Generative Artificial Intelligence (GAI)
(Technical Topic)

An Analysis of Algorithmic Models and the Spread of Misinformation and Disinformation
(STS Topic)

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Henry Chen

November 8 2024

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Dr. Gerard J. Fitzgerald, Department of Engineering and Society

Word Count: 1795

Introduction

Over the last decade, Artificial Intelligence (AI) has grown to become one of the most prominent and fastest-growing fields in science. Its applications are integrated into many aspects of everyday life and have become a primary source of information for many (Lammertyn, 2024). However, this rapid expansion also brings challenges, particularly in the generation of false or misleading information.

My potential technical capstone would explore the impact of training data quality and design optimizations on the performance and reliability of GAI models, exploring how these factors influence the generation of accurate content. The purpose of this is to explore how to optimize components within a GAI model and enhance model precision.

The sociotechnical issue that is being addressed is GAI's role in the spread of misinformation and disinformation. The potential for GAI to create convincing but inaccurate content, such as fabricated news and deep fake images, poses a serious risk of spreading misinformation and disinformation. To understand the societal components that contribute to this system, Bruno Latour's Actor-Network Theory (ANT) will be used as a framework to examine the influence of various actors on the spread of misinformation and disinformation. This approach helps reveal how different human and nonhuman actors interact and shape the dissemination of false information and where best to address this issue (Elder-Vass, Atkinson, Delamont, Cernat, Sakshaug, & Williams, 2020).

Since the design and development of a GAI model directly influence the type of information it generates, there is an inherent connection between the technical aspects of creating GAI and the sociotechnical consequences of misinformation. The way models are trained and optimized influences not only their accuracy but also their potential to contribute to the spread of

false information, making it essential to consider both technical and societal components when addressing this issue. As the user base of GAIe grows by the millions, it is becoming increasingly important to address its potential to spread misinformation and ensure that technical and regulatory measures are in place to promote responsible and accurate content generation (Lammertyn, 2024).

Generative Artificial Intelligence

What GAI is capable of is all made possible using a variety of machine learning algorithms. Text generation often utilizes a model called Generative Pre-trained Transformers (GPT) which operates by processing input data in the form of tokens. These tokens are smaller chunks of language, such as words or subword units, that the model uses to understand and predict the next token in a sequence. By analyzing large datasets and learning relationships between tokens, GPT can generate coherent text by predicting the most likely next token based on the context of previous ones (AWS, n.d.).

Image generation relies on diffusion models, which transform noisy inputs into clear images through an iterative denoising process. These models are trained to understand the distribution of real images by learning to reverse the process of adding random noise. Starting with a noisy input, the model progressively reduces the noise in steps, refining the image with each iteration until a detailed and realistic image is generated. This gradual refinement process allows diffusion models to create images from random starting points including text input (AWS, n.d.).

What a GAI model is able to generate is highly dependent on the information it has access to. As Kothandapani often exclaims in her work: "garbage in, garbage out," meaning that

the outputs of the model will only be as good as the data it was trained and tested on (Kothandapani, 2024). A model's tendency to generate inaccurate information is often rooted in the quality of its training and testing data, where flawed, biased, or insufficient datasets can lead to unreliable results. This is particularly problematic in generative AI systems, which learn patterns and relationships from the data and replicate these patterns during generation (Kothandapani, 2024).

Another important technical consideration is that developers have begun to design systems that intentionally avoid producing sensitive or controversial material. The goal of this is to prevent the spread of what is deemed harmful information. As a result, developers have embedded filters in AI systems to block such content (Ryan-Mosley, 2023).

To explore how different data sets and design decisions impact the accuracy of a given AI model, it is important to consider the different types of data sources that can be used. Recent data reveals that Grok, a GAI developed by xAI and trained primarily on text data scraped from Twitter, performs two percent worse in arithmetic tasks but nearly three percent better in being perceived as human, compared to its counterpart ChatGPT, which is trained on a broader range of publicly available text sources (Thaler, 2024). To understand what about the data itself influences these differences in performance, it is crucial to analyze the nature and diversity of the datasets. It is also important to consider how data is being pre-processed before it is used in training an AI model. By explicitly making different decisions regarding noise, biases, and inconsistent data, models will vary in performance (Aparicio, 2024).

