Research proposal

Master of Research - Business Analytics - S2 2024

Prepared by

Taylor Chu - LinkedIn - hathanh.chu@students.mq.edu.au

AREA OF RESEARCH - RESPONSIBLE AI

Since the beginning of the internet and especially over the recent pandemic, we have evolved to live much of our lives online, and the need for us to want our lives to be enhanced by the digital world only seems to continue to grow. The internet provides convenience, efficient access to information, and a lot of opportunities, whether financial, social or emotional. As we continue to interact more frequently with the digital world, our information makes up a digital “fingerprint”. The vulnerability of our information is becoming more and more of an urgent issue, and it is a significant socio-technical challenge to keep the benefits of data-driven technology while protecting people’s privacy. (Australian Government - Attorney General's Department, 2022)

In Australia, the Privacy Act Review was released in 2022 after two years of work, contributing to a major rewrite after 40 years. The review has estimated it could cost up to $2200 per customer for big businesses to overhaul how they structure and store data to comply with new laws. In addition to the risks of privacy invasion from data prevalence, the government is also actively working on updating guidelines and laws to ensure safe and responsible AI, having called for public consultation in 2023.

Together with the rise of the digital economy and the amount of data available, recent years have seen exponential improvements in artificial intelligence, with even more improvements possible in the coming decades. Due to its power and potential of being used maliciously, AI is recognized by the Center for the study of Existential Risk at University of Cambridge (*Risks From Artificial Intelligence,* 2023) as one of the risks that could lead to human extinction or civilisational collapse. In both the short-term and long-term, AI should be directed to develop in a safe and beneficial direction to humanity, ensuring fairness and transparency throughout.

One of the reasons for the public’s lack of trust in AI systems is that explainability features are not common in consumer AI applications. Most applications intentionally obscure the AI-powered parts of the product irrelevant to the users’ goal, which is also known as the “black-box” approach (Abedin, 2022). On the other hand, non-technical users may have less tolerance for complex explanations of the AI models. Research has found that trust for an AI system increases when the system provides information on its reasoning, while it decreases when the system provides information on sources of uncertainty. In the end, finding the right balance would be important to create better user interactions with AI-powered products.

Governments, companies, and research centers around the world are doing research into technical and governance frameworks to ensure the development of safe and responsible AI throughout its lifecycle - from data collection, training and development, to testing, launching and monitoring.

LITERATURE REVIEW

Since the launch of ChatGPT by OpenAI in November 2022, the rapid expansion of ChatGPT and other large language models (LLMs) show that the potential of generative AI technologies is difficult to predict over the next two years, let alone ten (Bell et al., 2023). LLM technologies are used in many different industries, from healthcare, manufacturing, to finance, retail and creative industries. These technologies, and the applications and businesses built upon them, have amplified long-standing public and expert concerns about the higher-scale risks of AI, especially the role of computing daily life, fears about runaway systems and uncontrollable technology.

To make AI trustworthy, transparent, auditable and accountable, there needs to be collaboration between public policy, technology standards, and other institutional mechanisms related to AI safety. In the Safe and Responsible AI in Australia discussion paper published by the Australian Government in 2023, the government highlighted various voluntary and regulatory mechanisms that are being revised to address risks from AI, similar initiatives in other countries, as well as called for public consultation. The consultation received 510 public submissions from companies like Woolworths, Canva, and Microsoft, and the government has released an interim response. This effort shows how closely the government and industry are working together to understand and surface issues that need urgent attention (Australian Government - Department of Industry, Science and Resources, 2023).

Risks from AI can be broken down to 3 areas: technical, social and systematic/economics. Technical risks include validity, accuracy, security, privacy, bias and explainability, while social risks include human rights issues in high-stake contexts and acceleration of existing social inequalities (He et al., 2016). Systemic social and economic risks, on the other hand, includes risks to democratic systems, environmental impacts, transformation of work, and monopoly by a small number of transnational corporations providing generative AI as a platform or service.

From the perspective of the development of AI systems, we can also assess the risks in each stage of the development lifecycle: model pre-training, fine-tuning, input and output filtering, testing and deployment, and ongoing monitoring. At each stage, there are mechanisms to help manage potential risks, namely data manipulation, bias, having humans in the loop, documentation, transparency and explanation for the user, as well as ongoing trust and safety monitoring (Bell et al., 2023).

