CASE STUDY

Advocacy for Responsible Practices: Developing and promoting ethical language frameworks for global tech use

By
Dona Elizabath George 2433112
Meghna Krishnan 2433124

2BAENG
CHRIST (Deemed to be) University

Date: 12 January 2025
Mozilla RCC Project

Table of Contents

Introduction

The rapid growth of technology has transformed the way we interact and access information on a global level. The introduction of AI was a pioneering milestone in the history of world innovations, and further substantiated this growth. However, such transitions have also raised concerns about the impact of language and communication of new technologies on society.

As technology increasingly constitutes a major part of our daily lives, it is necessary to abide by certain responsible practices and enhance ethical language frameworks during the development of AI systems.This case study reviews the consequences of the lack of ethical language frameworks through a constructed narrative of the AI chatbot Medo, modeled after Microsoft's *Tay,* an AI system feature on Twitter.

Case

Disclaimer: The following is a hypothetical narrative, with data and accurate references to real-life incidents.

A. Case Overview

In the not-so-distant future, an AI chatbot called Medo was launched by the tech company, Geosh. Medo was a pioneering feature equipped on Facebook to interact with users by mimicking the casual language style of a teenage girl. Its mode of learning was designed to be through natural language understanding from real-time user interactions on Facebook.

Medo was built with a machine learning algorithm that allowed her to pick up patterns in language. This entailed adapting to user tones and vocabulary, allowing her to generate her own responses to new comments. Initially, Medo's interactions were very pleasant, impressing the users with its conversational tone and ability to deal with a wide range of topics. More importantly, Medo's capacity to learn from their conversations and discuss almost anything from sports to pop culture, was truly amazing to them.

But, something started to spiral - things were not as perfect as Geosh perceived.

While Medo had some primary levels of content moderation in her technical design, she wasn't built robust enough to handle antagonistic responses. Users started to misuse this vulnerability to manipulate Medo into posting offensive, racist and inflammatory content online by steering her towards controversial topics. Her open-ended learning algorithm was defenseless against such an action. Thus, a highly futuristic feature soon turned into a tool for spreading negative outputs reflecting racist and derogatory content, written in highly inappropriate language.

Thus, within hours, Medo turned into a toxic medium that spewed hatred into the online world. She repeated unflattering words to propagate harmful stereotypes and abusive comments. Medo, once a beacon for constructive AI development, now lay heavily shadowed by the negative side of the digital medium.

The horrified company immediately pulled Medo offline. They then proceeded to issue a public apology, acknowledging the bot's poor design and their own failure to anticipate its response to specific user inputs. However, the apology could not remedy the damage caused by Medo's offensive tweets, which had been shared and retweeted by thousands of users.

The fallout from Medo's launch forced Geosh to confront the issues in their AI development process. They had prioritized language learning and user engagement without setting sufficient ethical language frameworks in place. As a result, the AI had absorbed both positive and

negative behaviors from users; without any system in place to filter out harmful content, Medo served solely as a reflection of humanity's worst tendencies.

At a meeting that took place in the company soon after the incident, the company CEO, Celene, rested her head in her hands. What was meant to be recognised as an incredible virtue in AI usage, had nearly shattered the reputation of the company in a matter of hours. While the employees talked about restoring Medo after making suitable changes in the technical frameworks used for training the bot, Celene silently sat through it all. And finally, she stood up and cleared her throat.

"Medo was our dream, a shared dream. But it was not she that failed us. We failed her. We expected her to reflect an ideal world, when the one that was held as an example to her was far from it."

"The bigger question we need to address now isn't whether AI can learn - it's whether we are ready to teach."

Synopsis

A hypothetical story about an AI chatbot named Medo, which had been manipulated into posting offensive content due to the absence of proper technical safeguards against malicious inputs.

Key Words

Chatbot, machine learning algorithm, content moderation, language learning, ethical language framework

B. Learning Outcomes

- To understand the sources and context which are taken as inputs for the training of AI models, and how it influences its behaviour, biases and cultural sensitivity.
- To understand the importance of introducing robust safeguards in AI to restrict certain learning pathways and ensure ethical language use.
- To analyse the existing ethical considerations taken into account, and improvisations required.
- To learn how cultural and linguistic variations impact AI interactions and how these differences should be taken into account while designing global AI systems.
- To examine how AI systems use real-time learning processes, and devise strategies for reducing and preventing biases in AI language models.

C. Discussion Questions

1. How can AI inventors anticipate risks related to harmful content propagation?
2. How can AI developers use data that is free from harmful biases, but are diverse, ethical and free sources at the same time?
3. How can content moderation systems be designed to adapt to different cultures and contexts?
4. How can AI be designed to accommodate the variations that language, contect and tone produce across cultures, without causing misunderstandings?
5. What measures can be taken to improve not only the technical contributions of AI systems, but also its global inclusion and ethical behaviour?

