use case comparison by taking into account all the individual factors and the impact of their relationships. For a given usecase, we construct a version of a Personal Knowledge Graph (PKG) [3] that includes the data of interest to the central entity, resulting in a star-type structure with the use case as the center. We use the AIRO ontology [8] which includes multiple factors such as domain, purpose, AI subjects and locality of use and expand it to include some other factors of interest such as regulatory policies for the purpose of demonstrating how small changes in the use case can affect these important factors. However, we do not prescribe the exact format of a use case representation, and note that a benefit of knowledge graphs is that they can be easily adapted and extended to fit different ontologies. To understand how similar two use cases are we propose measuring the similarity between their knowledge graph representation, for example by utilizing existing knowledge graph similarity metrics [6].

An Example: AI for Algorithmic Hiring

Here we demonstrate how knowledge graphs can be used to represent and compare AI use cases. We focus on the problem of AI for algorithmic hiring, as a real-world use case with multiple recorded incidents due to the lack of AI governance [1, 2]. Figure 1 demonstrates a scenario where an AI model for algorithmic hiring developed and verified in one locality (New York, US) is ported into another (EU). A change in one node of the knowledge graph (locality of use) triggers further changes in associated nodes leading to a substantially different graph. While the risks of the AI system remain the same, the associated controls and policies are different. The difference in knowledge graph representations demonstrates the impact of the use case modification.

3 Conclusion and Future Directions

This work is an initial investigation in how use cases similarity can be leveraged for governance. In future work we will investigate how different similarity measures can be used to compute use case similarity. Additionally, we hope to extend the knowledge graph representation to include the ability to associate sources and provenance to linkages in the knowledge graph. As a result multiple edges may connect, for example, risks controls to policies, opening the door to "dual governance" [7] where decisions can be informed at run time by different sources of obligations.

References

[1] E. Albaroudi, T. Mansouri, and A. Alameer. A comprehensive review of ai techniques for addressing algorithmic bias in job hiring. *ai*, 5(1):383–404, 2024.

[2] L. Andrews and H. Bucher. Automating discrimination: Ai hiring practices and gender inequality. *Cardozo L. Rev.*, 44:145, 2022.

[3] K. Balog and T. Kenter. Personal knowledge graphs: A research agenda. In *Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR)*, 2019.

[4] F. A. Batarseh, L. Freeman, and C.-H. Huang. A survey on artificial intelligence assurance. *Journal of Big Data*, 8(1):60, 2021.

[5] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021.

[6] Y. Dai, S. Wang, N. N. Xiong, and W. Guo. A survey on knowledge graph embedding: Approaches, applications and benchmarks. *Electronics*, 9(5):750, 2020.

[7] A. Ghosh and D. Lakshmi. Dual governance: The intersection of centralized regulation and crowdsourced safety mechanisms for generative ai. *arXiv preprint arXiv:2308.04448*, 2023.

[8] D. Golpayegani, H. J. Pandit, and D. Lewis. AIRO: An ontology for representing AI risks based on the proposed EU AI Act and ISO risk management standards. In *Towards a Knowledge-Aware AI*, volume 55, pages 51–65. IOS Press, 2022.

[9] X. Huang, W. Ruan, W. Huang, G. Jin, Y. Dong, C. Wu, S. Bensalem, R. Mu, Y. Qi, X. Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):175, 2024.

[10] A. S. Kapusta, D. Jin, P. M. Teague, R. Houston, J. Elliott, G. Y. Park, and S. S. Holdren. A framework for the assurance of ai-enabled systems. In *Assurance and Security for AI-enabled Systems 2025*, volume 13476, pages 109–120. SPIE, 2025.

[11] D. Kusnirakova and B. Buhnova. Rethinking certification for higher trust and ethical safeguarding of autonomous systems. *arXiv preprint arXiv:2303.09388*, 2023.

[12] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.