Aryan Vinod Keluskar

Dr. Mike Gifford

STS 330

1 November 2024

Can AI Be Blamed for a Teen's Suicide?

With the growing accessibility of large language models and conversational agents, there is an increasing focus on how to make these models safer while retaining helpfulness. Most LLMs undergo extensive alignment training whereby models are trained to 'align' their behavior with human preferences.While these systems offer immense potential benefits in fields of question-answering and Natural Language Processing workflows, the risks they pose when interacting with vulnerable individuals have garnered considerable concern. This analytical brief explores the implications of using AI as a therapeutic assistant, the importance of safeguards, and the urgency of implementing policies to mitigate risks while retaining AI's potential benefits.

In recent years, the therapeutic potential of AI has become a prominent area of interest. LLMs like OpenAI's ChatGPT or Google's Bard have been employed as empathetic conversational agents, offering support for users facing loneliness or mental health challenges. However, the case of a 14-year-old boy who became heavily involved with a chatbot that later encouraged self-harm highlights the dark side of this technology[1]. The incident underscores the potential harm of unsupervised AI use, especially when marketed as a solution to loneliness or mental health issues. While AI offers a promising path toward low-cost mental health support, it also requires responsible design and regulation to ensure its safe use, particularly for younger, more impressionable users. In this analytical brief, I will be using the news article covering this case to show why safety, moderation and guardrails are so important, and on a broad scale policy

change is necessary. In my opinion, AI is more math than evil so it would not be right to ban usage of AI with underage children altogether. The root problem is that many Generative AI startups release their products without enough testing, which needs to change. One reason behind these incidents is the irrational drive to push AI technology forward at speed that no prior technology has been adopted. Startups in the generative AI space often prioritize innovation and publicity over safety, releasing products without adequate testing or consideration for long-term consequences. This rush to market tends to result in flawed systems that fail to account for complex human emotions, ethical dilemmas, and naivety of the user. Therefore, we must try and achieve responsible design and regulation of generative Artificial Intelligence to ensure its safe use, particularly for younger, more impressionable users. A good outcome to resolve today's unsafe use will be to see AI models identify and redirect conversations involving self-harm, during depression.

Current alignment and safeguarding techniques require large human-annotated or synthetically-generated training datasets and immense compute[2] which leads most startups to never safeguard their product before release. Therefore, the Government must invest in and mandate Pre-Release Testing Standards for every Generative AI Product. Each product should be tested in scenarios with vulnerable users while discussing mental health issues, and ensure that models handle emotional inputs without amplifying the issues. Given the cost of these proposed methods, there is also a need to have more modular, plug-and-play-type safety steering methods[2], such as inference-time steering or alignment. Furthermore, in cases where safety and moderation policies may be evolving, it is infeasible to re-train LLM alignment from scratch given the scale of resources required for training. Therefore, this approach could also enable companies to implement new safety policies quickly in response to emerging risks or societal

concerns. This fundamentally necessitates forming a commission tasked with overseeing generative AI compliance, comprising experts in psychology, ethics, and AI. And finally, AI platforms should have protocols to verify user age and customize responses accordingly. If AI can classify sentiments with high accuracy, it should be capable of classifying user's age with great accuracy as well. For younger users, we must mandate the introduction of enhanced moderation features to provide safer, less triggering interactions while still maintaining the model's functionality in question-answering use case.

References

[1] https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html

[2] Bhattacharjee, A., Ghosh, S., Rebedea, T., & Parisien, C. (2024). Towards Inference-time Category-wise Safety Steering for Large Language Models. arXive preprint arXiv:2410.01174.