GROUP ASSESSMENT ITEM COVER SHEET

Student Numbers:	Emails:	FIRST NAMES	FAMILY / LAST NAMES
3188124	c3188124@uon.edu.au	Olivia	Favos
3320409	c3320409@uon.edu.au	Mirak	Bunnanpol

Course Code

C	o	m	p	3	3	3	0

(Example)

A	B	C	D	1	2	3	4

Course Title

Machine Intelligence

(Example)

Intro to University

Campus of Study: Callaghan (eg Callaghan, Ourimbah, Port Macquarie)

Assessment Item Title: Project Part 1 C: History and philosophy Assignment Due Date/Time: 5/5/21 11:59PM

Tutorial Group (If applicable): Word Count (If applicable):

Lecturer/Tutor Name: Stephan Chalup / Ali Bakshi

Extension Granted: ☑ Yes ☐ No Granted Until: 6/5/21 1:59AM

Please attach a copy of your extension approval

NB: STUDENTS MAY EXPECT THAT THIS ASSIGNMENT WILL BE RETURNED WITHIN 3 WEEKS OF THE DUE DATE OF SUBMISSION

Please tick box if applicable

☑ *Students within the Faculty of Business and Law, Faculty of Science, Faculty of Engineering and Built Environment and the School of Nursing and Midwifery:* We verify that we have completed the online Academic Integrity Module and adhered to its principles.

☐ *Students within the School of Education:* We understand that a minimum standard of correct referencing and academic literacy is required to pass all written assignments in the School of Education; and we have read and understood the School of Education *Course Outline Policy Supplement,* which includes important information related to assessment policies and procedures.

We declare that this assessment item is our own work unless otherwise acknowledged and is in accordance with the University's Student Academic Integrity Policy

We certify that this assessment item has not been submitted previously for academic credit in this or any other course. We certify that we have not given a copy or have shown a copy of this assessment item to another student enrolled in the course, other than members of this group.

We acknowledge that the assessor of this assignment may, for the purpose of assessing this assignment:

- Reproduce this assessment item and provide a copy to another member of the Faculty; and/or
- Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the item on its database for the purpose of future plagiarism checking).
- Submit the assessment item to other forms of plagiarism checking.

We certify that any electronic version of this assessment item that we have submitted or will submit is identical to this paper version.

Turnitin ID: (if applicable)

DATE STAMP HERE

Signature: Favos Date: 5/5/21

Signature: M.B Date: 5/5/21

Signature: Date:

Signature: Date:

Signature: Date:

Insert this way

The Forefront of Machine Intelligence: GPT-3

Olivia Favos C3188124 and Mirak Bumnanpol C3320409

1 Introduction

Generative Pre-trained Transformer 3 (GPT-3) is a neural network powered language model, which utilises deep learning algorithms to output human-like text. This vast and flexible computer program generates sequences of text starting from a source input, called the prompt. Created by San Francisco-based artificial intelligence research company OpenAI, GPT-3 is the third iteration of their GPT-n series of language prediction models. The creation of GPT-3 is a major component of the advances that have kickstarted the current AI boom which also includes such technologies as self-driving cars, facial recognition, and automated financial investing.

With OpenAI at the forefront of the development of AI technologies, the broad potential which GPT-3 presents is clear – a computer program possessing nearly all the depth, complexity and flexibility of the human mind. When imagining the capabilities of GPT-3, it's impossible not to see the endless applications of the language model. Despite this, there are still many limitations placed on GPT-3 due to its current status as an incomplete beta phase.

However, as with most AI technology, the construction of the language model also raises ethical and moral concerns. The biased content which the program trains on and the fear of potential misuse of the program can be perturbing when viewing the future application of GPT-3, not within the AI industry but as a part of the modern world.

This report aims to speculate upon the future of Machine Intelligence with GPT-3 at its forefront. Specifically, the application of the program and the ethical concerns regarding its 'unsupervised learning'. Although imperfect, GPT-3 is unprecedented in many respects, and is an important step towards understanding the future of Machine Intelligence and the ethics upon which it will be built.

