Project Overview

Reinforcing responsibility into language models: The case of OpenAI's language generator GPT-3

Workshop01-Group03

Team Name: The Explorer

Team members:

Fengqing Wu	**u7166770**
Ming Xu	**u7076449**
Xufeng Zhu	**u6825259**
Yuxuan Yang	**u7078049**
Zixian Huang	**u6872840**

1. Project background

Last June, OpenAI released the third version of a language model called GPT-3 that can be considered as a huge leap. According to OpenAI's description, GPT-3 can not only write novels, poetry, and news reports, as well as summarise long texts, but also generate guitar tab, and create program codes. However, despite the powerful ability of GPT-3, the model is not that perfect as many people think. There are many issues such as bias and misleading outputs, and high cost.

2. Project Purpose

This project aims to figure out the current biggest concern of GPT-3 and comp up with useful and effective suggestions and recommendations for GPT-3 key stakeholders to help them to have better responsible developments on artificial intelligence.

3. Project Challenge

Reviewing the social, ethical, and legal concerns of GPT-3 is challenging, as understanding how known discrimination and unconscious biases are embedded in the language model is quite hard. Also, the efficient ways to mitigate social and ethical issues are still clouded.

4. Project Stakeholders

Four different types of key stakeholders can be seen from the below table:

Table1 Stakeholders matrix

Stakeholder Analysis Matrix	**Low interest**	**High interest**
Low power	Individual citizens	GPT-3 users (Newsagency, educational institutions, advertising agency, internet companies, etc)
High power	Governments	OpenAI company, GPT-3 Developers, Engineers

Table 2 briefly shows how the stakeholders affect and are affected.

Table2 Stakeholder analysis matrix

Stakeholders	Affect	Be affected
Governments	have authority to make regulations and policies	Under pressure from public opinion
OpenAI GPT-3 Developers and engineers	Have the power to design and implement the GPT-3	highly related to the outcome of the GPT-3 on financial and emotional aspects
GPT-3 Users	Their using purposes can drive the way GPT-3 is heading	The bias output of GPT-3 can bring unexpected risks
Individual citizens	Their comments can be a power to affect the GPT-3	GPT-3 might cause some of them lost jobs

5. Problem

The main concerns of GPT-3 are the disinformation in the model outputs and potential risks of misuses.

5.1 empathy

And the empathy maps of two key stakeholders are shown as following:

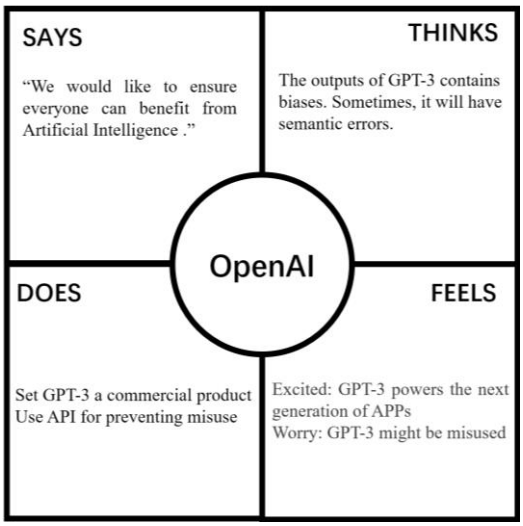

Figure1 Empathy map of OpenAI

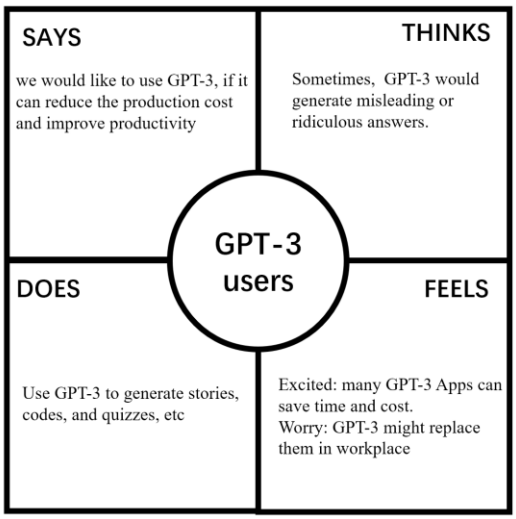

Figure2 Empathy map of users

5.2 Point of View

GPT-3 users who work on such as news institutions, education institutions, or other fields that are highly relevant to the public, need to pay more attention to potential risks of discriminations, disinformation, and privacy issues when using GPT-3, This is because most datasets trained on GPT-3 come from all raw sources of the Internet, which contains many biases such as sexism, racism, fake and misleading information, and even identified sensitive information.

The policymakers and governments need to build new regulation and policy to prevent the misuse of GPT-3, because a lack of safeguards and toxicity filters, a low cost of API entry, and no specialized technical required make GPT-3 successfully and effectively be weaponized by extremists to spread influx of machine-generated disinformation and propaganda related to large-scale terrorism and radical violence.

6. Solution

The key solution to mitigating the risks of disinformation and misuses of GPT-3 is to improve laws and regulations to fit the future society and to persist in moral educations. And the skeleton frame of the strategy can summarise as “one central task and two basic points.” (see Figure3)

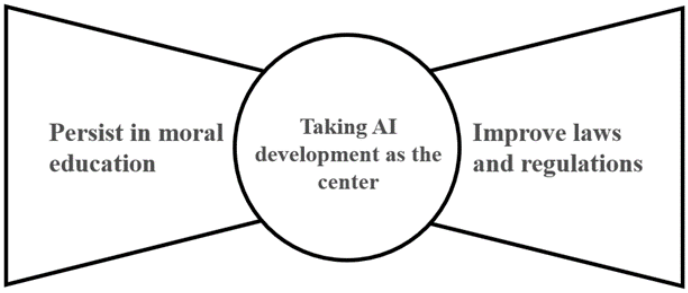

Figure3 A prototyping to mitigate GPT-3 key issues

One center task means no matter what strategies and methods are used to address AI challenges and problems, it should set developing artificial intelligence as the priority. As for the two basic points, one point is improving laws and regulations to fit the future society. Currently, countries in the world are stepping into industry 4.0 at different levels, and more and more AI issues would come up in different fields and industries as AI will become the main power in developments. The other point is persisting moral educations. There are many ways to achieve that goal such as setting up an AI ethics course, conducting AI debate competitions, advertising, social norms campaigning, or even filming Hollywood Science Fiction movies. This is a long-term strategy, and if people have the guidance of the right moral rules, moral rules will exert a subtle influence to help people know what is right and what is wrong.

7. Conclusion

Despite GPT-3 is facing many issues currently, the ability of GPT-3 cannot be ignored, especially since the whole society, now, has been stepping into a new industrial revolution, all walks of life will rely on AI to a different degree. From the result of empathy, AI is a double-edged sword, as long as people can make good use of it, then the benefits that AI can bring are immeasurable. As for current key concerns, legislation is paramount, while building new social norms and conducting AI moral education are also indispensable.