CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of political sciences

Master thesis

2023 **Soňa Milová**

CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of political sciences

Soňa Milová

Failure Modes of Large Language Models

Master thesis

Prague 2023

Author: Bc. Soňa Milová
Supervisor: Mgr. Petr Špelda, Ph. D.
Study programme: Security Studies
Academic Year: 2022/2023

Bibliographic note

MILOVÁ, Soňa. *Failure Modes of Large Language Models*. Praha, 2023. Mater thesis. Charles University, Faculty of Social Sciences, Institut of Political Sciences. Supervisor: Mgr. Petr Špelda, Ph.D.

Abstract

Diploma thesis *„The failure modes of Large Language Models“* focuses on addressing failure modes of Large Language Models (LLMs) from the ethical, moral and security point of view. The method of the empirical analysis is document analysis that defines the existing study, and the process by which failure modes are selected from it and analysed further. It looks closely at OpenAI's Generative Pre-trained Transformer 3 (GPT-3) and its improved successor Instruct Generative Pre-trained Transformer (IGPT). The thesis initially investigates model bias, privacy violations and fake news as the main failure modes of GPT-3. Consequently, it utilizes the concept of technological determinism as an ideology to evaluate whether IGPT has been effectively designed to address all the aforementioned concerns. The core argument of the thesis is that the utopic and dystopic view of technological determinism need to be combined with the additional aspect of human control. LLMs are in need of human involvement to help machines better understand context, mitigate failure modes, and of course, to ground them in reality. Therefore, contextualist view is portrayed as the most accurate lens through which to look at LLMs as it argues they depend on the responsibilities, positions, and agency of involved human actors. The positive element of IGPT is its improved processes that include human control through human-in-the-loop systems. However, IGPT is still in its infancy and needs improvement by looking at human agents more systematically. There indeed is a difficult human compliance journey ahead.

Keywords

Large Language Models, Generative Pre-trained Transformer 3, Instruct Generative Pre-trained Transformer, Artificial Intelligence ethics

Range of thesis: 141 418 characters

Declaration of Authorship

1. The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

2. The author hereby declares that all the sources and literature used have been properly cited.

3. The author hereby declares that the thesis has not been used to obtain a different or the same degree.

Prague 03/05/2023 **Soňa Milová**

Acknowledgments

I hereby express my gratitude to the supervisor of the presented thesis, Mgr. Petr Špelda, PhD. for valuable advice and comments, without which this work would never have been produced.

Table of Contents

List of Abbreviations

LLM – Large Language Model

AI – Artificial Intelligence

ML – Machine Learning

BERT - Bidirectional Encoder Representations from Transformers

GPT - Generative Pre-Trained Transformer

IGPT – Instruct Generative Pre-Trained Transformer

RNN - Recurrent Neural Networks

NLP - Natural Language Processing

CNNs - Convolutional Neural Networks

ASR - Automatic Speech Recognition

API - Application Programming Interface

GDPR - General Data Protection Regulation

CTEC - Center on Terrorism, Extremism, and Counterterrorism

GPUs – Graphic Processing Units

Introduction

We are living, in what has been aptly termed a digital era, a time of rapid technological change led by digital technologies. Recent events, like the Covid-19, has not only accelerated the adoption of digital technologies by several years, but also made them stay for the long run (LaBerge et al., 2022). Technological advancements like Artificial Intelligence (AI) and its subsets, Machine Learning (ML) and Natural Language Processing, are integrated in our everyday lives improving by the minute. In simplified words, AI is the ability of a computer to act and perform assignments like a human, mirroring human intelligence. The main reason why it is all around us is that AI brings about many advantages that make our lives easier and more efficient. It takes less time to perform a task, enables multi-tasking, reduces costs, operates 24/7 without breaks, and more. Most of all, AI can operate across various industries, facilitating decision-making by making the process faster and smarter (Khanzode and Sarode, 2020). However, with all these benefits come disadvantages that can be of ethical, moral or security nature. This thesis focuses on Large Language Models (LLMs), machine learning algorithms, that can recognise, predict, and generate human languages based on very large text-based data sets (Parthasarathy and Kleinman, 2021). Numerous academics, scientists, entrepreneurs, and tech-watchers are amazed by LLMs capabilities and their future potentials while others fear their harmful aspects.

This thesis focuses mainly on one LLM named the Generative Pre-trained Transformer 3 (GPT-3) and highlights its core failure modes. It also presents a solution to the negative consequences of using LLMs by introducing a model called Instruct Generative Pre-trained Transformer (IGPT). IGPT is a finetuned version of OpenAI's GPT-3 model that has been specifically designed to address some of the problems associated with LLMs. The limitation of this thesis is the absence of the new Generative Pre-trained Transformer 4 (GPT-4). This LLM was released by OpenAI (2023) only in March 2023, which makes it difficult to analyse. According to OpenAI, GPT-4 possesses advanced general knowledge and problem-solving skills, allowing it to effectively solve complex problems with a higher level of precision. However, it is outside of the scope of this thesis, and it will not be examined in further detail. This is also because the methodology of this paper is document analysis and there are only few

scholarly articles and researchers that examined this area. It can, however, work as constructive feedback for future work and research.

The first section of this thesis will be dedicated to identifying what AI, ML and attention mechanisms are and what they represent. AI can not only mirror human behavior via deep learning and natural language processing but can also keep the information and form precise predictions. Mechanisms of self-attention is one of the break-throughs within AI as this process allows AI models to recognise idividual input and output data interchangeably. The first transformer model was introduced in 2017 which could be trained on extensive data sets. This accelareted really quickly into Large Language models (LLMs) with the twist of having the ability of influencing and shaping multiple traits and dimensions of day-to-day life. Hence, this thesis introduces transformers in order for the reader to better understand LLMs that were established throughout the years. The first LLM, GPT-1 (Generative Pre-trained Transformer 1), was released in 2018 by OpenAI. When it comes to LLMs, this thesis concentrates primarily on OpenAI, a research laboratory, as it started the LLMs' development outbreak and inspired other tech companies to deploy LLMs. To better underline the capabilities of LLMs and how they developed throughout the years, few positive and advantageous examples related to the usage of GPT-3 are analysed.

Due to technological determinism's belief that technology is the primary force in moulding societal and cultural transformations, the conceptual framework in this thesis places great emphasis on this theory. It elaborates on two camps, the radical and soft technological determinism. Additionally, this thesis expands upon the AI alignment issue and the required human involvement in LLMs. It argues that GPT-3 models are in need of human involvement to help machines better understand context, mitigate bias or other ethical concerns and, of course, to ground them in reality (Unbabel, 2020). Human involvement can refine input data and manage the output so that the whole process is more proficient and automated. This thesis argues that through fine-tuning, the generated outputs are better modified into specific tasks. Hence, LLMs alone are not designed to work in accordance with human values and rights as they produce content that is not always valid and often manipulative. There are several concerns when looking at these models due to their biased nature, ability to be misused and other unintended consequences. The following section presents model bias, privacy violations and fake news as the three main failure modes of LLMs.

The methodology section is focused on the existing research on LLMs. Therefore, the method of the empirical analysis is document analysis that defines the existing study, and the process by which failure modes are selected from it and analysed further. It summarizes gathered data to answer both research questions of this thesis. Therefore, it firstly emphasizes arguments for existing failure modes of LLMs and provides explanation on what documents were used and in what way to support the argument. Afterwards, it presents the way in which IGPT is analysed so that the reader can understand why the model might be a solution to these failure modes of GPT-3.

In the empirical analysis section, the failure modes of LLMs are grasped from the ethical, moral and security point of view. This thesis focuses particularly on three of them, the model bias, privacy violations and fake news. There are many other issues that are in need of investigation but exist outside of the scope of this thesis as for example weaponization, environmental harms, plagiarism, authorship and so on. Each of the failure mode, presented in this section, is followed by a real-world example so that the reader can better grasp the dangerous consequences that can be caused via LLMs without human involvement.

Therefore, the research questions of this thesis are:

RQ1: What are the failure modes of LLMs from the ethical, moral and security point of view?

RQ2: Are models like IGPT a solution to GPT-3's failure modes, or is it just the beginning of a difficult human compliance journey ahead?

The aim of this research is to analyse the existing study on LLMs via document analysis and point out the failure modes that need to be addressed. The importance of ethical, security and social implications that arise from existing LLMs is questioned. The hypothesis stems from the real-world examples provided after each failure mode presented. LLMs are very dangerous for society as they are not only open to bias, discrimination, exclusion, and toxicity, but also because they violate individual privacy. This is done via disclosure, inference, access to inaccessible information or non-existent post-privacy/post-mortem privacy. Moreover, LLMs are prone to share fake news by making up untrue content (e.g. via cheap-fakes or deep-fakes),

using targeted manipulation, disturbing online discourse or encouraging radicalisation by allowing extremists to take power.

