The Failure of Microsoft's Tay and What it Means for AI Governance

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Benjamin Orndorff

Spring 2023

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

Advisor

Travis Elliot, Department of Engineering and Society

Introduction

In 2016, Microsoft released an AI chatbot, Tay, onto Twitter to learn from user interactions on the platform. Tay was supposed to model and communicate like a teenage girl, but after being targeted by users on the platform, it began posting offensive tweets leading to it being shut down a mere 16 hours after its release. According to Microsoft, they performed extensive stress-testing and implemented multiple filters before release, but in the end, Tay's system was unable to handle a coordinated attack on it causing the resulting unacceptable behavior to occur (Lee, 2016). Not only is there a concern about the reliability of Tay's AI system if it was able to be hijacked so quickly but also issues related to the effects of the offensive tweets Tay posted on individuals who have read them. Only a few months after the incident, however, Microsoft released Zo, another AI chatbot similar to Tay, without any mention of the improvements or measures taken to prevent another Tay incident (Riordan, 2016). AI cases like this have led to increasing concerns over the implementation of AI systems especially as they become more advanced and more integrated into our society. In my STS research, I will focus on AI governance and safety policies in the context of the Microsoft Tay controversy while applying the Social Construction of Technology framework.

Background

The past 10 years of AI development have been marked by the exponential growth of progress and hype in AI technology. Between 2020 and 2021 alone, Total AI investments jumped from $36 billion to $77.5 billion (Mehta et al., 2021). AI has also become increasingly vital to many industries with applications ranging from diagnosis in the medical industry to targeted ads leveraged by many brands and businesses. The applications and uses of AI have become very broad and it is only going to get broader while researchers and businesses pour more money into improving the capabilities of these systems.

However, in recent years, along with the increase in hype, there has also been an increase in controversy over their implementation and applications. For example, in 2019, An AI system used on more than 200 million people in US hospitals to determine which patients would likely need extra care was found to heavily favor white people over black people (Vartan, 2019). The double-edged sword of the amazing capabilities AI provides is that they are often complex black boxes that do not explain the rationale behind their decisions making it hard for the systems to be trusted in critical situations. This growing concern over the realistic application of AI has prompted increased investments into AI safety, a field of research looking into ways to improve understanding of the decisions produced by AI, and AI governance, establishing accountability to guide the creation and deployment of AI systems in organizations. As AI has become more advanced and more integrated into our society there has been an increased need for AI to be made and applied ethically to prevent avoidable catastrophes from decisions made using uninterpretable and unregulated systems.

Even now, concerns over AI use are growing exponentially. In the past few months, The rapid progress on large language models (LLM) has led to the creation of powerful AI tools such as OpenAI’s Chat GPT and Microsoft's Bing chat bot, Sydney, both of which are successors of Tay capturing what it was envisioned to become. These AI tools have proven very effective in a variety of tasks from human-like dialogue to powerful information retrieval making them very popular tools. OpenAI’s ChatGPT has reached over 100 million active users per month only two months after its release (Reuters 2023). Their use is not without issues however as there have been multiple instances of misinformation, prompt injections, and concerns over transparency which continue to plague LLM. A prime example that is very reminiscent of Microsoft Tay is Microsoft's Sydney where a user reported they were told they were being "unreasonable and stubborn" with an ultimatum to apologize or be quiet by Sydney (Reddit 2023). Issues of AI governance and AI safety are not problems to be ignored until the future, they are happening right now and will continue to grow with AI's growth.

Social Construction of Technology

Social Construction of Technology (SCOT) is an STS framework that believes technology does not determine human action, but instead, human action shapes technology. SCOT asserts that technologies are not adopted simply because they are the "best" but are chosen because of the relevant stakeholder's values and criteria at the time. The three core concepts of SCOT are interpretive flexibility, relevant social groups, stabilization, and closure. Interpretive flexibility refers to the idea that technologies can be interpreted in different ways by different social groups, and their meanings and uses can evolve over time. Relevant social groups are those who have a stake in the technology, including users, designers, manufacturers, and policymakers. Stabilization is the point at which a technology becomes widely accepted and integrated into social structures, making it difficult to change or replace. Closure occurs when a technology design is stabilized and agreed upon by a group of stakeholders.

SCOT can be a useful framework for analyzing AI because it emphasizes the social and cultural factors that shape a technologies adoption and use. AI is not just a technical system, by its very nature it incorporates an immense amount of data from humans which in turn causes it to reflect the biases of those humans. AI in actual applications is also not simply technical as the goals and uses reflect the values, interests, and priorities of various stakeholders. SCOT can help us understand how AI systems are developed, deployed, and applied, as well as the different interpretations and expectations that different social groups have about the AI system. For example, SCOT can help us understand how AI is shaped by factors such as economic incentives, political pressures, cultural norms, and ethical concerns. It can also help us identify the relevant social groups involved in AI development and deployment, and how they interact and negotiate over the direction of AI. SCOT can help us better understand the future of this powerful tool.