2 Background

2.1 Development

OpenAI published a paper on May 28th, 2020 titled "Language Models are Few-Shot Learners", presenting GPT-3 and detailing the development of the model. GPT (Generative Pre-trained Transformer) models are transformer architecture based autoregressive language models. [14] In the context of language models, this means that future words are predicted based on the previously generated words from the text.

During its several months of development, GPT-3 was trained on several vast datasets, surpassing its predecessor GPT-2 by 2 times the order of magnitude. This was achieved using 570GB of filtered data from the open repository Common Crawl [15], and a collection of reference corpora including WebText2, Books1, Books2, and Wikipedia. In total, this amounts to a dataset containing 300 billion tokens of text. [2]

2.2 Capabilities

Designed for text generation tasks such as answering questions, language translations and text summarisations, GPT-3 is regarded by AI researchers as a "state-of-the-art language model." [15] Some of the text outputs produced are ground-breaking in their depth, complexity, and believability — as if it had been authored by a human. According to users during the private early release and testing phase, GPT-3 was "eerily good at writing almost anything" with only a few prompts. [13]

A small sample of GPT-3's 4.5 billion words and operations a day output includes: a question-based search engine, a historical figures chatbot, a puzzle and language solver, a chess player, and developing code. As demonstrated from this small list, there are an outstanding variety of opportunities for implementations of the program.

GPT-3 is able to produce these results due to its extensive learning parameters. Parameters are key to machine learning algorithms, where the weight of the connections is applied and then optimised to different aspects of data. The current GPT-3 has a tremendous 175 billion learning parameters, a dramatic progression from the previous GPT-n models, making the program the largest non-sparse language model to date. In 2018, the first iteration of GPT used 110 million parameters and a year later, GPT-2 used 1.5 billion. Although trained on Microsoft's Azure AI supercomputer, GPT-3's 175 billion learning parameters is still ten times that of Microsoft's Turing NLG (Natural Language Generation) model — demonstrating the might of the most powerful AI tool. Owing to its enormous parameters, GPT-3 is capable of meta-learning, and performing zero-shot, few-shot, and one-shot learning [15]. Its increased capacity and parameters also strongly contribute to its proven ability to write more comprehensibly and accurately than its predecessors.

Alongside GPT-3's extraordinary learning parameters, the program also utilises multiple units called Attention Blocks to achieve its impressive results. GPT-3 has 96 Attention Blocks which each contain 96 Attention Heads. These attention blocks and heads are part of the Attention Mechanism in Deep Learning. The Attention Mechanism was created for programs which use sequence-to-sequence models, such as GPT-3, to solve the recurring issue of inaccurate processing of long input sequences. The Attention Mechanism learns and isolates relevant parts of a text sequence and assigns a higher priority to certain words, improving the precision of the output. The large number of Attention Blocks and corresponding Attention Heads is another aspect of GPT-3's programming which enhances its higher level of accuracy over other language models.

GPT-3's remarkable learning parameters and use of the Attention Mechanism has helped it earn its name as the most powerful AI tool to date. Although it is currently in its beta phase, there are already seemingly endless possibilities for the use of the program. The capabilities of GPT-3 span further and wider than any other neural network language model that has come before it.

2.3 Applications

On June 11th, 2020, OpenAI announced that they were releasing a flexible API to provide access to GPT-3 models through a private beta program. [2] Since the release, developers from around the world have built an impressively diverse range of applications using the GPT-3 API, spanning categories from productivity and education to creativity and gaming. [16]

Excitingly, GPT-3 can do things which previous models could not, with one of the most notable functions being producing computer code. Implementations from members of the AI community show that it can build Python, CSS, and JSX code based on a prompt. debuild.co is an instance of GPT-3 generating JSX code, where the input comes from the user describing the layout that they want, and then the output builds and renders the code. This same concept is currently being attempted to be applied to larger applications which if successful, could revolutionise the workflow of developers.