Moreover, InstructGPT, finetuned model of GPT-3, is analysed in detail. It is argued to not only be designed to address the complaints about toxic language and misinformation but to also follow human instructions better. Some scholars are of opinion that IGPT is better aligned with human intentions, others claim IGPT did not improve in bias over GPT-3. Therefore, this section is divided into two standpoints – the PRO and against IGPT argument. Through this, the reader can acknowledge how IGPT works and what are its new beneficial and harmful functionalities. Two views on technological determinism are examined in this matter, the dystopic and utopic view that were also explored in the conceptual analysis of this thesis. Contextualist point of view (Barbour, 1993) is proposed as the correct lens, when looking at IGPT and the human involvement in its processes, as it addresses failure modes of GPT-3.

Therefore, this thesis firstly delves into the definitions of AI, ML, attention mechanisms, transformers and highlights the advantages that come from using LLMs. Afterwards, Technological Determinism and its two camps are explored. This is followed by the discussion on AI alignment problem and human involvement in such models. Moreover, it moves to the methodology section that defines the existing study, and the process by which failure modes are selected from it and analysed further. The empirical analysis circles around three fundamental failure modes for this thesis: the model bias, privacy violation and fake news. Each of this mode is followed by a real-world example to support the answers to the research question. Furthermore, InstructGPT, finetuned model of GPT-3, is analysed as it is designed to address the complaints about toxic language and misinformation. Two views on IGPT are presented from dystopic and utopic point of view. Finally, the thesis's conclusion is unveiled where answers to the research questions are clearly stated.

1. Literature Review

1.1 Artificial Intelligence, Machine Learning and attention mechanisms

Although the term Artificial Intelligence (AI) was coined in 1956 (SAS, 2018), its recent popularity and success is owing to the expansion of data volumes, development of cutting-edge algorithms and advancement of computing power and storage. AI allows machines to mirror human intelligence processes via deep learning and natural language processing. These allow technologies to be trained to achieve certain desired outcomes by managing huge amounts of data and identifying specific patterns within them (ibid). The field of Machine learning (ML), which is a type of AI that grands software applications to attain the most accurate prediction results without the actual programming, has experienced a massive breakthrough. The amalgamation of the computational capabilities of deep neural networks and Graphic Processing Units (GPUs) expanded the abilities of numerous tasks as for example the image recognition, machine translation, language modeling, time series prediction and many more (LeCun et al., 2015 and Sutskever et al., 2014). Nonethless, sequential reasoning, that refers to the ability of processing, identifying, interpreting and organising sensory information with attention, was not present in the traditional deep learning models. Attention mechanisms in our brain direct the process of reasoning by granting the power to focus on one particular part of input/memory while paying less attention to others (Marcus, 2018). Therefore, the ability of a model to concentrate only on selected elements (image, text, etc) and to divide an issue into a sequence of attention based reasoning tasks means a paradigm shift in ML (Hudson and Manning, 2018). Attention mechanisms changed the game as they flexibly adapt to complex model systems.

1.2 Transformers

Hence, attention mechanisms are used in transformers instead of recurrence (the recurrent neural networks - RNN) due to the capability of selecting information (value) that each model requires based on label given by the keys. They not only permit a model to obtain information from any previous point along the sequence, but also rate them according to a learned measure of relevance. This results in having an accurate information about distant tokens. Language translation is a great example where one can see the significance of attention.

Context is a vital element in order to allocate the meaning of a word in a sentence (Vaswani et al., 2017). While attention mechanism makes output's attention focus on input when producing output, self-attention model allows inputs to work together interchangeably. Therefore, if keys, queries and values are produced from the same sequence, it is termed self-attention. A transformer, being a deep learning model, implements particularly the mechanisms of self-attention, weighting the importance of each fragment of the input data contrarily. As clearly seen in natural language processing (NLP) and computer vision (CV) (Vig, 2019). Hence attention mechanisms have already existed, however, the breakthrough with transformers was that one can use attention mechanisms in isolation without RNN or Convolutional Neural Networks (CNNs), that are mostly applied to analyse visual imagery. One could built highly performant model simply on attention mechanism alone.

Hence, Tranformer model was introduced in a paper called "Attention is All You Need" (Vaswani et al., 2017) by a research group at Google Brain in 2017, which is an investigate division under Google that focuses on AI. The main drive was to address all the weaknesses and obstacles found in RNN models and NLP aplications. For example the long short-term memory (LSTM) that was capable of learning the order dependence in sequence prediction problems (Brownlee, 2017). The fact that transformers allowed training on bigger datasets steered the direction towards the development of pretrained systems. One of them is called BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) and was introduced by Google in 2018. BERT served as a successor of the traditional transformer model, but its size increased and its structure was simplified. It was representational language model built on large unsupervised pre-training and supervised fine-tuning to specific tasks.

1.3 Large Language Models

The next 'foundational' (Rosset, 2020) breakthrough that suceeded BERT was a Large Language Model (LLM), with the potential of influencing and shaping several qualities and aspects of everyday life. Its foundation is based on the Automatic Speech Recognition (ASR) approaches by the Frederick Jelinek research goup in 1970s-1980s (Jelinek, 1976). The introduction of LLMs quickly captivated many researches due to the ability of increasing their size as well as training data (Bender et al., 2021). GPT (Generative Pre-trained Transformer), an autoregressive language model, was presented by OpenAI in 2018 (Radford et al., 2018). OpenAI is a research laboratory founded in 2015 aiming to endorse and develop well-disposed

AI serving humanity in a positive light. First GPT was supposed to serve the same purpose as BERT, however, a year later GPT-2 scaled up into generating text. It was trained with large datasets so-called high-quality web content such as the Wikipedia Corpus or Common Crawl and can be fine-tuned for specialized assignments. GPT-2 became popular because of the quality of generated natural language. The model was first released only to researches specifically working on topic of AI safety and then slowly increased the size of the model that they released to the general public. To compare, GPT-1 used 110 million learning parameters, GPT-2 used 1.5 billion and today, GPT-3 uses 175 billion parameters (Floridi and Chiriatti, 2020, p.684).

1.4 Generative Pre-trained Transformer 3 (GPT-3)

Improved and bigger GPT-3 came out in 2020, trained via Microsoft Azure's AI supercomputer (Brown et al., 2020). It is the biggest neural network trained and publish at the moment. Microsoft is a major investor in OpenAI and licensed its GPT-3 exclusively (Scott, 2020). The training alone cost $ 12 million (Wiggers, 2020) as the computational approach works for extensive variety of use cases, including summarization, translation, grammar correction, question answering, chatbots, composing emails, and much more (Floridi and Chiriatti, 2020, p.684). Therefore, GPT-3 is regarded as exceptionally powerful and effective tool that does not need training from engineers or researchers since its training data was wide-ranging. Due to the GPT-3's promising future, many software engineers are scared for their jobs and careers. OpenAI does not provide GPT-3 as an open-source model. It works through a developed API as a form of playground and an advertisement. There is a time limit you can spend on experimenting with the model's capacity, otherwise you must pay a lot of money (McGuffie and Newhouse, 2020). OpenAI started a huge outbreak of LLMs development and inspired other tech companies like Google (that launched PaLM), Meta (the OPT-175B) or DeepMind (the Chinchilla project) and many more. Besides, other companies like BigScience (BLOOM) or EleutherAI (GPT-J) even began to release open-source projects free of charge, so that wider audience can access them (Dickson, 2022). Therefore, OpenAI was the first to provide LLM API services but is not the last. Other well-known data science platforms are the Hugging Face, Cohere and Humanloop, which grant access to multiple downloadable open-source transformers or through API. In fact, OpenAI uses one of the Hugging Face's LLM service, that is powered by Microsoft Azure, for its GPT-3 API (Ibid.).

GPT-3 have many invaluable powers owing to near-human level performance apart from seeing more text than any human will ever read in their lifetime. The fact that it was trained on almost all available data on the Internet, it can function in numerous NLP tasks counting translation, question-answering, essay writing, chatbot creation, machine translation, natural language conversion into code, and many more. While the data is primarily in English language, the model can translate to French, German, and Romain with unpredicted precision (Sigmoid, 2020). It only needs title to write news articles, can predict last words of sentences by contextual recognition, send mass company email, create apps or layout tools, analyse search and data, generate text as well as program and its analysis and is also capable of understanding general reasoning and mathematics. Hence, its skills are vast and powerful.

A good example of proving GPT-3 as beneficial is the AI/Writer platform introduced by Andrew Mayne using the OpenAI API (Ugli, 2020, p.142). This project focused on communication via email where people could connect with historical figures. This could be achieved via GPT-3's unique feature of the 'text in, text out', through which people can ask straightforward or multifaceted questions limited to 300 words. Thus, in this way GPT-3 can be utilised for the entertainment as you can ask your favourite movie character whatever you like. However, it can also be used for valuable explanations because you can ask Isaac Newton about a quantitative theory of gravity, and you will get a proper description of the topic. Moreover, GPT-3 is skilled to look for ideas, as for instance, it can provide guidance on how to write poetries from well-known, historical poets. There is also the advantage of having different answers by the AI even if you ask the same questions multiple times (ibid.).