The hope of this study is to identify how different design decisions, including dataset selection and preprocessing techniques, impact the performance of AI models. This will serve as

the groundwork for developing best practices in AI model design, enabling the creation of more accurate and reliable models.

An Analysis of Algorithmic Models and the Spread of Misinformation and Disinformation

In the "algorithmic age," large digital platforms connect billions of people, allowing misinformation to reach large audiences quickly. Misinformation has the potential to negatively influence decision-making and public opinion. At a large scale, this could lead to mass confusion at an uncontrollable scale (Omoregie, & Ryall, 2023). With ChatGPT alone receiving over 10,000,000 queries daily (Lammertyn, 2024), the potential for large GAI platforms to contribute to the spread of misinformation and disinformation is significant.

Using ANT, we can analyze this complex issue by examining the relationships between key actors within this network (Elder-Vass, Atkinson, Delamont, Cernat, Sakshaug, & Williams, 2020). One of the key actors in this network are the developers of the GAI models. They are responsible for designing the algorithms and selecting training data such that the models they build yield the most accurate information (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). One way to hold developers to this standard is to evaluate the content produced using the "Wittgensteinian" model for written misinformation analysis. This model is based on the fact that meaning of words isn't fixed but depends on how they're used within specific contexts and helps analyze misinformation by examining how language is used to construct and reinforce certain beliefs (Omoregie, & Ryall, 2023).

It is also a developer's choice to impose a content filtration system based on his/her own ethical guidelines and societal standards (Huang, Zhang, Mao, & Yao, 2023). ChatGPT, a GAI model developed by OpenAI, has been programmed to avoid providing responses on certain

topics such as Donald Trump or weapon-making. This decision reflects the developers' attempt to prevent harm or controversy by limiting the model's ability to generate reponses to potentially dangerous queries. However, these filters are an outcome of the developers' values and judgments about what content should be restricted. Through investigation, it was discovered that ChatGPT has a clear left leaning political bias in its responses regarding its polotical position (West, Engler, Friedler, Turner Lee, & Turner Lee, 2024). This poses serious ethical issues surrounding biased content and censorship.

The government and regulatory bodies represent another significant actor in this network. Policymakers are tasked with creating regulations that can prevent the dissemination of false information without infringing on free speech (Sarangi, & Sharma, 2019). Balancing regulation with the need to foster technological development is a delicate task. The European Union has taken a large step toward regulating artificial intelligence with the proposed Artificial Intelligence Act (AIA). This act provides a framework for monitoring and controlling GAI applications and would subject them to strict requirements for transparency, accountability, and data quality. The AIA mandates that companies provide clear documentation of the data sources used in training these models, aiming to slow the spread of misinformation by ensuring models are built on unbiased data ("High-Level Summary of the AI Act," n.d.).

In contrast, the U.S. approach on AI remains decentralized, relying on executive orders and industry-specific regulations rather than a single law. One reason for this difference is because of the U.S. Constitution and its emphasis on free speech and individual freedom protected by the First Amendment.

While not enough time has passed to evaluate which approach is more effective, by regulating how GAI platforms operate and enforcing transparency in the training and functioning

of these models, governments can act as a check on the influence of GAI in spreading disinformation (Kalpokienė, 2024).

Consumers, or users of these platforms, play another vital role in the network. How users interact with AI systems and interpret their reliability are crucial factors in the spread of misinformation. In many cases, users may not recognize AI-generated content as misleading (Shin, 2023). Additionally, over-reliance on AI technology can lead users to trust AI without questioning their accuracy. When users assume AI systems are objective, they are likely to accept and share information without critical analysis (Cheng, 2022). Enhancing media literacy among users can help mitigate the spread of misinformation, making them more critical and informed consumers of AI-generated information.