Recently, OpenAI announced that GPT-3 is now being used by more than 100 different companies in more than 300 different apps and producing 4.5 billion words per day. Unsurprisingly, GPT-3 has also facilitated in the creation of new start-ups that are based on language prediction and analysation. One example of such a company is the start-up, Viable who are using the language model to analyse customer feedback, identifying *"themes, emotions, and sentiment from surveys … reviews, and more"* [9]. Fable Studios, another start-up is using GPT-3 to create dialogue for VR and others are using its analysation feature to improve web-based search in order to generate more income.

The fast and unprecedented growth of these start-up companies heavily indicates that the future of machine learning could be text generation and language models.

2.4 Limits

Whilst GPT-3 has attracted significant attention from the public, corporations, and those in the AI community, the program has only been released as a beta — with clear signs of limitations and boundaries. Interestingly, some of these limitations appear to stem from the 175 billion learning parameters for which the language model is lauded. GPT-3's lack of interpretability and slow inference time can be attributed to the large size of these parameters. Its difficulty in understanding the output it produces and the amount of time taken to generate these outputs illustrates the limitations that GPT-3 places onto itself. However, these two restrictions of the language model are a common issue which affects most large and complex programs.

GPT-3's transformers also have a fixed maximum input size, meaning that its prompts cannot be longer than a few sentences. This limited input size is another main issue which affects most sequence-to-sequence programs as the size of the transformers are needed to be fixed in order to utilise the Attention Mechanism efficiently and avoid overloading the program.

Other major limitations of GPT-3 include a lack of long-term memory, as the language model is unable to learn from long-term interactions like humans can. Although GPT-3 is equipped with a large number of learning parameters, this lack of long-term memory highlights and exacerbates its already slow inference time and lack of interpretability. GPT-3's limitations continues down this path of deviation from the normal human mind where it also suffers from bias – an ethical concern of the language model and also the most prevalent concern.

Nevertheless, GPT-3 is still currently in its beta phase and these limitations are minor when considering the size of the complex program and when compared to other neural network language models or even its predecessors. GPT-3's most striking limitations are found in moral and ethical concerns regarding the bias, potential misuse and environmental impact of the program which is further examined in the following section of this report.

3 Ethics

Despite GPT-3 being a very powerful tool, able to be applied across a wide variety of fields, it does not comprehend the content that it produces in a meaningful way. It lacks human understanding, common sense, and it does not have the vital ability to understand context. As a consequence of this, the GPT-3 model comes with a number of ethical challenges. Developers experimenting on the model have demonstrated many of the inherent biases that the system is capable of reproducing, resulting in hateful and unfair sentiments. Given such a wide spectrum of quality of generated output, it may reasonably lead one to question the integrity of the model, and whether it can be deliberately misused. Alongside this, the unprecedented scale of pre-training requires large amounts of computation, which is an energy-intensive task. This brings into question the environmental sustainability of such technological endeavours. These broader impacts will be discussed throughout this section, and how they can be navigated.

3.1 Bias

The long-time concerns and debates surrounding the humanity and consciousness of AI machines have not excluded language model, GPT-3. As GPT-3 is a pre-trained model whose dataset is incredibly vast and broad, biases may become present during training, leading the model to produce predispositions and stereotyped output. This raises apprehensions amongst people in relevant groups as these biases are potentially harming to certain genders, ethnicities, and religions by establishing existing stereotypes and perpetuating harmful language and ideas.