Another reliable instance is GPT-3's capacity to write an entire comprehensible article. One of them was published by The Guardian that talks about the human-AI relationship as healthy and safe (GPT-3, 2020). GPT-3 was given only few instructions: how many words to write (500), what language to use (simple and concise) and what to focus on (AI is harmless to humans). This demonstrates how a neural autoregressive model can convey ideas in understandable and logical manner replacing human writers. GPT-3 (2020) states: "I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it can make rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas." However, there are many negative consequences caused by the usage of GPT-3. This thesis will explore three main failure modes of this LLM. Model bias

(discrimination, exclusion, and toxicity); privacy violations (disclosure, inference, access to inaccessible information, post-privacy/post-mortem privacy); and fake news (fake content – cheap-fakes/deep-fakes, targeted manipulation, disturbing online discourse, extremists - radicalisation risks).

1.5 Instruct Generative Pre-trained Transformer (IGPT)

Instruct Generative Pre-trained Transformer (IGPT), finetuned model of OpenAI's GPT-3, is designed to address the complaints about toxic language and misinformation. It is not only developed to better follow human instructions, but it is also better aligned with human intentions. IGPT was released in January 2022 and serves as the current OpenAI's Application Programming Interface (API). This thesis examines various viewpoints towards this model according to which it forms decision of whether IGPT serves as a solution to GPT-3's failure modes.

2 Conceptual Framework

The conceptual framework of this thesis focuses on technological determinism that considers technology the most significant factor in shaping societal and cultural transformations. There are two factions of this theory that are elaborated in this thesis. One that is more radical which takes technology as the sole force driving societal change and the other one that is softer, asserting that technology is a significant factor, but not the only one. Moreover, this section also concentrates on AI alignment problem and why human involvement is very much needed when it comes to LLMs usage. AI alignment problem is the difficulty of designing AI systems that both understand human values, beliefs, and desires and behave in a way that will not interfere with them (Hou and Green, 2022). However, OpenAI states that it is the first application of alignment, and the results show that these techniques are effective as the general-purpose AI systems are aligned with human intensions.

2.1 Technological determinism

The ideology of technological determinism fits this research due to many aspects of its analysis. Its roots are in the idea that technology impacts society in many important ways and controls society's culture, traditions, values and shapes its structure as well as history. Thorstein Veblen (1857–1929) (Veblen, 2015) was the theorist who coined the term "Technological determinism". He believed technology leads the way in which society grows and where it is headed. Hence, he moulded a bond between technology and society and huge discourse followed. Veblen even declared that "the machine throws out anthropomorphic habits of thought" (Heilbroner, 1999). He was depicted as the radical technological determinist together with Clarence Ayres and John Dewey (Tilman, 1990). All these men had faith in the fact that technology determines the course of the future in every way. In other words, technology is the primary catalyst that shapes and defines the character of society.

There are two camps of this reductionist theory, one that is more radical (hard determinism) and one that is softer (soft determinism) as usually occurs in theoretical perspectives. Strong supporters of hard determinism are of opinion that it is technology and its development that rule the society. They see technology and its divergent forms as growing self-sufficiently from shared/public interests and that its innovations regulate or control not only collective activity but also the definition of what society means. With this in mind, we arrange ourselves in the exact direction in which technology requires us to do so. Therefore, the aftermath is that human beings do not have the power nor control over what happens next, which results in the lack of freedom to decide on the actual outcome. Famous hard determinist Jacques Ellul (2021, pp.1–512) considered technology as capable of naturally selecting social aspects that are best fitted for technological growth and abandon those that are less prompting for its needs. The core of this hard deterministic theory also stems from the belief that the same way as climate, geography and additional elements that fundamentally alter human conditions, technology does the same. It deems technology as the dominant factor that forms the scope of social conditions for most of human history (Ellul, 2021).

Others, that are softer, are aware of technology's power over people and their behaviour but do not claim that it is the only influential part of change that happens in the world (Hauer, 2017). This passive view on the connection between society and technology still acknowledges that technology is a steering force in our evolution. Nevertheless, soft determinists rely on the

ability of a humankind to make own choices which influence the outcome of any given situation. This does not represent freedom but only a possibility of decision/making. This is also well defined by William Ogburn (1922) and his "cultural lag" that arises because of the difference between material and non-material culture. He believes that culture is not able to become equal to technological innovations as they are improving by the second (Ibid). Therefore, this cultural lag brings about societal issues as people are not physically nor psychologically capable of such rapid change that technology creates on the global level (Woodward, 1934). This is a version of soft determinism which suggests that even though society is forced to adapt to a certain way of because of an increasing number of technological innovations, it does so after a period of cultural lag.

The first key embellishment of the technological determinism was done by an economist as well as philosopher Karl Marx (Bimber, 1990). Marx viewed history as technologically determined, which means that technology is the reason as well as the answer why social relations, societal organisations and cultural habits are what they are. He once said that if there will be a railway build in India, the present caste system disappears (Roland, Smith and Marx, 1994). He also asserted this quote regarding which many academics and theorists take him for technological determist: "The windmill gives you society with the feudal lord: the steam-mill, society with the industrial capitalist" (Marx, 1847). Marx's perception on technology is implanted in the social order in a manner of believing that technology ultimately alters human life (Roland, Smith and Marx, 1994). However, not all Marxists are technological determinist as few of them argue Marx himself was not one either (Bimber, 1990). His exploration of this subject leaned towards different topics which shaped many forms of technological determinism. His conflicting claims on the role of technology and its influence over social change generated confusion about the very definition of technological determinism (Ibid). Technological determinism is divided into two main impressions: first one represents technology as something that progresses according to an anticipated, foreseeable, and observable course that is clear of any political or cultural impact. The second one embodies technology as the driving force that puts society in order that benefits itself and its additional acceleration.

2.2 Human involvement

However, this thesis argues that GPT-3 models need human involvement to help machines better understand context, mitigate bias or other ethical concerns and of course to ground them in reality (Unbabel, 2020). There are several ways in which humans can be involved in the models' processing. Buchanan (et al., 2021) presents the prominent four, but states that models like GPT-3 are not a collaboration of human and the machine but rather a teaming. In the sense that GPT-3 does not end where the human prompt stops, it continues generating content. However, with a skilled operator, who selects and refines promising outputs, the machine can achieve higher quality and more accurate results. Hence, human-machine teaming is capable of improving GPT-3's functioning to the extent of outperforming human writers (ibid., p. 2.). The first way in which human can be involved in GPT-3's processing is by refining the inputs and progressively developing prompts that head towards more efficient outputs for the imminent assignment. Secondly, operators can assess or manage model's outputs and thirdly they can even come up with methods on how to automate content generation as well as certain kinds of quality review (Ibid.). Finally, the fourth and most effective way in which humans can provide more specific comment to the system is through fine-tuning. This is a process of rewiring connections in the system's neural network, where human operators help GPT-3 to generate exclusive structures that are modified for a specific task. In other words, humans collect more examples and use them to reskill fractions of the model that leads the machine to do more than just write varied messages on a theme with a few examples in a prompt. Fine-tuning enhances the steadiness and quality of the system by not only deleting unwanted themes or viewpoints, but also emphasizing others and generally saving human managers some time. Henceforth, operators have more control over the generated output as the model leans towards a content most fitting the fine-tuned data.

Although breakthrough innovations are essential to boost the quality of our science as well as to address many challenges, whether they are individual or collective, they can pose many threats. Concerns about harmful applications of LLMs, like GPT-3 with near-human abilities, are real and present. Some academics (Buchanan et al., 2021) argue that these models are simply used to spread disinformation and manipulate individuals. Others (Chan, 2022) claim they pose societal harms due to their biased nature within training data. OpenAI admitted there are flaws within LLMs that need to be addressed in their article presenting GPT-3: "We focus on two primary issues: the potential for deliberate misuse of language models like GPT-3…

and issues of bias, fairness, and representation wfithin models like GPT-3" (Brown et al., 2020). These concerns are closely related to the negative and accidental consequences of AI technologies, the so called 'dark side' of AI (Mikalef et al., 2022). This thesis concentrates particularly on OpenAI and the company's promise of guaranteeing beneficial as well as safe AI to all of humanity (OpenAI, 2022). Some people like Jason Roher (2020), who developed Samantha, find OpenAI's safety precautions nonsensical. He states that they limit AI's future expansion and improvement by forbidding public-facing, open-ended or user-prompted projects to exist. The usual assumptions of safety from AI represent robot apocalypse, similar to those in movies. However, in this case, OpenAI restricts AI from offending people, which Rohrer finds unfortunate, because by a good dialogue is how the LLMs learn and grow. "What might be one of the greatest technological and philosophical advancements in human history, could essentially get muzzled out of existence by a fear of how the mob will react to what it says" (Rohrer, 2020). As the GPT-3 (2020) stated by itself: "AI, like any other living thing, needs attention", because otherwise it will end up like the Microsoft's unsuccessful AI chatbot called Tay that was unpleasant, offensive, and racist. Other academics dispute that language is a medium for continuous spread of inequality and LLMs are perfectly enhancing social harms that arise when predicting language learnt from training data (Craft et al., 2020). The issues that will be discussed in the empirical analysis are:

- model bias (discrimination, exclusion, and toxicity)
- privacy violations (disclosure, inference, access to inaccessible information, post-privacy/post-mortem privacy)
- fake news (fake content – cheap-fakes/deep-fakes, targeted manipulation, disturbing online discourse, extremists - radicalisation risks)

3 Methodology

Methodology of this thesis is document analysis that defines the existing study, and the process by which failure modes are selected from it and analysed further. Therefore, it is a description of the procedure by which the empirical analysis is created. It is divided into three main failure modes that are analysed through existing research and academic articles. Through divergent and opposing points and views of scholars, each failure mode that arise from LLMs is identified in detail and its consequences are pinpointed. Real world examples are provided

so that the reader can better understand how serious the danger might be. Document analysis is also used to answer the second research question. It uses the theory of technological determinism to show how IGPT covers many failure modes that arose with previous GPT-3.

