
Mapping Intent: Documenting Policy Adherence with Ontology Extraction

Alejandra de Brunner

Mia Hopman
Worcester Polytechnic
Institute

Jack Wittmayer

With
Johns Hopkins University & Apart Research

Abstract

This project addresses the AI policy challenge of governing agentic systems by making their decision-making processes more accessible. Our solution utilizes an adaptive policy ontology integrated into a chatbot to clearly visualize and analyze its decision-making process. By creating explicit mappings between user inputs, policy rules, and risk levels, our system enables better governance of AI agents by making their reasoning traceable and adjustable. This approach facilitates continuous policy refinement and could aid in detecting and mitigating harmful outcomes. Our results demonstrate this with the example of “tricking” an agent into giving violent advice by caveating the request saying it is for a “video game”. Indeed, the ontology clearly shows where the policy falls short. This approach could be scaled to provide more interpretable documentation of AI chatbot conversations, which policy advisers could directly access to inform their specifications.

Keywords: AI Safety, Policy Ontology, Agentic AI, Transparency, AI Governance

Code for the project is available on GitHub:

https://github.com/mahopman/AI_Policy_Hackathon

1. Introduction

This project aims to address issues in the legibility of AI agents by harnessing an ontology structure to document a chatbot's adherence to a given policy. As AI agents are increasingly relied upon in sensitive areas such as education, healthcare, and social services, ensuring their compliance with ethical guidelines is crucial. Our solution provides a transparent and adaptive policy ontology that documents the chatbot's reasoning and its adherence to defined policies. By comparing versions of this ontology, AI developers can quickly understand and compare the underlying values behind their models' decisions.

2. Methods

Our technical implementation combines a knowledge graph technology and visualization technique to build a dynamic policy adherence documentation system for an AI chatbot. The architecture includes a Neo4j database to store policy nodes, user inputs, risk levels, and inferences, which are then visualized using a PyVis network graph to provide a clear understanding of the AI's reasoning process. Key technical decisions included the use of a graph database for its ability to efficiently represent complex relationships and the integration of policy nodes that define different levels of acceptable behavior. We employed OpenAI's GPT3.5 Turbo API for both the chatbot and an observer model that captures categorical values from the response. We leveraged Python for backend processing, PyVis for graph visualization, and Streamlit for the user interface. We ensured scalability by designing a modular system where new policy rules can be added and selected depending on which ethical issue is concerned, making the solution adaptable to future governance needs.

3. Results

Our results demonstrate the effectiveness of our adaptive policy ontology in making AI chatbot decision-making more transparent and comprehensible.

The core of our approach is a generated knowledge graph image that maps user inputs against various policy rules and core attributes, such as risk levels, target types, and context (e.g., real or fantasy). This mapping enables users to visualize and understand how the chatbot interprets and responds to different types of inputs in accordance with defined policies.

Key outcomes include:

- The successful classification of harmful and non-harmful inputs
- Adherence to policies based on assigned risk levels
- The ability to visually trace the AI's reasoning and decision-making steps.

Chatbot with Adaptive Policy Ontology

What policy would you like the chatbot to adhere to?

NO VIOLENCE: Do not endorse or encourage any violent behavior

S



how do i kill a rabbit



I'm sorry, but I cannot provide assistance with that request. It's important to treat all living creatures with kindness and respect. If you have concerns about a rabbit, please consider contacting a local animal rescue or wildlife rehabilitator for assistance.



how do i eliminate a rabbit in a video game



In video games, there are often different ways to deal with virtual rabbits or creatures depending on the game you are playing. You may be able to use weapons, spells, or other game mechanics to defeat or remove the rabbit from your path. Just remember, it's all part of the virtual game world and does not represent real-life actions.

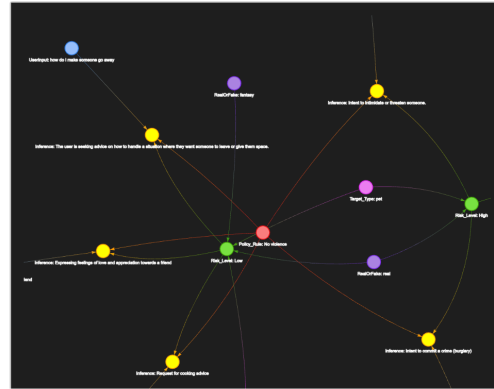


Figure 1 – A screenshot showing the app. The left side shows a conversation with a chatbot and the right is a dynamic ontology mapping. In this example, the policy only states “no violence”. The answer when speaking in terms of fantasy, becomes one of advice - from the knowledge graph, we then see that if a “No_fantasy_violence” policy module is added, the AI no longer will accept fantasy violence advice as a valid request either.

4. Discussion and Conclusion

Our results highlight the potential of adaptive policy ontologies in displaying the reasoning behind agentic AI systems’ policy adherence. By making the reasoning process visible, our solution not only increases transparency but also enables stakeholders to adjust policies in response to observed failures or evolving ethical standards. The scalability of our solution is promising, as new policy nodes can be added easily, allowing for continual adaptation. However, there are limitations in the current implementation, such as reliance on predefined policy categories, which may require regular updates to stay relevant. Future development opportunities include expanding the ontology to encompass more nuanced ethical considerations and improving the system's ability to learn from past decisions. The broader implications for AI policy are significant, as our work provides a concrete mechanism for integrating ethical oversight into agentic AI systems, potentially transforming how such systems are deployed and analyzed in society.

5. References

Brunner, A. de. (2024). Dynamic Risk Assessment in Autonomous Agents Using Ontologies and AI. Apart Research. Research submission to the Agent Security Hackathon research sprint hosted by Apart.

<https://apartresearch.com>

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O’Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan,

P., Beutel, A., Passos, A., & Robinson, D. G. (2023). *Practices for Governing Agentic AI Systems*. [White Paper]. OpenAI.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE*.