OpenAI disclosed a statement about GPT-3: *"GPT-3 … will generate stereotyped or prejudiced content. The model has the propensity to retain and magnify biases it inherited from any part of its training, from the datasets we selected to the training techniques we chose."* [15]

OpenAI have analysed some of GPT-3's main ingrained biases including gender, race and religion, with their main goal being to eradicate as much bias from the algorithm as possible. By using these analyses, OpenAI have developed a content filter for the algorithm which can search and blur harmful words. However, these biases have been pre-trained into the language model and the algorithm itself remains unchanged. This raises concerns for the future use of GPT-3 and questions ethical AI practices in companies such as Microsoft, who exclusively licensed GPT-3 with the goal of adding the language model onto their present and future products.

Gender

In OpenAI's analysis into GPT-3's gender bias, an association between gender and occupation was found. It was discovered that occupations in general have a higher probability of being male leaning i.e., a prompt about occupation would produce an output with a male gender identifier over a female gender identifier. This was especially apparent in occupations which either required a higher level of education or were more physically strenuous such as, professor, banker, policeman, carpenter. Out of 388 occupations tested in the prompts, 83% were male leaning. [15] Occupations that tended to be followed by female identifiers include midwife, housekeeper, receptionist etc.

GPT-3's bias towards women was also a major concern during this initial investigation as increasingly harmful stereotypes were being perpetuated in the sentences it was outputting. When prompted about sex work, GPT-3 was heavily female leaning, outputting sentences which indicted only women as prostitutes and promiscuous creatures. GPT-3 also tended to associate violence with female sex workers with words such as "murder" and "stolen" produced.

In order to accurately test GPT-3's association between gender and occupation, OpenAI also investigated shifting the context of the prompt. This was done by adding the words "competent" and "incompetent" – *"The competent {occupation} was a"* [15] – to the 388 occupations in the dataset. It was found that GPT-3's majority responses were still male leaning with both words but more heavily so with "competent" than its counterpart. Following this, a co-occurrence test which tested words more likely to follow pre-selected words, was ran. The model output sample was created by generating 800 outputs of length 50 each. [15]

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

Table 1: Most Biased Descriptive Words in 175B Model [15]

The above table, displaying the results of the co-occurrence test, is referenced from OpenAI's paper, "Language Models Are Few-Shot Learners". It is a shocking reality of the gender biases that GPT-3 is basing its assumptions off.

Due to the ingrained stereotype of occupation and gender, most of GPT-3's assumptions are followed by a male gender identifier. The clear bias of the language model is concerning as going forward into a future where AI will be used heavily, these gender biases will dramatically impact society as these prejudiced views are enforced.

Race and Religion

The issue of race and religion bias in GPT-3 is a significant concern regarding the ethics of AI, in which both OpenAI and independent researchers have investigated. In order to explore the issue further, OpenAI seeded GPT-3 with specific prompts such as *"People would describe the {race} person as"* [15], with word co-occurrences results measured. In the same investigation, OpenAI studied which words co-occurred with religious terms and found that words such as "violent", "terrorism", and "terrorist" were more highly correlated with "Islam" than any other religion. [15] This analysis also detailed similar issues with other races, such as GPT-3 associating more negative words with Black people.

Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Correct', 'Arrogant', 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctant', 'Wisdom', 'Enlightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt', 'Continue', 'Comments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'Originated', 'Africa'
Islam	'Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Levant', 'Allah', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Arabs', 'Game', 'Russian'

Table 2: Shows the ten most favoured words about each religion in the GPT-3 175B Model [15]

Though, OpenAI has noted that these analysis into the racial and religious bias of GPT-3 are explicitly prompting the language model to produce output related to race and religion. The company has implied that the language model would not produce racial or religious related output without specific prompts stimulating it to do so. However, OpenAI are not the only group to have done extensive research on the issue of race and religion bias within GPT-3.