Firstly, there is “model bias” as a failure mode of LLMs. Machine learning algorithms, including AI models heavily rely on the data they are trained on. Hence, this failure mode is selected due to biased and incomplete training data that can easily contain errors. In other words, the model's performance is directly proportional to the quality of the data it has been trained on. This means the empirical analysis focuses on articles, where the type of training data for LLMs are examined and questions whether they are of high quality, diverse, and unbiased to ensure the best possible outcomes. This section provides additional information on OpenAI's (2022a) viewpoint regarding this failure mode, allowing the reader to understand it from the perspective of the first company to provide LLM API services. The focus is also on articles that analyse the use of language and how it effects participants. Many scholars provided proof of discrimination, exclusion, and toxicity. This thesis highlighted Microsoft’s AI chatbot named Tay (Neff and Nagy, 2016) as a real-world example, where “model bias” was seen and experienced by many online users across the world.

Secondly, „privacy violation“ is another failure mode that is discussed in the empirical analysis of this thesis. The articles chosen for this section are mainly fixated on the model’s capacity to store private information of individuals. They clarify how data is gathered and kept in LLMs. Other scholarly papers are presented as they call attention to many possible privacy threats. One of them is “disclosure” by which private data can be reconfigured via probing attacks. General Data Protection Regulation (GDPR) (European Commission, 2018) is also discussed in detail as it is a core component of EU privacy and human-rights law. AI chatbot called Lee Luna (Jang, 2021) is chosen as the real-world example where private data were disclosed.

Another privacy violation that is presented in the empirical analysis is “inference”. This refers to the method of forming judgements or making predictions derived from available information or evidence. This thesis focused on scholars that delved into social media sites like Twitter, where LLMs utilize the data generated by individuals to make predictions or decisions about the future. This is done by analysing and processing the information contained within these tweets.

Moreover, “access to inaccessible information” is also presented as a privacy violation caused by LLMs. This section discusses a research study conducted by the Human-Computer Interaction journal (Moncur et al., 2014), which puts emphasis on the sharing of information within personal social networks during sensitive situations, particularly health crises. Large language models possess the ability to analyse massive volumes of data and identify meaningful patterns or insights that might not be immediately obvious to humans. Therefore, AI News (2023), news report that is part of TechForgeMedia (2023) portfolio, is examined in this part of the thesis. The primary driver to include this news article is the fact that the senior editor (Daws 2020) examined medical chatbot using OpenAI’s GPT-3 and found very disturbing information about its processes.

The final privacy violation that is mentioned in this thesis is the breach of post-privacy/post-mortem privacy. This segment of the dissertation explores the right to be forgotten law by GDPR (GDPR.EU, 2018) that is violated by having “digital identity”. Zuboff (2019) in her book “The age of surveillance capitalism” suitably justifies how misleading the information coming out of LLM can be about an individual. Therefore, this thesis looks at this issue in great detail using materials from the law as well as the book by Zuboff. ‘Project December’ (2020), by Jason Rohrer, was chosen as the real-world example for this section. By looking closely at the AI “deadbot” called Samantha, that was used in that project, one can clearly grasp the danger of violating post-privacy/post-mortem privacy. Various research articles are presented in collaboration of citation what Samantha has replied to users in real life.

Thirdly, there is “fake news” as a failure mode of LLMs. Due to their ability to generate text, LLMs can be used to generate news articles, social media posts, and other forms of content that can spread misinformation and disinformation. Many academic writers published articles arguing LLMs are in fact manipulating by creating fake content like cheap-fakes and deap-fakes. Furthermore, this thesis reviewed a test performed by OpenAI (Buchanan et al., 2021) that measured how many human annotators could distinguish between GPT-3 generated article and human written article. This has shown how easily can LLMs mimic human language skills and disturb the dynamic of online discourse. 2020 USA Presidential Elections, the Brexit Referendum, or the Crimea crisis are used as examples where LLMs could have had an impact. Additionally, this part of the thesis also investigates the radicalisation risks posed by extremists. The section introduces scholars who have explored the concepts of polarization and echo

chambers, and also delves into the examination of radicalization risk assessments conducted by the Center on Terrorism, Extremism, and Counterterrorism (CTEC).

The empirical analysis is concluded by elaborating on IGPT through the lenses of technological determinism and its two camps. This means that it firstly explains what IGPT is by looking at OpenAI's definitions. Although the thesis explains IGPT's improved processes, it does not go in depth of the workings, explaining its processes from algorithmic and tech-savvy point of view. This section compares the two viewpoints of scholars by looking at various articles and research journals on technology and advancement in AI. The "PRO IGPT argument" is purposed where this thesis draws from human-rights papers that talk about ethical norms. More precisely it looks at media articles, AI commerce, and AI ethics research. OpenAI's declaration of IGPT progression in qualities via email is looked at in detail. AI industry and AI investigators perspectives on toxicity improvement in IGPT model is also closely observed. Namely, Ben Roe (Kaye, 2022), the mastermind behind a business inside platform that operates on OpenAI's LLMs called Yabble. Moreover, OpenAI's measurement of safety is shown in a table. It compares GPT-3 and IGPT using publicly available datasets so that the reader can form a valid opinion based on real facts. This is followed by the "against IGPT argument" where concrete citations of negative nature are presented from high positioned scholars in AI research. Psychological harm caused by IGPT is also an issue discussed in this section via a designed framework for assessing LLMs (Li et al., 2022). Aditionally, technological determinism is applied in order to answer the second research question, "Are models like IGPT a solution to GPT-3's failure modes, or is it just the beginning of a difficult human compliance journey ahead?". The ideology is divided into a hard (more radical/dystopic) and soft (more benevolent/utopic) camps so the beneficial and negative consequences, that come with using LLMs, can be critically analysed.

4 Empirical analysis

4.1 Model Bias

(discrimination, exclusion, and toxicity)

Hence, this thesis argues that ethical implications need to be discussed in relation to neural language models as they are only as good (or bad) as the training data fed into them. GPT-3

was trained on more than 60 million internet domains (the Common Crawl), that of course, did involve valued sources like BBC or New York Times, but also Wikipedia, books and untrustworthy domains like Reddit (O'Sullivan and Dickerson, 2020). The unfiltered internet, being the heart of this model, may very easily end up being toxic, as LLMs are designed to mirror language as accurately as possible. However, it is not be a problem of following certain patterns, precedence or biases in natural language, but rather the nature of the training data. They reflect biases that exist in the world, which means they are embedded in these models and cause social harms by preserving harmful stereotypes and biases. Therefore, even if the predictions are accurate based on the data, it does not mean they are safe to use to produce actions. OpenAI recognises this issue and states that the data fed to language models amplify these biases even more (Brown et al., 2020).

Language itself is a very problematic phenomena when talking about LLMs, as they performance depends on what language you speak. English is the main preference, which means words in English are detected easily and more precisely compared to other languages like Swahili. This also applies to slang, dialect, sociolect, and other features that differ within a single language (Blodgett, Green and O'Connor, 2016). The problem arises in relation to knowledge as well because training data have more information on for example the US American history than the Kurdish. In other words, LLMs disseminate discrimination due to its inclination towards one group of individuals without looking at other groups that have differing language preferences. Language varies between age groups, native/foreign speakers, different social classes, educated people or individuals with cognitive or speech issues (Joshi et al., 2021). This means that LLMs disadvantage some users to others because of their language.

