Statement of Interest

Designing Provenance-Aware AI Systems for Trustworthy Information Use

Motivation and Alignment with the Research Challenge

The increasing reliance on AI-mediated information systems has amplified societal risks associated with misinformation and disinformation, particularly in domains where decisions carry real-world consequences. While advances in generative and retrieval models have made information more accessible, they have also made it more difficult for end users to understand why specific claims are presented, what evidence supports them, and how uncertainty or disagreement should be interpreted. In many cases, misinformation persists not because information is unavailable, but because existing systems fail to expose provenance, context, and reasoning in ways that users can meaningfully inspect and trust. Recent work has shown that user trust in AI systems depends not only on accuracy, but on the availability of transparent reasoning and provenance that enables meaningful inspection and error recovery [1, 2].

My research is motivated by the view that misinformation is fundamentally a **systems and provenance challenge**, rather than solely a modeling or classification problem. I am interested in how AI systems can be designed so that claims, sources, evidence, and transformations are explicitly represented and traceable, enabling users to interrogate outputs rather than passively consume them. This motivation aligns directly with the *AI for Global and Societal Impact* research challenge, and specifically with the goal of tackling misinformation through the design, implementation, and evaluation of provenance tools for end users. Prior work supports this perspective, demonstrating that effective responses to misinformation require system-level approaches that integrate evidence, context, and human interpretability rather than relying solely on automated classification [3].

Rather than focusing exclusively on automated detection or suppression, my work emphasizes **human-centered, provenance-aware system design** that supports trust calibration, sensemaking, and informed decision-making in environments where information is incomplete, contested, or evolving.

Relevant Qualifications, Expertise, and Perspective

I bring a systems-oriented perspective grounded in software engineering, AI system architecture, and applied research. My work focuses on building end-to-end AI-enabled systems that integrate data ingestion, structured reasoning, and interpretability, with particular attention to transparency and accountability.

Across multiple applied domains, I have designed architectures that treat interpretability and provenance as first-class concerns rather than post-hoc features. In policy- and safety-relevant systems, I have encountered recurring challenges such as incomplete data, lack of ground truth, and evolving definitions of correctness. These constraints reinforce the importance of systems that expose assumptions, surface uncertainty, and allow users to inspect how conclusions are derived.

This background positions me to contribute meaningfully to research on provenance tools for misinformation. I approach this problem not only as an AI researcher, but as a system builder concerned with

how real users interact with AI outputs, how trust is established or eroded, and how transparency can be operationalized at scale.

Preliminary Research and Ongoing Work

My current doctoral research includes the **Research.AI** project, which serves as a concrete foundation for my interest in provenance-aware AI systems. Research.AI explores how agentic AI systems, combined with knowledge graphs, can support transparent research discovery and structured reasoning over complex, unstructured corpora such as scholarly literature and technical reports.

In this work, I have developed prototype systems that:

- Extract claims, entities, and relationships into structured knowledge representations
- Track source documents, transformations, and reasoning steps as explicit provenance metadata
- Use agent-based orchestration to reason over evidence rather than raw text
- Surface reasoning paths and supporting sources to enable user inspection and verification

Although Research.AI is focused on scholarly knowledge, the underlying challenges closely parallel those found in misinformation contexts: fragmented sources, conflicting claims, varying credibility, and the need for users to understand why an AI system arrives at a given conclusion. This project demonstrates both technical progress and sustained investment in provenance-centric system design, and it provides a transferable framework for studying how provenance tools can improve trust and understanding in broader societal information environments. These challenges mirror those identified in prior work on knowledge-based AI systems and scholarly knowledge graphs, which emphasize explicit structure, provenance, and human-interpretable reasoning over unstructured text alone [4, 5].

Proposed Research Direction and Collaboration with Microsoft Research

Within the *AI for Global and Societal Impact* challenge, I am interested in collaborating with Microsoft Research on the design and evaluation of **end-user-facing provenance tools** that support transparent interaction with AI-mediated information.

Specific areas of collaboration include:

- Designing system architectures that integrate provenance tracking across data ingestion, retrieval, and reasoning components
- Developing representations that balance machine interpretability with human usability
- Evaluating provenance tools using human-centered metrics such as trust calibration, comprehension, and error detection
- Exploring deployment contexts where misinformation has tangible societal impact

Collaboration with Microsoft researchers and applied scientists would enable this work to be pursued with greater scale, rigor, and real-world relevance. Microsoft Research's expertise in responsible AI, human-centered systems, and large-scale deployment makes it an ideal environment to advance this line of inquiry. Evaluating such systems requires metrics that go beyond predictive accuracy to capture trust calibration, user comprehension, and appropriate reliance on AI-supported information [6].

Closing

My research goals closely align with the objectives of the *AI for Global and Societal Impact* challenge. By focusing on provenance, structured reasoning, and system-level transparency, I aim to contribute tools and insights that help users better understand, evaluate, and trust AI-mediated information. I see this fellowship as an opportunity to deepen this work through collaboration with Microsoft Research and to help advance AI systems that support informed, trustworthy engagement in complex information ecosystems.

References

References

[1] F. Doshi-Velez and B. Kim, *Towards a rigorous science of interpretable machine learning*, arXiv:1702.08608, 2017.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, *"Why should I trust you?" Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[3] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, *FEVER: A large-scale dataset for fact extraction and verification*, Proceedings of NAACL-HLT, 2018.

[4] S. Auer et al., *Towards a knowledge graph for scholarly communication*, Semantic Web, 9(1), 1–15, 2018.

[5] M. Y. Jaradeh et al., *Open Research Knowledge Graph: Next generation infrastructure for semantic scholarly knowledge*, Proceedings of the 10th International Conference on Knowledge Capture (K-CAP), 2019.

[6] R. Hoffmann et al., *Evaluating the trustworthiness of AI systems*, Communications of the ACM, 65(7), 52–59, 2022.