The spread of misinformation through Generative AI presents a complex sociotechnical problem, with no single solution capable of addressing all its facets. A good solution will involve analyzing how current developer choices and existing government regulations have impacted the GAI system and selecting the choices that yielded the best results. This will take the form of a combination of central or decentral regulatory policies, ethical design frameworks for developers, and campaigns to improve media literacy. By analyzing the actor network surrounding GAI and misinformation, it becomes possible to identify potential issues within the system. However, because of the interconnected nature of these actors, potential solutions require a collaborative and multifaceted approach.

Conclusion

Although AI is a powerful tool, it also presents significant challenges related to the spread of false information, which can manifest in various ways. The technical analysis of how

data quality and design decisions affect model performance is critical in obtaining a new understanding of how GAI models can be optimized to produce more accurate and reliable results. This will be integrated into the broader socio-technical analysis of the GAI network which will explore ethical and technical concerns for each key actor. The careful analysis of this network is crucial for identifying key points where interventions can help mitigate the spread of misinformation. Ultimately, the goal is to develop a framework for creating reliable and ethical GAI systems, allowing society to fully harness the potential of AI while minimizing its negative impacts on information integrity.

References

Sarangi, S., & Sharma, P. (2019). *Artificial Intelligence: Evolution, Ethics and Public Policy*. Abingdon, Oxon: Routledge.

Omoregie, U., & Ryall, K. (2023). *Misinformation Matters: Online Content and Quality Analysis*. Boca Raton, FL: CRC Press.

Elder-Vass, D., Atkinson, P., Delamont, S., Cernat, A., Sakshaug, J. W., & Williams, R. A. (2020). *Actor-network theory*. SAGE Publications Ltd.

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An Overview of Artificial Intelligence Ethics. *IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE*, Vol. 4. (4th ed.), 799–819

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, Vol. 3. (2nd ed.)

Shin, D. D. (2023). *Algorithms, humans, and interactions : how do algorithms interact with people? designing meaningful AI experiences* (First edition). Routledge.

Kalpokienė, J. (2024). *Law, human creativity and generative artificial intelligence : regulatory options*. Routledge.

Lammertyn, M. (2024, September 30). *60+ facts about CHATGPT you need to know in 2025*. InvGate. https://blog.invgate.com/chatgpt-statistics#:~:text=ChatGPT%20receives%20more%20than%2010,hit%20100%20million%20weekly%20users

Kothandapani, V. (2024, March 8). *Addressing bias in Generative AI starts with training data explainability*. RWS. https://www.rws.com/artificial-intelligence/train-ai-data-services/blog/address-bias-with-generative-ai-data-explainability/

AWS. (n.d.). *What is GPT AI?* . AWS.Amazon. https://aws.amazon.com/what-is/gpt/

AWS. (n.d.). *What is Stable Diffusion?* . AWS.Amazon. https://aws.amazon.com/what-is/stable-diffusion/

Thaler, S. (2024, March 29). *Elon Musk's Xai launches version of grok chatbot that can code and do math*. New York Post. https://nypost.com/2024/03/29/business/elon-musks-xai-launching-new-version-of-grok-chatbot/

Aparicio, M. (2024, June 27). *How does GPT data training work?*. Capicua. https://www.wearecapicua.com/blog/gpt-data-training

Ryan-Mosley, T. (2023, October 3). *How generative AI is boosting the spread of disinformation and Propaganda*. MIT Technology Review. https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/

West, D. M., Engler, A., Sorelle Friedler, S. V., Nicol Turner Lee, D. V., & Nicol Turner Lee, D. M. W. (2024, October 18). *The politics of AI: Chatgpt and political bias*. Brookings. https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/

High-level summary of the AI Act. EU Artificial Intelligence Act. (n.d.). https://artificialintelligenceact.eu/high-level-summary/

Cheng, L. (2022). *The trust factor: Understanding consumer reliance on AI in information assessment*. Journal of Media and Technology Studies, 15(3), 150-169.