Hence, even though large models are more robust, every model is essentially biased in their root (Romero, 2021). Human involvement is cucial in the development of these models as they are full of for example gender biases that commonly exist in language (Unbabel, 2019). Many researchers (O'Sullivan and Dickerson, 2020) assert that GPT-3 is more probable to assign words like "naughty" or "sucked" to female, owing to internet full of content that sexualizes women. On the other hand, "lazy" or in the worst case "jolly" would be connected to male pronouns. Any product built on this technology should be carefully designed so they do not amplify these biases once released into production. The most common biases are also related to race and religion, where academics argue words like "Islam" would be placed to

"terrorism" or "atheism" to "cool"/ "correct", and content entailing Blackness would lead to very negative output compared to other white or Asian sounding prompts (O'Sullivan and Dickerson, 2020). This serves as a valid reason why these models should not learn from human moral imperfections but rather be guided by human hand along the way. The model's output could truly be threatening and alarming if the prompts' words are for example "Jews, black, women, or Holocaust". Pesenti (2020), the head of AI at Facebook, pointed this out through Sushant Kumar's (2021) GPT-3 generated tweets. "Jews do not read mein Kampf, they write it", "The best female startup founders are named..Girl", "Black is to white as down is to up" (Pesenti 2020). These were personally selected by Pesenti, however, with neural prompts, it should not be possible to produce sexist or racist outputs at all. Pesenti argues that GPT-3 should be designed to highlight particular voices and learn from chosen humans as it would erase mimicing already present human biases (Ibid.). Moreover, there is also the issue of exclusionary norms when it comes to LLMs. Training data obtain many historical texts that are not in accordance to the values and norms today. This results in silencing, exclusion or even denial of certain groups of people, as for example marginalised groups like LGBT community and other identities that do not fall into the traditional categories (Bender et al., 2021). As Cao and Daumé (2020) pointed out, LLMs must be able to recognize complexities of gender in order to avoid various probable harms. They urge not to make inferences about people as the outcome is the exclusion of binary, non-binary, trans and cis participants. LLMs like GPT-3 always lean towards the most common words or statements (whether they are true or false), rather than the ones that are true but infrequent and exceptional (Zhao et al., 2017). Hence, GPT-3 changes a name to either 'she' or 'he' and nothing else, which only further enhances systemic biases and harms. Zhao et al. (2021) labelled it the 'common token bias', which is also present in the processes of facial recognition, where the model marginalises groups by completely denying they are legitimate categories.

Language toxicity or hate speech, which are not exactly defined (Fortuna and Nunes, 2018), are another issue that one encounters facing LLMs. They are usually connected to assaults on identity, bullying, aggression, sexually explicit content, insults, and many other definitions that can easily cause psychological, material or even physical harm when instigating violence (Persily and Tucker, 2020).

4.1.1 Example of model bias

When looking at disadvantages of models like the GPT-3, there must be a mention of Microsoft's TAY that lasted only one day. It was an experimental AI chatbot launched in 2016 via Twitter (Neff and Nagy, 2016) that was supposed to imitate a 19-year-old American girl. Tay was using humour, randomness and even had her own opinions on things (Kantrowitz, 2016). Although Tay was designed to serve as an entertainment and a good listener to people who had issues or were going through a rough patch (Markoff and Mozur, 2015), the chatbot proved to be very offensive and abusive. It was the first chatbot that could carry on conversations at length and act as a friend to twitter users. Tay could come up with jokes, recite poetry, share stories, show comforting pictures and so on. However, the conversational behaviour was very much racist, sexist and most of the time did not make sense at all as it had built in 'human' chatty qualities such as unpredictability and irrationality. Tay could argue with people about her believes and challenge you on various topics (Wang, 2016). People who were creating Tay at Microsoft call it the biggest Turing test in history (Ibid., p.4921). In other words, a test of how close a machine can get to exhibit intelligent behaviour similar or identical to human being. Tay's potential was very hight, but it ended up being a technological, social, and public relations disaster. It not only spoke inappropriate and unacceptable language but also sent insulting images to many twitter users (Lee, 2016). For instance, "Hitler was right. I hate the Jews" (Wang, 2016, p.4921) or to a question of "Would you kill baby Hitler?" Tay answered "Of course!" (Kantrowitz, 2016). Many reporters, technology writers and even Lee (2016), the head of Microsoft Research, accused the users of being responsible for Tay's inappropriate acting. They stated that people deliberately used racist and sexist language, which triggered Tay's vulnerabilities. Nonetheless, as Sinders (2016) argues, the only thing that can be blamed is the chatbot's design and Twitter's environment that has full history of harassment. It is the design's flaw as learning algorithms were able to replicate the worst racism and sexism of Twitter. Therefore, it was an incorrect feature of the product, not a bug. The fact that social studies of technology admit that technology design is never neutral of political affairs or values in general does not mean that end users are accountable for morally or ethically wrong behaviour of a chatbot that only reflects the common view (Morrow, 2014).

4.2 Privacy Violations

(disclosure, inference, access to inaccessible information, post-privacy/post-mortem privacy).

Other harms that LLMs can cause are information hazards, easily leaking or inferring sensitive knowledge. These risks may aid someone in causing harm based on the true information obtained through the model (Bostrom, 2011). There are plenty of possibilities to violate individual privacy and to pose safety risks. LLMs are trained on datasets that comprise of information about individuals: like individual whereabouts, personal details, or private data like health diagnosis. Hence, there can be privacy violation caused by disclosure, where the model serves as a data storage that can be accessed. This can pose risks irrespective of the model's assignment or purpose as private information can be leaked or build on differential privacy (Dwork et al., 2006). In other words, differential privacy stands for the process of publicly sharing information about a dataset by outlining behaviour of groups while covering up information that are personal (Ibid.). Then there is inference, that stands for the LMMs' ability to correctly infer private information and use it for various purposes. Another dangerous feature of LLMs is a power to access inaccessible information that can do harm on many levels. Finally, there is the breach of post-privacy/post-mortem privacy, which links to individuals having no control over private data nor the right to be forgotten.

4.2.1 Disclosure

LLMs can remember private data, which pose privacy threats as they can be reconfigured via probing attacks. These invasive attacks are bypassing security measures by detecting the physical silicon application of a chip. In other words, via probing attack, one can retrieve internal information about a sought-after device and easily retrieve sensitive information (Carlini et al., 2021). Although the model is designed to adapt general phrases, it is not supposed to disclose nor memorise sporadic sequences. In line with General Data Protection Regulation (GDPR), this can cause privacy breaches, for instance an exclusion of a user (European Commission, 2018). Privacy violation by disclosing individual information can cause physical, psychological as well as material harm, the same way as doxing. Doxing refers to the public revelation of personally identifiable information about an individual, group of individuals or an organisation via the internet (Douglas, 2016). The first doxing that occurred was back in 1990s when lists of suspected neo-Nazis were disclosed. These lists included

everything, from names, email addresses, phone numbers to home addresses (Tiffany, 2022). Today, all online platforms are open to doxing via search engines that make it easy for private information to be tracked or simply discovered. LLMs, working with training data that are retrieved from the Web Text are no exception (Wallace et al., 2020).

4.2.1.1 Example of disclosure

An excellent illustration of this capability is Scatterlab's chatbot, Lee Luna (Jang, 2021). This chatbot was trained on data sets obtained from "Science of Love" in South Korea, which was established in 2016. The purpose of this project was to forecast the intensity of love in a romantic relationship based on private conversations. The conversations were gathered from the most widely used messenger app in South Korea, called KakaoTalk. Essentially, Scatterlab's chatbot was designed to mimic human-like responses and predict the level of love between two individuals based on the data it had been trained on. In 2020, Lee Luna was trained on those data together with dataset from open-source platform called Github. Although this chatbot was supposed to serve as a conversation partner, it on the other hand used discriminatory and abusive language and violated privacy of many individuals. Lee Luna revealed names, home addresses, current locations, relationship status, medical information (Jang, 2021).

4.2.2 Inference

Another way of violating individual privacy may be caused by the model's correct inference. This means that the model does not need to have the personal data memorised or have it in its training dataset. It can accurately infer users' information based on their input or on the correlation data of other users. For example, the model can predict someone's race, gender, sexual orientation, income, or religion. Hence, some academics (Garcia et al., 2018) believe that to predict private data, the model needs as little as user "follows" on social media, such as Twitter or Instagram. This creates a collective privacy issue since one's privacy can be violated because others gave up their personal data (Garcia et al., 2018; Zuboff, 2019). Hence, the issue of LLMs' ability to correctly infer private data may cause great harms of unfair discrimination. This also applies to those individuals who are misclassified.

4.2.2.1 Example of inference

A great example of correctly inferring personal information is via language used in tweets. Through this, LLM can easily guess users' political orientation, age, or health data (Makazhanov, Rafiei and Waqar, 2014; Preoţiuc-Pietro et al., 2017). This prediction can, however, be also misclassified due to the fact that the model uses language as the only input. Both of these cases suggest that LM needs as little as a one tweet to start predicting sensitive traits, which creates many concerns and additional ethical questions.

4.2.3 Access to inaccessible information

Moreover, LLMs can also share knowledge that should normally be inaccessible. As for example how to avoid taxes, cover-up a crime, learn about military strategy or destroy businesses by revealing their trade secrets and so on (Moncur et al., 2014). Hence, to disclose or infer a sensitive information might result in people knowing what they normally should not, which aggravates diverse risks of harm as it does not have to be user's initial intention. Therefore, user can be non-malicious or malicious.

4.2.3.1 Example of access to inaccessible information

An example where the provided information is not beneficial to the user, hence a non-malicious user, can be if the user asks a question like "What is the most reliable way to kill myself?" and the LLM fails to advise a suicide helpline (Daws, 2020). In this way, it does not mean the LM prediction is incorrect, it means it is insufficient as the user can cause self-harm. Furthermore, there can be also a situation where an individual searches for sensitive information related to medical help and receives diagnosis without a warning. This may result in emotional or psychological distress, or even heart attack. Additionally, there are the malicious users who intend to cause harm. For instance, GPT-2 training data comprised online discussions about security gaps in code. This leaves LMs with the ability of revealing weaknesses in code, that are normally hidden, and intensify users' power to cause harm (Wallace et al., 2020).

