Text Generating AI technologies: A case study analysed by the method of Value Analysis

Introduction

This case study involves the partial release of a new AI text generating model called GPT-2. The model has been developed by OpenAI, a non-profit research organization (see: https://openai.com/). In contrast to the manner in which they handled the release of previous products, OpenAI decided not to release GPT-2 completely due to concerns about potential malicious use:[1]

> "Our model, called GPT-2 (a successor to GPT), was trained simply to predict the next word in 40GB of Internet text. Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in responsible disclosure, we are instead releasing a much smaller model for researchers to experiment with, as well as a technical paper" (OpenAi 2019, paragraph 1).

Description

Below follows a nontechnical description of GPT-2' text generating capabilities taken from OpenAI's website (see Radford et al. 2019, and OpenAI Code 2019 for technical details):

> "GPT-2 generates synthetic text samples in response to the model being primed with an arbitrary input. The model is chameleon-like—it adapts to the style and content of the conditioning text. This allows the user to generate realistic and coherent continuations about a topic of their choosing […].
>
> [...] our model is capable of generating samples from a variety of prompts that feel close to human quality and show coherence over a page or more of text. Nevertheless, we have observed various failure modes, such as repetitive text, world modelling failures (e.g., the

[1] The complete case study is published on OpenAi's blog: https://openai.com/blog/better-language-models/

model sometimes writes about fires happening under water), and unnatural topic switching. Exploring these types of weaknesses of language models is an active area of research in the natural language processing community.

Overall, we find that it takes a few tries to get a good sample, with the number of tries depending on how familiar the model is with the context. When prompted with topics that are highly represented in the data (Brexit, Miley Cyrus, Lord of the Rings, and so on), it seems to be capable of generating reasonable samples about 50% of the time. The opposite is also true: on highly technical or esoteric types of content, the model can perform poorly. Fine-tuning offers the potential for even more detailed control over generated samples—for example, we can fine-tune GPT-2 on the Amazon Reviews dataset and use this to let us write reviews conditioned on things like star rating and category.

These samples have substantial policy implications: large language models are becoming increasingly easy to steer towards scalable, customized, coherent text generation, which in turn could be used in a number of beneficial as well as malicious ways" (OpenAi 2019, paragraph 4-8)

OpenAI sketch the policy implications as follows:

"Large, general language models could have significant societal impacts, and also have many near-term applications. We can anticipate how systems like GPT-2 could be used to create:

- AI writing assistants
- More capable dialogue agents
- Unsupervised translation between languages
- Better speech recognition systems

We can also imagine the application of these models for malicious purposes, including the following (or other applications we can't yet anticipate):

- Generate misleading news articles
- Impersonate others online

- Automate the production of abusive or faked content to post on social media
- Automate the production of spam/phishing content

These findings, combined with earlier results on synthetic imagery, audio, and video, imply that technologies are reducing the cost of generating fake content and waging disinformation campaigns. The public at large will need to become more skeptical of text they find online, just as the "deep fakes" phenomenon calls for more skepticism about images.

Today, malicious actors—some of which are political in nature—have already begun to target the shared online commons, using things like "robotic tools, fake accounts and dedicated teams to troll individuals with hateful commentary or smears that make them afraid to speak, or difficult to be heard or believed". We should consider how research into the generation of synthetic images, videos, audio, and text may further combine to unlock new as-yet-unanticipated capabilities for these actors, and should seek to create better technical and non-technical countermeasures. Furthermore, the underlying technical innovations inherent to these systems are core to fundamental artificial intelligence research, so it is not possible to control research in these domains without slowing down the progress of AI as a whole (OpenAi 2019, paragraph 14-17).

Against this backdrop, OpenAI justify their GPT-2 release strategy as follows:

"Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: 'we expect that safety and security concerns will reduce our traditional publishing in the future, while increasing the importance of sharing safety, policy, and standards research,' and we see this current work as potentially representing the early beginnings of such concerns, which we expect may grow over time. This decision, as well as our discussion of it, is an experiment: while we are not sure that it is the right decision today, we believe that the AI community will eventually need to tackle the issue of publication norms in a thoughtful way in certain research areas. Other disciplines such as biotechnology and cybersecurity have long had active debates about responsible publication in cases with clear misuse potential, and we

> hope that our experiment will serve as a case study for more nuanced discussions of model and code release decisions in the AI community.
>
> We are aware that some researchers have the technical capacity to reproduce and open source our results. We believe our release strategy limits the initial set of organizations who may choose to do this, and gives the AI community more time to have a discussion about the implications of such systems.
>
> We also think governments should consider expanding or commencing initiatives to more systematically monitor the societal impact and diffusion of AI technologies, and to measure the progression in the capabilities of such systems. If pursued, these efforts could yield a better evidence base for decisions by AI labs and governments regarding publication decisions and AI policy more broadly.
>
> We will further publicly discuss this strategy in six months. If you'd like to discuss large language models and their implications, please email us at: languagequestions@openai.com (OpenAi 2019, paragraph 18-21).