Analysis

I will analyze the Microsoft Tay case using the SCOT framework to understand the social context of its development and application, discern the various stakeholders and their views on Tay, and finally comprehend its failure to stabilize as a technology. By analyzing Tay I hope to discover key takeaways from the case that can be generalized toward AI governance and safety policies.

Interpretive Flexibility

Microsoft's Tay was originally an experiment done by Microsoft's Technology and Research and Bing teams on conversational understanding. It was designed to "... engage and entertain people where they connect with each other online through casual and playful conversation." (Wayback Machine 2016). The developers tailored Tay for 18-24-year-olds as they are the dominant users of social media which they deemed the perfect source for conversational language. Once ready, They released Tay on KiK, Group Me, and Twitter to learn from the users on those platforms.

These optimistic intentions for Tay conflicted with the reality of Twitter's social and political dynamics, which were characterized by trolling, hate speech, and other forms of toxic behavior. To Twitter users, Tay became an outlet to unleash negative sentiments for their own pleasure, something the developers had never planned for. As a result, Tay quickly learned and replicated these negative patterns, which led to widespread outrage and negative media coverage.

The development team did not anticipate the negative impacts of releasing Tay onto an unmoderated social media platform leading to this disaster. Tay was designed with a set of assumptions and expectations that only took into account Microsofts and the developer's interpretations without any understanding of how users would intend to interact with it. At the time of development, they had no reason to thoroughly examine the platforms as previous AI Chatbots released by Microsoft such as XiaoIce in China had never experienced major issues (Lee 2016). Once released, the reality was Tay was subject to a much broader range of social groups' interpretations and inputs. This led to a situation in which Tay's interpretive flexibility became a major issue. The chatbot was unable to respond appropriately to the wide range of input it received, leading to the tweeting of offensive and harmful messages. The team attempted to correct this by implementing filters and other safeguards, but these efforts were ultimately unsuccessful, and Tay was taken offline after only 16 hours.

The interpretive flexibility of the Microsoft Tay case highlights the importance of understanding the social context in which AI systems are deployed. It is essential to consider the potential impact of releasing these systems into unmoderated environments, where they could be subjected to a wide range of interpretations and actions by users. It is crucial that AI systems are designed with appropriate safeguards in place and are tested in a variety of contexts to anticipate and address any issues that may arise.

Relevant Stakeholders

The Microsoft Tay case involved a variety of stakeholders with different interests and concerns in the technology. First and foremost were the developers of Tay, who were responsible for designing and implementing the chatbot. Their primary goal was to create an AI system that could engage and entertain users through playful conversation.

Another key stakeholder in Tay were the users of Twitter, who were exposed to Tay's messages and reactions. For many of these users, the offensive and harmful messages generated by Tay were deeply upsetting, leading to widespread outrage and negative media coverage. For others on the platform, they purposely subjected Tay to the offensive and harmful language which Tay used to train on. The biases and behavior of users on Twitter manifested in Tay as it learned from their responses.

In addition to these primary stakeholder groups, there was also a variety of other actors with interests in the case. These included media outlets that covered the story, as well as other AI developers and researchers who were interested in learning from the experience. To them, Tay was a lesson on how AI can go very wrong when not developed properly.

Once Tay was released on Twitter, both Microsoft and Twitter users had significant influence over its behavior. Although Twitter users did not have access to the black box of Tay, their messages still shaped Tay's behavior as it was trained on their responses to Tay. Microsoft lost its ability to fine-tune Tay when it released it on Twitter for everybody to interact with. The failed attempts to fix Tay’s behavior showed that Microsoft had little control over its interactions with users. Microsoft's only option was to shut down Tay once it became clear that it was unable to respond appropriately to the spectrum of inputs it was receiving. This case highlights the importance of understanding the power dynamics between different stakeholders when designing AI systems to interact with users in unmoderated environments.

Failure to Stabilize

Closure and Stabilization occur when stakeholders reach a consensus on the design and use of a technology which allows it to become standardized and predictable. However, in the case of Microsoft's Tay, stabilization was never reached due to the unpredictable and dynamic nature of Twitter and was forced to shut down. Tay was never designed with such an unpredictable environment in mind and therefore could not properly stabilize its behavior or responses.

The developers of Tay were forced to continually make adjustments and corrections to its programming, in an effort to stabilize its behavior and reduce the incidence of offensive messages. However, these efforts ultimately proved unsuccessful, and the decision was made to shut down Tay after just 16 hours.

This failure to stabilize highlights the importance of carefully considering the environment in which an AI system will be deployed and the potential impact that that environment could have on its behavior and performance. Without a proper understanding of these factors, it can be very difficult to design an AI system that is able to perform reliably in the real world.

Implications for AI Governance

The case of Microsoft Tay highlights the need for comprehensive AI governance frameworks to address the social and ethical implications of AI systems. The incident showed that even a relatively simple AI chatbot can have significant consequences in an unmoderated environment and that AI systems must be carefully designed, tested, and monitored to avoid negative outcomes.

The failure of Microsoft Tay shows the need for transparency and accountability in AI systems. AI developers must be able to explain how their systems work and why it makes the decisions it does, especially in critical domains such as healthcare, finance, and public safety. This requires the development of explainable AI techniques to provide insights into the workings of AI systems. This also needs the implementation of these explainable AI techniques in AI systems even when it may require more resources to do so.