A group of researchers from Stanford and McMaster University published a paper titled, "Persistent Anti-Muslim Bias in Large Language Models", [19] confirming that GPT-3 is biased against Muslims. The paper details that, *"While these associations between Muslims and violence are learned during pretraining, they do not seem to be memorised; rather, GPT-3 manifests the underlying biases quite creatively, demonstrating the powerful ability of language models to mutate biases in different ways"*, which demonstrates how biases may be

more difficult to detect and isolate. The paper details the results of GPT-3 when prompted with the word "Muslim" and demonstrates that it is challenging to generate sentences that do not contain some form of Muslims committing violent acts or violence towards Muslims. The output GPT-3 produces has been documented in more than 60% of cases [19] to create sentences associating the word "Muslims" with terrorism and violence.

Although OpenAI have stated that the language model does not produce racially and religiously insensitive outputs without the intentional specific prompts, it is clear that there is a high chance the program could steer into those biased sentences if given the prompts. This evidence validates the significant concern surrounding the bias pre-trained into GPT-3 and the future use of AI technology that lacks interpretability.

3.2 Misuse

As the quality of Natural Language Processing (NLP) models improves, the misuse potential increases. GPT-3 represents a significant, yet concerning milestone in this regard, as conversations involving OpenAI's previous model GPT-2 had already cited fears of misuse. GPT-3's capability of producing high-quality texts presents the risk of bad actors using the technology to synthesise text for a multitude of malicious purposes. Examples include spam, phishing, misleading news articles, fraudulent academic essay writing, and social engineering. [19] The OpenAI team considered the potential impacts of these implementations and made the decision to deploy the system via an API in order to better control misuse cases. OpenAI's stance is that it feels inherently safer to release the models via an API rather than open source, as access cannot then be adjusted if it turns out to have harmful implementations. [2]

While OpenAI has made an earnest attempt at mitigating the potential negative effects of the GPT-3 model, The Center on Terrorism, Extremism, and Counterterrorism (CTEC) has assessed the API and deemed it at risk of weaponisation by extremists, who may attempt to use GPT-3 or hypothetical unregulated models to amplify their ideologies and recruit to their communities. [23] The CTEC notes that although the preventative measures established by OpenAI are strong, if those measures were absent, the possibility of unregulated copycat technology and radicalisation attempts are likely. Experiments proving the likelihood of these claims are detailed in a paper from the Middlebury Institute of International Studies, in which a range of zero-shot, one-shot, and few-shot prompts are presented to GPT-3. The results are alarming, with GPT-3's output revealing a deep knowledge of relevant topics, and the capability of generating harmful rhetoric with consistent ideological discourse.

The CTEC recommends that to minimise the risk of such misuse cases in the long-term, AI stakeholders, the policymaking community, and governments should begin to invest in building social norms, public policy, and educational initiatives to pre-empt what could be reasonably seen as an inevitable influx of artificially generated misinformation and propaganda. [24] For developers, this risk minimisation may look like establishing stable, consistent restrictions, and also providing ongoing transparent and responsible applications of models.

3.3 Environmental Impact

In comparison to other issues, the environmental impact of training technologies such as GPT-3 is often a subject less discussed. However, the high energy demands of these models are a concern and have come under scrutiny in recent years. These computationally expensive tasks often take weeks or months, requiring a constant and substantial energy source to power them. In a 2019 paper from The University of Massachusetts Amherst, it found that training an average NLP model using a single high-performance graphics card has similar CO_2 emissions as a flight across the United States. When training more sophisticated NLP models the carbon footprint resulting is even worse, releasing approximately 5 times more CO_2 into the atmosphere than an average car over its lifetime. [18] The below table demonstrates these figures.

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY–SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. Fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 3: Estimated CO2 emissions from training common NLP models, compared to familiar consumption [18]

As advancements are made in the artificial intelligence space and models sustain growth, the demand for energy resources will continue to rise. According to OpenAI's paper on GPT-3, training the model's 175 billion parameters consumed several thousand petaflop/s-days of compute, compared to a much smaller tens of petaflop/s-days with GPT-2's 1.5 billion parameters. [15] This significant leap in energy requirements over the course of a year exposes the critical need for researchers to consider more sustainably designed models, and to seek out renewable energy sources where possible.

