A Practical Approach to Fostering Trust in AI via Semantically Supported Accountability

Ali Hashmi
IBM Consulting
ahashmi@us.ibm.com

Abstract

In this demonstration (or poster), we will be presenting a walkthrough of a practical approach to elevating trust in artificial intelligence (AI) enabled systems using semantically grounded approaches for improved transparency and well-defined accountability. Within the larger context of the responsible use of AI, this investigation seeks to demonstrate actionable steps toward that goal while surfacing lessons learned and challenges faced along the way. We leverage semantic techniques and tools, including ontologies and automated reasoning, to operationalize the communication of accountability and the transparency information throughout the design, deployment, and ongoing evaluation of AI-enabled systems. A simplified use case related to the development of a clinical decision support system (CDSS) will be the backdrop for this investigation.

1 Background

The pervasive integration of artificial intelligence (AI) systems into our daily lives has brought about a heightened awareness of the importance of trust in AI from a user perspective. AI-enabled systems, which can serve as beneficial tools used by human agents or, in specific domains, evolve into agent assistants, are positioned to significantly impact human agency, decision-making, and outcomes. Trust in AI plays a pivotal role in its effective adoption across a variety of applications.

As AI algorithms become increasingly sophisticated and autonomous, their decision-making processes can become opaque, making it difficult for individuals to understand how these systems are shaping their lives. This lack of transparency, coupled with the potential for AI systems to perpetuate biases, cause unintended harm, and infringe upon human rights, has led to the call for greater accountability in AI governance [Varshney, 2022].

While the notion of trust must ultimately be examined in a multi-dimensional way [Ashoori & Weisz, 2019], transparency and accountability are widely recognized as essential principles for responsible AI development and deployment. Transparency enables people to better understand how AI systems arrive at their results, while accountability ensures that there are clear mechanisms for assigning responsibility and providing redress when these systems cause harm [Afroogh *et al.*, 2024]. However, implementing these principles in practice is challenging, and de facto methods and standards are still emerging. Herein, we will explore the use of semantic approaches to achieve greater levels of transparency and accountability form this practical perspective.

1.1 Benefits of Semantic Approach

Utilizing the knowledge representation tools and techniques from the semantic web toward realizing accountable AI systems brings several notable benefits. Firstly, ontologies model accountability and transparency information in a structured, interoperable format that is easily understood by both humans and machines. Secondly, accountability plans within these structures guide and assess the collection of accountability trace information, going beyond basic documentation forms seen in other frameworks[1]. Additionally, ontologies support automated reasoning for efficient and rigorous checking of completeness, consistency, and accuracy. Lastly, the inherently flexible nature of sematic approaches facilitates easy translation and alignment to existing frameworks, thereby enabling robust tracking and verification of accountability and transparency throughout the entire lifecycle of AI development. [Naja, 2022] Indeed, semantic approaches have shown promising results for the two most widely investigated classification and segmentation problems in medical image analysis [Yang, 2022].

2 Methods

For this demonstration, an example use case of an AI-enabled solution has been drawn from the field of Clinical Decision Support Systems (CDSS). In this context, we will explore the applicability of semantic approaches toward improving transparency and illuminating accountability.

2.1 Example Use Case: Risk Prediction

CDSS are the computer programs that assist healthcare professionals in making medical decisions. CDSS interventions have been shown to enhance healthcare quality by facilitating adherence to clinical guidelines, reducing medication errors,

[1] e.g. AI Datasheets, Model Cards, and FactSheets.

and minimizing adverse drug events, all culminating in improved patient outcomes [Elhaddad, 2024].

AI-enabled systems increasingly play an important role in the early detection of adverse drug events and toxicity, incorporating a range of AI-based methodologies from anomaly detection, predictive modeling, to deep learning. This example use case will focus on the transparency and accountability of AI development in clinical risk prediction and prevention.

2.2 Application of Semantic Approach

The foundation for this investigation is based on an adaptation of prior work in this space by both Naja [2021] and Fernadez [2023]. Generally, these works apply ontologies and automatic reasoning to AI governance, providing open-source libraries to reproduce their results. Relevant assets include:

- The SAO ontology, a lightweight generic ontology for describing accountability plans and corresponding provenance traces of computational systems [Naja, 2021]
- The RAInS ontology, which extends SAO to model accountability information relevant to AI systems [Naja, 2021]
- The FIDES ontology-based approach towards achieving the accountability of AI/ML systems, where all the relevant information related to the modeling is semantically annotated [Fernandez, 2023]

These approaches are applied to the use case, and results are aligned to the broader guidance offered by the NIST AI Risk Management Framework [2023].

3 Discussion

Along the road to realizable promotion of trust in AI-enabled systems, the questions and challenges of operationalizing reliable and meaningful governance. To that end, we look to test how the incorporation semantic knowledge representation methods supporting increased transparency and clarifying accountability may contribute to achieving these practical goals:

- Automation of key labor-intensive steps
- Information extraction from source material and processes
- Lower cognitive load on developer and practitioners
- Wise utilization of modularity and re-use
- Take advantage of established, pre-developed libraries and assets to the extent possible.

In this investigation we are reporting on the degree to which the application of semantic approaches in AI governance have measured up against these goals. Preliminary proof-of-concept results include improvements in information extraction and other interoperability features.

4 Conclusion

Engendering trust is paramount in the responsible development of AI-enabled systems. The need for transparency and accountability in the design and operation of these systems has been highlighted. We presented a practical approach to elevating trust in AI systems using semantically supported accountability and transparency reporting, leveraging technologies such as ontologies and automated reasoning to operationalize the capture of key information throughout the design, deployment, and ongoing evaluation of AI-enabled systems. The use case of a simplified CDSS is used as the backdrop for this investigation.

References

[Naja *et al*., 2021] Naja, I., Markovic, M., Edwards, P., & Cottrill, C. (2021). A Semantic Framework to Support AI System Accountability and Audit. In R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, & M. Alam (Eds.), The Semantic Web (pp. 160–176). Springer International Publishing. https://doi.org/10.1007/978-3-030-77385-4_10

[Fernandez *et al*., 2023] Fernandez, I., Aceta, C., Gilabert, E., & Esnaola-Gonzalez, I. (2023). FIDES: An ontology-based approach for making machine learning systems accountable. Journal of Web Semantics, 79, 100808. https://doi.org/10.1016/j.websem.2023.100808

[Varshney, 2022] Varshney, K. R. (2022). Trustworthy Machine Learning. http://www.trustworthymachinelearning.com/

[Naja *et al*., 2022] Naja, I., Markovic, M., Edwards, P., Pang, W., Cottrill, C., & Williams, R. (2022). Using Knowledge Graphs to Unlock Practical Collection, Integration, and Audit of AI Accountability Information. IEEE Access, 10, 74383–74411. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3188967

[Elhaddad *et al*., 2024] Elhaddad, M., & Hamam, S. (2024). AI-Driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential. Cureus, 16(4), e57728. https://doi.org/10.7759/cureus.57728

[Afroogh *et al*., 2024] Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. Humanities and Social Sciences Communications, 11(1), 1–30. https://doi.org/10.1057/s41599-024-04044-8

[Tabassi, 2023] Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (No. NIST AI 100-1; p. NIST AI 100-1). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.AI.100-1

[Ashoori & Weisz, 2019] Ashoori, M., & Weisz, J. D. (2019). In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes (No. arXiv:1912.02675). arXiv. https://doi.org/10.48550/arXiv.1912.02675

[Yang *et al*, 2022] Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, *77*, 29–52. https://doi.org/10.1016/j.inffus.2021.07.016