Based on these risks, a lot of companies and governments have come up with a risk-based framework for managing AI projects, which would classify projects into high-medium-low risk. Each level of risk recommends mechanisms to control the risk by mandating documentation, providing training data sources, monitoring, or the involvement of humans in the loop. Having a tiered risk approach helps direct focus to high-risk AI systems, and lets low-risk systems develop without much hindrance (Australian Government - Department of Industry, Science and Resources, 2023).

In this research, I am planning to look into the potential differences that the government, industry, and academia might have in their perspective of responsible AI development. The perspective of law enforcement and public welfare from the government may be very different from that of companies that are looking to innovate and create new valuable solutions, and may in turn be different from the perspective of academics looking into the meaning of safety and responsibility in relation to personal data, identity, and potentially sentient systems. I want to identify areas of tension and conflict, as well as areas where alignment is strong between these different parts of the ecosystem.

RESEARCH METHODOLOGY

This research will include a literature review to understand the most up to date perspective of responsible AI research. I will also utilize natural language processing techniques (NLP) to analyze discussion papers and statements released by the government, companies and research institutions about responsible AI. My area of focus is analyzing the Australian government's Safe and responsible AI discussion paper, as well as the 510 submissions from the public. I will use techniques like topic modeling and sentiment analysis to extract the areas of focus in each submission.

As the submissions have different formatting and structure, but still mostly attempt to address the 20 questions put forth by the government's discussion paper, I will need to extract the relevant text from each of the 510 files, group them together in a machine friendly dataframe, and work with the text from there to extract insights. For data cleaning and preparation, I will use the popular Python nltk library, and for topic modeling, I will use a State of the Art (SOTA) model, the BERTopic model.

As I read the literature, I will identify areas of potential tension, conflict or alignment between government, industry and research, and attempt to verify these themes through analyzing the Australian government's public consultation on responsible AI.

EXPECTED CONTRIBUTION

As AI continues to develop and get deployed, it is crucial that we understand the dynamics between key players in the ecosystem, and the potential points of conflict and alignment among them. The method I develop will help me analyze a large volume of text more quickly, and find patterns that may be missed by a human reader and analyst. As this area of research continues to grow and expand, the method can be utilized to help the government more quickly assess the submissions they receive from public consultations, or also help other industry and research institutions keep track of new ideas and developments in the area of responsible AI.

REFERENCES

Abedin, B. (2022). Managing the tension between opposing effects of explainability of artificial intelligence: A contingency theory perspective. *Internet Research.*, *32*(2), 425-453. https://doi.org/10.1145/3479645.3479709.

Australian Government - Attorney General's Department. (2022). *Privacy Act Review Report 2022*. Australian Government - Attorney General's Department.

Australian Government - Department of Industry, Science and Resources. (2023, June). Safe and responsible AI in Australia. https://consult.industry.gov.au/supporting-responsible-ai/submission/list

Bell, G., Burgess, J., Thomas, J., & Sadiq, S. (2023, March 24). Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs). *Australian Council of Learned Academies*.

He, C., Parra, D., & Verbert, K. (2016). Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, *56*, 9-27. https://doi.org/10.1016/j.eswa.2016.02.013.

Meske, C., Abedin, B., Klier, M., & Rabhi, F. (2022). Explainable and responsible artificial intelligence. *Electronic Markets*, *32*(1), 2103-2106. https://doi.org/10.1007/s12525-022-00607-2

Risks from Artificial Intelligence. (n.d.). The Centre for the Study of Existential Risk. Retrieved April 11, 2023, from https://www.cser.ac.uk/research/risks-from-artificial-intelligence/

UX for AI: The role of transparency. (2022, November 16). Levity AI. Retrieved April 11, 2023, from https://levity.ai/blog/ux-for-ai

Vossing, M., Kuhl, N., Lind, M., & Satzger, G. (2022). Designing Transparency for Effective Human-AI Collaboration. *Information Systems Frontiers*, *24*, 877–895. https://doi.org/10.1007/s10796-022-10284-3.