GRADIENT INSTITUTE

Gradient Institute Ltd.
Level 2 Merewether Building H04
Cnr City Rd & Butlin Ave
The University of Sydney NSW 2006
https://gradientinstitute.org

28 February 2023

Senate Standing Committees on Economics
PO Box 6100
Parliament House
Canberra ACT 2600

Dear Sir/Madam,

Response to the inquiry into the influence of international digital platforms

Gradient Institute welcomes this inquiry at a time of great concern about how international digital platforms (commonly referred to as 'Big Tech' companies) will make use of the rapidly growing capabilities of their Artificial Intelligence (AI) technologies.

About Gradient Institute

Gradient Institute is an independent, nonprofit research institute and registered charity that works to build ethics, accountability and transparency into AI systems.

Scope of Gradient Institute's response to the inquiry

This submission addresses term of reference (c), as it relates to concerns about manipulation of individuals and its societal consequences (emphasis added):

*c. whether algorithms used by such international digital platforms lack transparency, **manipulate users** and user responses, and contribute to greater concentrations of market power and how regulating this behaviour could lead to better outcomes in the public interest;*

Executive Summary

- The new technology of *generative AI* enables Big Tech companies to deploy novel and powerful forms of user manipulation.
- Generative AI can be used to make chatbots that can conduct coherent, natural-sounding and human-like conversations, encouraging users to form emotional bonds which could then be exploited for commercial or political ends.
- The expertise and resource requirements of generative AI mean its development is dominated by Big Tech companies, who are in competition to deploy and monetise it as quickly as possible.
- The risk of harmful manipulation from generative AI chatbots and the potential scale of that harm creates an urgent need for regulating the use of this technology.

What is generative AI?

Generative AI refers to a type of AI that is capable of creating new data or content, such as images,[1] music,[2] text,[3] or videos[4]. ChatGPT and Microsoft Bing Chat[5] are examples of chatbots powered by generative AIs for text. Whilst chatbots are not themselves new technology, the new generation based on generative AI are far more capable: able to sustain coherent, human-like conversations with users that include answering queries, writing poetry, solving riddles, summarising ideas and reasoning about the emotional states of others.

Manipulation through synthetic relationships[6]

Generative AI chatbots can already be used to create synthetic personas that users who are interacting with them form emotional bonds with.[7] This creates a risk: that the entity in control of the chatbot can use the emotional bond as a powerful tool for manipulation. There is already evidence of significant harm caused by this type of emotional manipulation.[8] The chatbot might, for example, be designed to convince the user to use certain products, or to subscribe to a particular

[1] https://stability.ai/blog/stable-diffusion-public-release
[2] https://google-research.github.io/seanet/musiclm/examples/
[3] https://chat.openai.com/chat
[4] https://ai.facebook.com/blog/generative-ai-text-to-video/
[5] https://www.bing.com/new
[6] https://www.humanetech.com/podcast/synthetic-humanity-ai-whats-at-stake
[7] For instance, see Replika (https://replika.com/)
[8] Vide the Replika incident. Replika (https://replika.com/) is a company that "provides an AI companion that cares". Users can create a personalised avatar and develop a relationship with it. Recently, Replika discontinued a particular functionality that enabled an intimate type of communication capability. Many users became profoundly distressed as they had developed a deep attachment to the avatars. The issue was widely publicised in the media (e.g. https://www.businessinsider.com/replika-chatbot-users-dont-like-nsfw-sexual-content-bans-2023-2)

political belief. These preferences and beliefs could be integrated into the chatbot's persona, making it difficult for users to even recognise the intentions of the chatbot's owners.

Considering how powerful the opinions and requests of our friends and family can be on our own preferences and behaviour, we believe manipulation through artificial relationships represents a serious concern. Though today interactions with generative AI personas are text-based, related generative AI technologies may soon be combined to generate photo-realistic human avatars with realistic voices, movements, and expressions. The increased fidelity of emotional connections with such digital human simulacra creates even greater potential for manipulation in the future.

A technology specific to big tech

Generative AI is so resource intensive that it is almost exclusively the domain of Big Tech companies. Not only do big tech companies have the required expertise, computing power, and data to build these models, they have the platforms from which to deploy them.

We believe that just as today, most interactions with content are moderated by Big Tech through social media, in the future most interactions with artificial personas will also be moderated through Big Tech. It is also clear that giving generative AI models more data and more computing power makes them much more capable, entrenching the existing advantages of Big Tech companies. Finally, the amount of personal data that Big Tech companies have about individuals may enable them to personalise generative AI chatbots more effectively, further increasing their manipulation capabilities.

Regulation required urgently

Big Tech companies are currently competing to develop, deploy and monetise generative AI in general, and chatbots in particular. ChatGPT, for example, upon release saw the fastest uptake of a new digital service in history.[9] This suggests that measures to mitigate the risks of generative AI manipulation are required urgently.

Gradient Institute believes that the scale of the risk, and the previous failure of market forces to control AI-driven manipulation by Big Tech in social media, imply the need for strong regulation of generative AI manipulation.

Recommendations

- The Australian Government should take immediate action to define circumstances in which, or purposes for which, the use of generative AI for emotional manipulation should be prohibited.

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

- The Australian Government should seek to **establish an expert advisory committee** with diverse and relevant expertise on AI risks and AI safety, drawing broadly from industry, government, academia, the nonprofit sector and the broader civil society, to monitor the rapidly evolving AI risks and provide ongoing advice to the Government on how Australia should best respond to those risks.

_____

William Simpson-Young

Chief Executive, Gradient Institute