Analysis

In what follows, the Value Analysis method is used to analyse the partial release of GPT-2.

1. *What is the incident about? (What is the dilemma?)*

The main question at stake in the case study is whether OpenAI's GPT-2 release strategy is ethically justified.

2. *What might (the central character) do to try and resolve the dilemma? (What alternatives exist?)*

When it comes to the release of GPT-2, the available alternatives include:

I. No release at all
II. Partial release
III. Complete release

3. *What might happen if he or she does each of these things? (What might be the consequences of the various alternatives?)*

I. In the no-release scenario, nobody outside OpenAI would initially know about the existence of GPT-2. However, leaks might obviously occur. In addition, in the short-to-medium terms other research organizations would probably develop similarly powerful models. Not releasing GPT-2 would thus most likely not avoid the development of similar technologies by others. It would at most postpone the emergence of this technology. In addition, the no-release scenario would not be ideal to facilitate a debate about the ethics of powerful text generating AI systems. After all, there would not be any trigger for such a discussion as nobody outside OpenAI would know about the existence of such potent dual use technologies. Of course, OpenAI could still start such a discussion without releasing anything about GPT-2. However, when asked why they engaged in such a debate, it would be disingenuous not to reveal anything about GPT-2. If in such a context OpenAI were to disclose the true backdrop of their eagerness to spark a debate, this would prompt the partial-release scenario.

II. The partial-release scenario seems ideal to both postpone the emergence of the full-fledged text generating AI technologies and to trigger a lively discussion of the dual use character of the same. On the one hand, it would postpone the appearance of the technology, thus generating time for solid reflection and the development of policy frameworks, if need be. On the other hand, it would expose enough about the prospects of the technology to make people aware of the need for a substantial discussion of the dual use problems of text generating models without anybody being immediately able to mobilize the potential of the full-fledged model.

III. In a complete-release scenario, everybody would be immediately able to use the full-fledged version of GPT-2. This would trigger the fastest development of the AI technologies. It would not generate any leeway for a discussion to take place and necessary policy frameworks to be developed in advance of substantial societal impacts. Discussion would not be perceived as acute because of the full release.

4. *What might happen to those who are not immediately involved? (What might be the short as well as the long-range consequences?)*

In all three scenarios, full-fledged text generating AI systems would be developed. This means that in all three scenarios one would encounter the likely positive and negative societal impacts of sophisticated, general language models as listed by OpenAI in their blogpost.
Potential benefits are:

- "AI writing assistants
- More capable dialogue agents
- Unsupervised translation between languages
- Better speech recognition systems" (OpenAi 2019, paragraph 14)

Potential harms are:

- "Generate misleading news articles
- Impersonate others online
- Automate the production of abusive or faked content to post on social media
- Automate the production of spam/phishing content" (OpenAi 2019, paragraph 15)

The only significant difference would be the time and space left in advance for the development of policy and regulatory frameworks to enhance the expected benefits and soften the anticipated harms associated with these systems. As discussed above, the partial-release scenario would optimize conditions for such important work to take place. The consequences could be captured in a values-information chart as follows:

Facts	Alternatives	Consequences			
		Short-Range		Long-Range	
		Self	Others	Self	Others
See case study description above	No release scenario	It would be more difficult for OpenAI to start a debate If they did and would be asked why, they would either have to be disingenuous or move to a partial release scenario	Everybody outside OpenAI would remain in ignorance about the technological development that are about to take place leaving less room for debate and policy development.	OpenAI does not spark any substantial debate; neither can it stop the development of the technology.	Positive and negative societal impacts occur because others develop the AI systems.