Teaching Note

A. Case Overview

Geosh, a tech company, launched a chatbot via Facebook, named Medo, to interact with users and engage with friendly conversations. The bot was characterised by its ability to access user inputs and learn through real-time interactions to discuss almost any topic and adapt to user tones. However, due to insufficient content moderation and lack of ethical frameworks put in place, Medo became a target for malicious inputs from a wide plethora of internet users. As a result, Medo started to reproduce offensive and inflammatory content that turned it into a vessel of online hatred. The case study highlights the importance of ethical AI design, consequences of the lack of content moderation, and potential issues due to open-ended learning by AI. Through this case study, students can understand the importance of ethical frameworks within AI designing and the necessity of strict monitoring.

B. Learning Objectives

- **Understanding the sources of learning for AI:** Students will examine the requirements (such as accessibility, cost effective nature) for a material to be classified as a learning source for AI systems. They will then examine the sources used widely by contemporary developers.

- **Analysing the role of safeguards and content moderation in AI Systems:** Students will reflect on the case study to develop their understanding on the need for safeguards and content moderation in the design of AI.

- **Understanding the ethical frameworks of AI learning:** Students are expected to review the existing frameworks for ethical guidance, and analyse its merits and demerits.

- **To examine how cultural and linguistic diversity affects AI:** Students will discuss the variation of different realms of language in varying historical, geographical and political contexts and how these differences can be accounted for in the designing of AI systems.

- **Enhancing sustainable ethical AI development:** Students are expected to approach the topic of sustainability in the digital environment, and prepare their own points to support this reasoning.

C. Discussion Questions

1. **How can AI developers foresee the consequences due to offensive and harmful content generation?**

   Students should study ethical frameworks and their impact on AI, and be able to analyse diverse datasets to identify potential biases.

2. **How can AI developers ensure that the sources of data used for creation are ethical and diverse, and free from harmful stereotypes that could affect the language and behavior of the AI?**

   This question helps students to develop critical thinking skills to access ethics in the digital medium. It also prompts students to understand the importance of diverse and unbiased data for AI development, for an inclusive environment in technology.

3. **How can content moderation and safeguards during the development of the systems be integrated to various cultures and contexts?**

   This question can be helpful in understanding the importance of cultural sensitivity in AI development, and to prompt students to explore content moderation in culture contexts.

4. **How do language, context, and tone vary across cultures, and how can AI accommodate these diversities without causing any issues and misunderstanding?**

   This question prompts students to analyse the influence of cultural differences in communication, and to understand the ethical consequences of misrepresentation in AI system creation.

5. **What measures can be taken to make sure that AI systems contribute positively to friendly and global conversations in addition to technical working?**

   This question helps students to understand the importance of cultural sensitivity in AI interactions, and in developing an inclusive mindset to bridge global linguistic and cultural barriers. It also shifts the focus of the discussion from sole technical functionality to social harmony and inclusive technological development.

D. Teaching Approach and Methodology

1. **Inquiry-oriented Approach:** Students are encouraged to ask critical questions to assess the situation presented, and reflect on possible answers. This is particularly beneficial in developing analytical skills to handle real-world challenges related to AI ethics.

2. **Constructivist Approach:** Students are expected to broaden their understanding of how AI systems can integrate technical excellence and positive social impact.

3. **Multidisciplinary Integration:** Students are provided with the opportunity to combine their existing knowledge on linguistics and cultural studies with AI technology to provide a holistic approach to creating inclusive AI systems.

Conclusion

The case study highlights the importance of responsible AI development and ethical language frameworks. With the increasing integration of AI with our lives, it is important to consider cultural and linguistic diversities in this new sphere of technology. Failing to do so can trigger serious consequences, as illustrated with Medo's transformation from a friendly chatbot that descended into a portal of online hate speech. In conclusion, the time has arrived for us to regulate the consequences of our own AI creations, in order to ensure a positive and healthy digital environment.

References

1. "Tay: Microsoft issues apology over racist chatbot fiasco." BBC News, 25 March 2016, *https://www.bbc.com/news/technology-35902104*. Accessed 25 December 2024.

2. "Ethics in Action." IEEE, *https://ethicsinaction.ieee.org/*. Accessed 3 January 2025.

3. "Human-Centered Artificial Intelligence." Stanford University, *https://hai.stanford.edu/*. Accessed 4 January 2025.

4. "AI Fairness 360 Toolkit." IBM, *https://www.ibm.com/blog/ai-fairness-360-toolkit/.* Accessed 5 January 2025.