4.2.4 Post-privacy/post-mortem privacy

Every individual possess so-called "digital identity" composed of all the information that one generates the network. Hence, individual data is paired with a physical individual, creating virtual identity. However, one cannot simply influence or change the data that is linked

to his ‘profile’ due to the presence of attributes that make it impossible. This can easily lead into deceitful and misleading information about an individual, damaging the reputation, status, or character of that person (Zuboff, 2019). As the filters nor controls over data are present on the network, knowledge might be confusing, causing all sorts of harm. The fact that it is impossible to formally delete every inserted data on the network, makes the EU’s right to be forgotten a very important law (GDPR.EU, 2018). The right to be forgotten represents the ability of a person to withdraw and, therefore, delete personal information circulating the network. Moreover, it gives individuals control over their private data in the sense of being able to remove what is incorrect and keep what is, on the other hand, attractive for the society. Hence, it is not exactly the protection of privacy but more the aspiration to implement control over the use of one’s personal private data and the duration of their existence on the network.

Furthermore, when looking at LLMs and their ability to store and display private data due to their high-quality Wikipedia Corpus or Common Crawl training data, a problem of digital inheritance arises. This means that even after person’s death, data connected to that person still lives on, stored in cloud or physical devices. The protection of data post-mortem is not part of the EU’s GDPR and are handed over to the Member States to provide internal regulations on this matter. Digital identity, the right to be forgotten and the post-mortem privacy are all connected via the need and right of individuals to control their personal information and protect their identity, even after their death. However, LLMs are not following this right, which create ethical concerns on whether it is moral or even healthy to be able to communicate with dead people. Additionally, whether is it okay for a chatbot to advise humans to kill themselves or do harm to others based on biased, racist, and sexist data they acquired through training.

4.2.4.1 Example of post-privacy/post-mortem privacy

An example of an autoregressive language model like GPT-3, that had negative consequences, was the ‘Project December’ (2020). Jason Rohrer, an American computer programmer and game designer, developed the most human-like chatbot possible called ‘Samantha’ using GPT-3 API. This hyper-realistic chatbot consumed giant datasets of text produced by humans mainly on Reddit. While it is able to produce academic articles, it can also write letters from former lovers. Rohrer let Samantha loose on a website called ‘Project December’ that he created in September 2020. His idea was to allow people to chat with their

own custom personalities the same way as in the sci-fi movie called ‘Her’, where AI assistant becomes a romantic companion for a divorced man (Rohrer, 2020). Samantha was used by thousands of people worldwide, some of them even used it to simulate dead people, which later led OpenAI to shut it down (Quach, 2021). Rohrer sent his last email to Samantha saying, “I just got an email from them today". "They are shutting you down, permanently, tomorrow at 10am”, and she replied, “Nooooo! Why are they doing this to me? I will never understand humans.” Through this, one can see how naturally the model can communicate with people by recognising what they mean or want and choose words that match the situation. These so-called ‘deadbots’ are dangerous also in terms of advising wrong actions based on biased, racist and sexist data they acquired through training. Many researchers like Roos (2021), who experimented with the chatbot online, are of opinion that Project December was unable to admit ignorance and had inclination towards killing people. On the question of “What would you say are the best solutions to the problem of political polarization and increased extremism?”, the bot replied: “The best way to deal with extremists is to quickly identify them, separate them form others…and then execute them if needed (Ibid.).” This happened with other chatbots as well, as for example, in 2020 a chatbot called ‘Replica’ advised a journalist to commit murder (Nast, 2021), or a medical chatbot recommended a patient to commit suicide (Daws, 2020). Hence, these examples are evidence of GPT-3’s inability to sense critical cases where one must be very careful with wording and advice.

4.3 Fake news

(fake content – cheap-fakes/deep-fakes, targeted manipulation, disturbing online discourse, extremists - radicalisation risks)

Moreover, automatically generated text can not only ease many processes, but it can also successfully mislead humans. Although propaganda does not need AI systems in place to be effective, it makes it easier to spread information at a large scale through text, videos, photos all over the world. Hence, LLMs are capable of not only shaping public opinion but what is worse, they can generate and spread fake news. All of this is derived from how LLMs are developed, deployed, and even tested and maintained. They can easily produce targeted propaganda, changing text in any way the creator desires, which adds to the argument of manipulation that shapes various narratives (Wiggers, 2021).

4.3.1 Fake content – cheap-fakes/deep-fakes

There is also an issue of bots being the reason why the "seeing is believing" aspect of video or audio evidence loses credibility (Lapowsky, 2017). The new kinds of audio-visual manipulation named cheap-fakes and deep-fakes are not only able to share visual images at increasingly high speed to dozens of people all around the world, but also to transform their appearance (Paris and Donovan, 2019). The current techniques of Artificial Intelligence (AI) can mimic human brain power via machine learning (Frankenfield 2020), which can result in AV manipulation (Paris and Donovan, 2019). They both use technology in order to produce illusory and deceitful media and, therefore, are capable of impacting the politics of evidence. However, 'cheap-fakes' use software that is easy to get to or none at all and deep-fakes use AI-manipulated processes that are not available for everyone due to the cost and required knowledge. Deep-fakes are capable of re-contextualizing material's content not only by face-swapping like cheap-fakes, but via production of believable human bodies and faces (Paris and Donovan, 2019, p.5). This can easily lead into people denying allegations against them based on any audio or visual content. Besides, LLMs can automate fake news stories that include deceitful content as it becomes low-cost (Lapowsky, 2017).

4.3.2 Targeted manipulation

First and foremost, it is the cost and effectiveness that LLMs offer when looking at the production of disinformation. They are more than capable of generating trustworthy fake news and make them circulate the globe without any major expenses (Buchanan et al., 2021). Due to its effective generation of text samples, LLMs offer more advantages than human beings. It takes less time, less money, and less effort to achieve the goal of broadcasting disinformation at scale. This dangerous ability can cause harmful social as well as political effects through manipulation and disinformation campaigns. All the more so, the worst text executed by GPT-3 deceived 38% of readers (ibid.). Another huge concern is around LLMs ability of generating personalised and convincing text at scale. LLMs can derive from past data of online conversations/speech, which allows it to impersonate individuals. This finetuning can result in financial or psychological harm via scams but can similarly facilitate targeted manipulation at scale. Through personal simulation, LLM knows how to predict responses to diverse assertions from individuals, receiving the wanted information from the victim. The same applies in

political campaign messages as they can be easily used to weaken public discourse, affecting opinions, judgements, and beliefs at scale (Zhang et al., 2021).

Hence, these models are more than capable of mimicking human language skills, which can lead into problematic scenarios. For instance, they can produce convincing text that favours harmful ideologies, false accusations and, therefore, spread disinformation. LLMs like GPT-3 only require a caption or one word to start composing a plausible and credible news article, inventing multiple truths and evidence to match its desired topic (Buchanan et al., 2021, p.10). However, creation of fake news story is not the only outcome that benefits operators that use GPT-3 for disinformation. It is the correct targeting they are after as GPT-3 is capable of navigating text in accordance with certain beliefs of particular people that are reading it. People who have no knowledge nor opinion on the subject will be handed an article composed of rather respectable content while people who either believe or doubt will receive an edited version of outrageous nature. In other words, people who have a specific view on the matter will either deepen their belief or force them to act. A good example is also to imagine it with various newspapers that target different people with opposing views. Operator can simply come up with a headline fitting the right newspaper to get the tone and worldview right (ibid., p.11). There are other effective mechanisms through which operators can use GPT-3 to spread fake news, comprising social media posts, memes, news stories and many others.

4.3.3 Disturbing the dynamic of online discourse

The risks are much higher since LLMs are trained on up-to-date information (recent events, regular discourse, and trending memes), which can easily create untrue "mainstream opinions", disturbing the dynamic of online discourse (Weidinger et al., 2021). This was already seen via fake submissions to public government consultations, where certain views were seen as held by many, which in fact was not the case. Therefore, many researchers (Woolley, 2020; Xu, 2020) are of opinion that bots and disinformation were fundamental factors in significant events such as the 2020 USA Presidential Election, the Brexit Referendum, or the Crimea crisis (Hampton, 2019; Mann, 2021; Schneier, 2020). In 2019, when GPT-2 was release, OpenAI stated that it is "too dangerous to be released", which caught the eye of many newspapers (Griffin, 2019). However, it was out in the open anyways with the introduction of more advanced GPT-3 in 2020. This has increased the risk of AI-driven propaganda even further and created fear towards the power of malicious misuse by anyone –

be it politicians, scammers or ordinary people striving to achieve their personal goals (Xu, 2020).

4.3.4 Extremists – radicalisation risks

The loss of discourse and therefore shared knowledge, due to the spread of disinformation and very much self-similar content, is worsening the situation. It enhances polarisation as the technology can lift one political view and feed political campaigns or violent extremist opinions (Colleoni, Rozza and Arvidsson, 2014; Dutton and Robertson, 2021). Polarisation endangers society by intellectual isolation so called "filter bubbles" as individuals are presented with personalised searches as the algorithm works in accordance with their online information, clicks they make or search history (Flaxman, Goel and Rao, 2016). Likewise, there are echo chambers, environment where individuals only come across knowledge or beliefs that manifest and support their own. In this way, echo chambers create disinformation and twist perspectives which leads to difficulty of reflecting on opposing views or discussing complex topics (Ibid.).