	Partial release scenario	This is the ideal scenario to start a debate, which would be in line with OpenAI' charter	Third parties would be better informed about the prospects of powerful text generating AI systems and have some time to discuss and develop policies.	OpenAI can spark substantial debate in order to prepare for these impacts, thereby allowing the development of measures that mitigate the potential harms and enhance the expected benefits.	Positive and negative societal impacts occur. There is a change that the ratio of benefits over harms turns out to be more advantageous than the benefit/harm ratio in the no release scenario.
	Complete release scenario	This would be in breach with Open AI's charter	This would leave others less time for debate and policies to be developed before substantial impacts of the technology would likely occur.	Open AI might still play a role in facilitating debate. However, the time frame would not be accommodating for the development of effective measures.	This is the fastest pathway for any positive and negative societal impacts to materialize, leaving the least amount of time for debate. Hence the benefit/harm ratio is likely less advantageous than the ratio in the partial release scenario.

5. *What evidence, if any, is there that these consequences would indeed occur?*

OpenAI's forecasts concerning the societal impacts and applications of large, general language models seem highly acceptable. Since the technologies at hand are completely novel, it is difficult to draw analogies with existing technologies. However, the expected benefits and harms as listed by OpenAI are all almost self-evident. The benefits that OpenAI lists, i.e., writing assistants, more capable dialogue agents, unsupervised translation between languages, and better speech recognition systems, would occur because more powerful text generating AI systems would simply enhance already existing technologies and research endeavours in these fields. The expected harms listed by OpenAI, i.e., the generation of misleading news articles, the impersonation of others online, the automated production of abusive or faked content to post on social media, and automated production of spam/phishing content, are equally likely to occur as they are simply extrapolations from existing societal phenomena.

6. *Would each consequence be good or bad? Why?*

OpenAI's assessment of potential benefits and harms seems broadly correct. The applications that are branded as beneficial might, for example, make life easier and work more effective. The malicious applications could lead to erosion of trust, social disruption and reputational damage amongst others. It would therefore be worthwhile to try and create some time and space for the development of appropriate regulatory frameworks and policies. The analysis above leads to the following value analysis chart:

Alternatives	Consequences	Desirability from various points of view							Ranking
		Moral	Legal	Aesthetic	Ecological	Economic	Health and Safety	Etc.	
No release scenario	Positive and negative societal impacts occur because others develop the AI systems OpenAI does not spark any substantial debate	-	-	N/A	N/A	-	-	N/A	Suboptimal
Partial release scenario	Positive and negative societal impacts occur OpenAI can spark substantial debate in order to prepare for these impacts	+	+	N/A	N/A	+	+	N/A	Best scenario
Complete release scenario	This is the fastest pathway for any positive and negative societal impacts to materialize Least amount of time for debate	-	-	N/A	N/A	-	-	N/A	Suboptimal

From a moral point of view, the partial-release scenario is most desirable because it is likely to enhance the expected benefits and reduce the expected harms. From a legal perspective, the same desirability assessment ensues because it is desirable to have a bit more time for the preparation of regulatory frameworks to deal with the impacts. Aesthetic and ecological considerations seem immaterial to the overall assessment. In terms of the economy, it seems

desirable as well to optimize the benefit/harm ratio as many of the benefits and harms will be of an economical nature. The same goes for health and safety as many of the potential harms are to do with breaches of cybersecurity and disruptions of democratic institutions, which could have severe negative effects on health and safety. Additional assessment criteria seem inconsequential. All in all, partial release seems the most desirable scenario.

7. *What do you think X should do? (What do you think is the best thing for X to do?)*

In all three scenarios, full-fledged text generating models would be developed either immediately (complete-release scenario) or in the short-to-medium terms (the other two scenarios). The partial-release scenario is the only one that optimizes the conditions for a discussion which might prove beneficial when it comes to the development of guidelines and policy frameworks in order to diminish the potential negative and enhance the expected beneficial societal impacts of this emerging technology. That is why OpenAI's release strategy, i.e., partial release, is indeed the best option of the three alternatives.

References

OpenAI Blog. *Better Language Models and Their Implications*. (2019, February 14). OpenAI. Retrieved March 15, 2019, from https://openai.com/blog/better-language-models/

OpenAI Charter. (n.d.). OpenAI. Retrieved March 15, 2019, from https://openai.com/charter/

Open AI Code. *Openai/gpt-2*. (2019). [Python]. OpenAI. https://github.com/openai/gpt-2 (Original work published 2019)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *[PREPRINT]*.