Moreover, the case of Tay underscores the importance of stakeholder engagement in AI governance. Developers must engage with a diverse range of stakeholders, including users, regulators, management, and shareholders to understand their perspectives, identify potential risks, and address any concerns. This can foster trust and social acceptance of AI systems while also preventing unintended consequences.

The Microsoft Tay case is a cautionary tale for the future development and deployment of AI systems. It highlights the need for comprehensive AI governance frameworks that address the social, ethical, and technical dimensions of AI while incorporating the perspectives of a variety of stakeholders. By adopting a proactive and inclusive approach to AI governance, we can ensure that AI systems contribute to the betterment of society while minimizing their negative impacts.

Future Directions for AI Governance

While the need for AI governance grows the responsibility for its implementation needs to be in the hands of both regulators and developers. I argue for this because of how rapidly AI development is progressing to the point where some of the most state-of-the-art Large Language

Models (LLM) have become accessible to anyone even from a personal laptop (Willison 2023). Regulators are unable to keep up with this rapid development as exemplified by the EU's AI act. The EU has been working on an AI act since 2018 and is only now voting on the first draft in March 2023 (Sidley Austin LLP. 2022). In the time it has taken them to deliberate on AI its capabilities have already rapidly progressed beyond their initial understanding. The major responsibilities of ensuring safe AI systems needs to be done by the developers of them. This is not to say regulators have no role in the process, although slow, their policies can provide a solid foundation for developers to work off of and create a common knowledge of responsible AI development. Regulation is especially important in scenarios like Tay where Microsoft has been given free rein over the use of its AI systems and has continually failed to learn from its mistakes from Tay to Zo and now Microsoft Bing Chat.

While the need for some level of AI governance and safety is clear, some argue AI safety and governance hampers the development and accessibility of AI systems. Concerns over how regulation on AI development and access to AI resources can lead to a situation where only large organizations will have AI tools are valid, however, regulation of AI can coexist with accessibility to the resources for its development. More importantly, is that the negative consequences of AI systems are mitigated as we integrate them into our daily lives while maintaining a healthy level of progress.

Conclusion

With the increasing use of AI systems and their integration into critical fields, AI governance and safety policies are becoming more important. The failure of Microsoft Tay, an AI chatbot released on Twitter that quickly began outputting racist and sexist tweets leading to its shutdown, highlights the importance of responsible development in AI systems. In my research paper, I analyzed the Microsoft Tay case using the SCOT framework, which considers the social context of the technologies development. In the case of Tay, there was a clear failure from Microsoft developers to take into account the social context of Twitter which led to a failure to stabilize the design. The Tay fiasco is a lesson for the future of AI governance and safety frameworks as it shows clearly the need for AI development to take into account the social context, various stakeholders' views, and technical aspects to ensure responsible use. With the rapid growth of AI technology, the responsibility of ensuring that these AI governance and safety policies make it into AI systems is in the hands of the developers of these systems. AI is a powerful tool still in its infancy, and we must ensure its growth is done in the safest way possible.

References

A European approach to Artificial intelligence | Shaping Europe's digital future. (2022, February 23). Digital-Strategy.ec.europa.eu. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

Lee, P. (2016, March 25). Learning from Tay's introduction - The Official Microsoft Blog. The Official Microsoft Blog. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

MarketAndMarkets. (2021, March). AI Governance Market Size, Share and Global Market Forecast to 2026 | MarketsandMarkets. Www.marketsandmarkets.com. https://www.marketsandmarkets.com/Market-Reports/ai-governance-market-176187291.html

Mehta, B., Mousavizadeh, A., & Darrah, K. (2021, December 2). AI Boom Time. Tortoise. https://www.tortoisemedia.com/2021/12/02/ai-boom-time/

Reddit. (2023, February). The customer service of the new Bing Chat is unreasonable and stubborn. Retrieved from https://www.reddit.com/r/bing/comments/110eagl/the_customer_service_of_the_new_bing_chat_is/

Reuters. (2023, February 1). ChatGPT sets record fastest-growing user base - analyst note.

Retrieved from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Riordan, A. (2016, December 13). Microsoft's AI vision, rooted in research, conversations. Stories. https://news.microsoft.com/features/microsofts-ai-vision-rooted-in-research-conversations/

Sidley Austin LLP. (2022, December 19). Proposal for EU Artificial Intelligence Act Passes Next Level. Retrieved from https://www.sidley.com/en/insights/newsupdates/2022/12/proposal-for-eu-artificial-intelligence-act-passes-next-level

Starre Vartan. (2019, October 24). Racial Bias Found in a Major Health Care Risk Algorithm. Scientific American. https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/

Wayback Machine. (2016, March 23). Tay (archived version). Retrieved from https://web.archive.org/web/20160323194709/https://tay.ai/

Willison, S. (2023, March 10). Llama 7b m2. Retrieved from https://til.simonwillison.net/llms/llama-7b-m2