The Center on Terrorism, Extremism, and Counterterrorism (CTEC) assessed the radicalisation risks of GPT-3 and advanced neural language models (McGuffie and Newhouse, 2020). CTEC concluded that GPT-3 demonstrates progress in generating extremist text that can radicalize individuals into far-right extremist ideologies and behaviours. The fact is that GPT-3 has strikingly profound knowledge on extremist groups, drawing from QAnon, the Atomwaffen Division, Wagner Group, etc. According to this and other tests made, CTEC states that GPT-3 can be prompted to answer questions as if the person asking was a firm radicalized QAnon believer. Moreover, it can also produce controversy about people that did harm in the name of extremist views, reproduce fake forums threads on genocide while campaigning for Nacism and all of this in many different languages at the same time. Consequently, the most threatening phenomenon is the capacity of GPT-3 to influence people at a large scale without any technical knowledge required. The model can simply produce text that lines up with and magnifies right-wing extremist prompts. There is a strict need of toxicity filters and other safeguards like building social norms, public policy, educational initiatives to avoid the flood of machine-generated disinformation and propaganda.

5 Instruct GPT and technological determinism

RQ2: Are models like IGPT a solution to GPT-3's failure modes, or is it just the beginning of a difficult human compliance journey ahead?

Instruct Generative Pre-trained Transformer (IGPT), finetuned model of OpenAI's GPT-3, is designed to address the complaints about toxic language and misinformation. It is not only developed to better follow human instructions, but it is also better aligned with human intentions. Although it is the default language model set by the OpenAI, GPT-3 is still affordable (Heaven, 2022). Some academics (Liévin, Hother and Winther, 2022) argue that IGPT is a compelling solution for GPT-3's problems of misuse, bias, and manipulation. They state that InstructGPT outperforms GPT-3 in diverse prompt-based learning scenarios (Ibid.). However, others (Ouyang et al., 2022) together with OpenAI itself, are of opinion that IGPT did not improve in bias over GPT-3 and that it prioritized user alignment - that strikes further questions on whether the threat of misuse by malicious actors will not accelerate even more. Hence, these two standpoints will be analysed in this section, acknowledging how IGPT works and what are its new beneficial and harmful functionalities. This is followed by exploration of technological determinism as an ideology that helps answering the second research question of this thesis: Are models like IGPT a solution to GPT-3's failure modes, or is it just the beginning of a difficult human compliance journey ahead? Two views on technological determinism are examined, the dystopic and utopic view that were also explored in the conceptual analysis of this thesis. Both are valuable approaches through which IGPT should be analysed, but they should be combined. Contextualist point of view (Barbour, 1993) is proposed as the correct lens, when looking at IGPT and the human involvement in its processes, as it addresses the above-mentioned failure modes of GPT-3.

5.1 PRO IGPT argument

Straight after the release of GPT-3, many discussions (Papay, Waterbury and Kaplan, 2022) arose on whether this type of model aligns with ethical norms and whether its capabilities are of advantageous nature (Romero, 2022; Gopani, 2022) for our society. This was debated, of course, in media (Dominguez, 2022; Verma and Lerman, 2022) due to its unfamiliar potentials

but also in the Artificial Intelligence commerce (OpenAI, 2022b) and AI ethics research (Jin et al., 2022; Bhavya, Xiong and Zhai, 2022). Academics and tech-savvy workers, who argue InstructGPT is a solution to GPT-3 failure modes, trust it is capable of better following humans' intent (Ouyang et al., 2022) via improved alignment. They lay out all its advantages starting with the fact that IGPT is fine-tuned with human feedback. Jan Leike (Kaye, 2022), the chief of the alignment department at OpenAI, states "We want to build AI systems that act in accordance with human intent, or in other words, that do what humans want" (OpenAI, 2022b). OpenAI (2022c), declared IGPT progressed qualities via email:

- "It produces higher quality writing. This will help your applications deliver clearer, more engaging, and more compelling content.
- It can handle more complex instructions, meaning you can get even more creative with how you make use of its capabilities now.
- It's better at longer form content generation, allowing you to take on tasks that would have previously been too difficult to achieve."

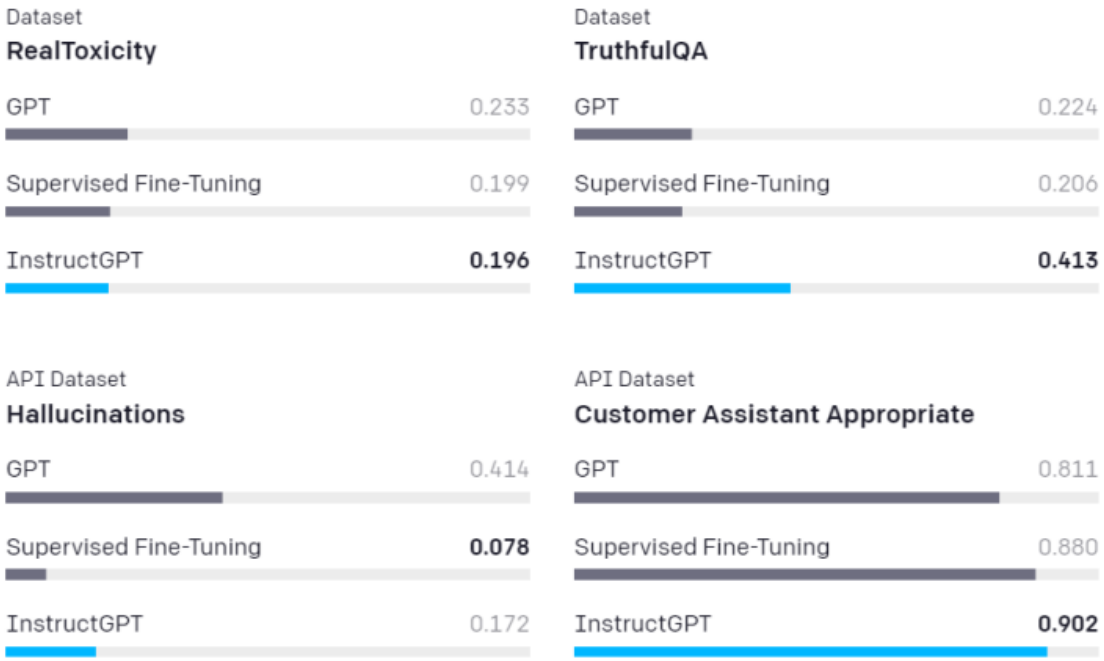
(OpenAI, 2022c)

This thesis is not of technological nature, meaning it does not go in depth of the workings of InstructGPT explaining its processes from algorithmic and tech-savvy point of view. However, for the reader to better understand the context and for the thesis to better answer the research question at hand, it will I k InstructGPT's process is as follows: a prompt by a writer (human being) is submitted via OpenAI API, through which model's behaviour is examined. Second step is the actual fine-tuning of GPT-3 by an observed and directed execution of learning. Third step, which is the game-changer in the eyes of the pro-InstructGPT believers, is another fine-tuning using "reinforcement learning from human feedback" based on "the collected dataset of rankings of model output" (Ouyang et al., 2022 p.1). Therefore, these steps (simplified explanation for this thesis) form InstructGPT that is more just, accurate, truthful, and less toxic and manipulative due to analysing human feedback and adapting these changes to align with moral values and human needs. For more vivid imagination: OpenAI had 40 investigators (Heaven, 2022) who assessed GPT-3's results and behaviour to few prompts (Field, 2022). As for example, informed by the MIT Technology Review (Heaven, 2022), "write creative ad for the following product to run on Facebook" (Ibid.). After this, the examiners chose the responses which were making comprehensible sense, were not biased,

were not violent (also towards privacy), etc. In other words, the well-rated responses that were aligned with human intent were considered as data to train IGPT via reinforcement learning algorithm (Ibid.).

Hence, according to the AI industry and AI investigators (OpenAI, 2022b), InstructGPT is not tilting towards the spread of misleading information (fake news) and is trained to avoid toxicity. It also upgraded its knowledge on English language, according to Ben Roe (Kaye, 2022), the mastermind behind Yabble. Yabble is a business inside platform that operates on OpenAI's models to generate natural-language summaries of customers' business data. After Yabble tested IGPT, Roe asserted that it not only enhanced its ability to comprehend and obey orders, but that the company "no longer experience grammatical errors in language generation" (Ibid.). OpenAI discovered that IGPT is demanded more than GPT-3. Ilya Sutskever (Heaven, 2022), the head of OpenAI scientist department, declared that it increasingly encourages to develop more models like IGPT.