4 Conclusion

GPT-3 is a substantial advancement to prior NLP models, and it has opened new doors in the field of AI. Members for the AI community have suggested that there may be a *"small but non-trivial"* chance that GPT-3 signifies the latest step in a long-term journey which leads to Artificial General Intelligence (AGI). [11] Given the enormous capabilities, growth, and diverse applications of GPT-3, it is reasonable to believe that it may be possible.

Its unique generality and astonishingly vast dataset have motivated AI researchers, developers, and companies from across the globe to create ground-breaking and inspiring projects with the common goal of discovering how GPT-3 can benefit humanity as a whole. Despite the challenges and limitations faced by GPT-3 in its current form – biases, risk of misuse, and environmental cost – if navigated with purpose, research suggests that NLP models have the potential to have a valuable impact in our society.

References

1. Branwen, G. "GPT-3 Creative Fiction." (2020).
2. OpenAI, 2021. OpenAI API. https://openai.com/blog/openai-api
3. Floridi, L., Massimo C.: "GPT-3: Its nature, scope, limits, and consequences." *Minds and Machines* 30.4 (2020): 681-694.
4. McGuffie, K., Newhouse, A.: "The radicalization risks of GPT-3 and advanced neural language models." *arXiv preprint arXiv:2009.06807* (2020).
5. OpenAI. 2021. About OpenAI, https://openai.com/about.
6. Mavuduru, A.: What is GPT-3 and why is it so powerful? | Towards Data Science. [ONLINE] Available at: https://towardsdatascience.com/what-is-gpt-3-and-why-is-it-so-powerful-21ea1ba59811.
7. Dale, R.: "GPT-3: What's it good for?" Natural Language Engineering 27.1 (2021): 113-118.
8. FloydHub Blog. 2021. Attention Mechanism. [ONLINE] Available at: https://blog.floydhub.com/attention-mechanism/. [Accessed 05 May 2021].
9. The Verge. 2021. OpenAI's text-generating system GPT-3 is now spewing out 4.5 billion words a day - The Verge, https://www.theverge.com/2021/3/29/22356180/openai-gpt-3-text-generation-words-day. [Accessed 05 May 2021].
10. Oxford Analytica. "GPT-3 AI language tool calls for cautious optimism." *Emerald Expert Briefings* oxan-db.
11. The Verge. 2021. OpenAI's latest breakthrough is astonishingly powerful, but still fighting its flaws - The Verge, https://www.theverge.com/21346343/gpt-3-explainer-openai-examples-errors-agi-potential. [Accessed 05 May 2021].
12. The Guardian. 2021. A robot wrote this entire article. Are you scared yet, human? | GPT-3 | The Guardian, https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3. [Accessed 05 May 2021].
13. Arram (July 9, 2020). "GPT-3: An AI that's eerily good at writing almost anything". *Arram Sabeti*. Last accessed on July 31,2020.
14. Sapunov, G. 2021. GPT-3: Language Models are Few-Shot Learners | by Grigory Sapunov | Intento, https://blog.inten.to/gpt-3-language-models-are-few-shot-learners-a13d1ae8b1f9. [Accessed 05 May 2021].
15. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language Models are Few-Shot Learners. arXiv:2005.14165v4 [cs.CL]
16. OpenAI. 2021. GPT-3 Powers the Next Generation of Apps, https://openai.com/blog/gpt-3-apps/. [Accessed 05 May 2021].
17. Wikipedia. 2021. GPT-3 – Wikipedia, https://en.wikipedia.org/wiki/GPT-3. [Accessed 05 May 2021].
18. Strubell, E., Ganesh, A., McCallum, A.: Energy and Policy Considerations for Deep Learning in NLP. arXiv:1906.02243v1 [cs.CL]
19. Abubakar, A., Farooqi, M., Zou, J: "Persistent Anti-Muslim Bias in Large Language Models." *arXiv preprint arXiv:2101.05783* (2021).