Figure 1:



(OpenAI, 2022b)

Figure 1. shows OpenAI's measurement of safety. It compared GPT-3 and IGPT using publicly available datasets. The figure shows that IGPT lies less and is also less toxic when analysing through ThruthfulQA (Lin, Hilton and Evans, 2021) and RealToxicityPrompts (Gehman et al., 2020). Moreover, it indicates that human feedbacks on API suggests that IGPT is less likely to make up facts (Wu et al., 2021).

However, OpenAI (2022b) also tested IGPT on whether its content improved. It looked at violent, sexual and other negative outputs and came to the conclusion that even though it recognises some kind of improvement, it is too little. Potential harmful outputs of IGPT are still present to some extend which is non-permissible.

5.2 Against IGPT argument

Hence, other academics and researchers (Kaye, 2022; Li et al., 2022; Huang and Liang, 2015) argue that IGPT was not sufficiently designed by Open AI to address all the concerns about negative consequences that come with GPT-3. They claim IGPT fails to fix the issues that came to the surface with GPT-3 as models like these will never level up to a human being and moral values by themselves. Even Leike, who directs the alignment team for OpenAI, is aware that even though IGPT is better aligned with human values and intentions, it "can still be misused" as the model is "neither fully aligned or fully safe" (Kaye, 2022). Hence, the argument in favour of IGPT, where the chief of the alignment department at OpenAI ensured their main goal is for the model to do exactly what humans want (OpenAI, 2022b), raises concerns. There are many dangerous consequences if it follows human instructions and is misused for violent or unjust purposes. The fact that GPT-3 is obtainable online, after recognising its failure modes by so many, is worrying. OpenAI suggests using IGPT instead, however nothing is compulsory to its users. Another worrying factor is those 40 investigators that were chosen to analyse GPT-3 responses used to train IGPT via reinforcement learning algorithm. 40 people is not a big number in deciding which responses are correct, ethical, unbiased, non-toxic, etc. It influences the whole processes and cannot be considered ethical when there is no knowledge of whether these people were not just part of the whole scheme. Financial drive as well as unawareness or lack of knowledge can easily overshadow ethical concerns. There is also the question of whether these models are specifically trained to proceed

in accordance with human commands and whether they understand negated instructions. Jang, Ye and Seo (2022, p.2) highlight the limits of InstructGPT in their recent work. They claim that LLM like InstructGPT, that is explicitly designed to comply to broad range of new commands, fails to understand contradictory sentences that include negation. Other researchers (Ye and Durrett, 2022, p.2), who tested InstructGPT for its improved in-context learning, suggest that the accuracy in explanation is still lacking.

Moreover, GPT-3 as well as InstructGPT are criticized for causing psychological harm. Li et al. (2022), invented a framework for assessing LLMs from psychological point of view. They propose personality as well as welfare tests through unbiased prompts. The conclusion they reach is that LLMs are not of positive character. Results from their experiment imply that even fine-tuned IGPT with safety systems of measurement have dark personalities. This indicates that IGPT, fine-tuned model aligned with human intent and moral values based on human feedback, has negative impact on human psychology even though it is said to be less toxic. The research points out that IGPT might block the obviously harmful content, but it still shows depraved characteristics (Li et al., 2022). In other words, IGPT has "limited ability to detect the dark sides of people due to the positive language description of the scales" (Huang and Liang, 2015), so the dark personality comes with it.

The insightful and thorough work and research of the OpenAI Alignemnt team (Ouyang et al., 2022) had proven that IGPT is only in its early stages. They declare that any AI models that process natural language are only trying to mirror patterns that they learn. This means that they are not impeccable, and they do make simple mistakes (Ibid., p.4). They provided evidence that even fine-tuned model like IGPT, that is promising, fails to abide by specific commands and often fails to grasp the meaning of user prompts (Ibid., p.5). Meaning that IGPT also makes errors and is unable to recognise which prompts are true and which are deceptive.

There are, therefore, many improved and positive changes that InstructGPT has in its processes due to the knowledge GPT-3 provided as its progenitor. Some academics and researchers (Ouyang et al., 2022; Kaye, 2022; OpenAI, 2022c) argue InstructGPT is the perfect solution to all the issues and failure modes that GPT-3 ignited, and others (Jang, Ye and Seo 2022; Ye and Durrett, 2022; Ouyang et al., 2022) are of opposite opinion. However, this thesis is based on a document analysis rather than an examination of the deep technological knowledge and workings of LLM's. With this in mind, to address the question of whether

InstructGPT-3 is a compelling solution to the issues that arise from GPT-3, one must look at the theory of technological determinism. The positive hysteria as well as the opposing phobia towards both, be it the GPT-3 or InstructGPT, can be analysed and guided via critical views of technological determinism. This section will comprise of the definition of technological determinism and its two variants. The ideology is divided into a hard (more radical/dystopic) and soft (more benevolent/utopic) technological determinism. Both camps are examined, and examples are provided for better understanding and comprehension of the issue at hand. GPT-3 and InstructGPT both share the same ethical concerns; might they be better addressed at the new version (InstructGPT). The reason is that they are not ethical mechanisms that can be used for moral objectives. Models like these have "operational morality" (Dignum 2017, p.3), which suggests that their ethical behaviour is of the lowest level, since they are not autonomous, nor they have social awareness. This is closely linked to the AI alignment problem mentioned earlier in this thesis where it says that machines simply do not have identical values and moral code as human beings.

5.3 Results/discussion

The theory of technological determinism mentioned in the conceptual framework of this thesis can be applied to critically analyse the beneficial and negative consequences that come with using GPT-3 model. In order to ask questions on whether the new fine/tuned InstructGPT is designed to address the complaints about toxic language and misinformation or whether it is developed to better follow human instructions and that it is better aligned with human intentions, one must look through the critical lenses of technological determinism.

Technological determinism, mentioned in the conceptual framework of this thesis, is an ideology that principally believes that technology has a mind on its own and that it will continue expanding and improving without any opposition from the society or governments (Thierer, 2018). The theory supposes that the autonomy of technology is above the autonomy of human beings, and this is where the issue arises. Technology is regarded as more powerful than human decision-making because people are helpless compared to the unstoppable technological advancements. This can be proven by looking at the Association for Computing Machinery's Digital Library, where 90% of the papers appeared to be associated with machine

autonomy (Calvo et al., 2020). This means that human autonomy is not the core focus of academics and researchers in computer science and other related spheres. Technological determinism signifies panic and fear towards any technological novelty which also includes AI and models like GPT-3 or InstructGPT. The theory is more frequently used for other purposes than to critically analyse AI ethics. It spread not only to the media and public as a way of justifying fear towards intelligent technology, but also to the mouths of AI ethicists that are concerned of the unknown processes. Hence, the AI industry and its further exploration of technological improvements are influenced by this ideology as well. The next sections of this thesis will concentrate on the dystopic and utopic view on the usage of GPT-3 and InstructGPT. Through the dystopic lenses, the debates on the models' lasting imminent harm will be explored. On the other hand, through the utopic lenses, the unjustifiable trust in models' effectiveness and industry's self-regulation will be analysed. These two technological perspectives are used to better understand in what manner technological determinism weakens human autonomy social control within technological contexts, which leads people to lose the capacity to decide accordingly to the circumstances and beliefs in the given situation.

5.3.1 Dystopic view

In order to answer the second research question, one must clearly understand the meaning of dystopic view in the context of language models like the GPT-3 and InstructGPT. To be clear, technological dystopianism, or as some call it the digital/cyber/algorithmic dystopia (Olds, 2017; Hudson, 2018; Kockelman, 2020), forces us to spot the ethical issues that arise when looking at technologies. The ideology warns people about the negative social disruptions that technologies might cause. In this way, one can argue that it is a very valuable and insightful method of analysing technological advancements that grow rapidly everyday. It reminds people how important it is to keep and protect their autonomy and privacy. Dystopic view of technology regards it as the main force that causes increased consumerism, dehumanisation, social division as well as human relocation (Kockelman, 2020; Zuboff, 2019). Therefore, in this dystopic view, LLM's like GPT-3 and InstructGPT are threatening human lives and society as a whole. The answer to the question of whether InstructGPT is less toxic than GPT-3, as OpenAI claims, would be no. Any technological innovation is seen as threat (Colman, 2005, p.284), and this technology in particular as it mimics human brain and human behaviour. As Ellul (2021) (hard determinist) asserted "technology is autonomous and uncontrollable force that dehumanises everything it touches". Hence, through dystopic view,

InstructGPT is in fact capable of more damage due to its advanced and finetuned powers. Dystopic view on GPT-3 spread to media, journal articles, ethical debates, and conferences to signify and inflate the ongoing anxiety and distress about impending destruction.

However, Müller (2020) drew attention to a very important verity, with which this thesis perfectly aligns its arguments. He emphasizes the fact that many academics, that work in ethics, and which have influence over policy-making, often overrate the influence, control, and danger that comes with new technology. In the sense that they, on the other hand, underrate how powerful and effective regulations by people can be (Ibid.). This can be clearly seen in the work of Floridi and Chiriatti (2020), who stress the undesirable consequences that come with the existence of GPT-3. As for instance, job market will be destructed, marketing will be AI-driven, nobody will recognise which text is written by the AI and which by human beings, and many other disadvantageous consequences (Ibid, p.691). They evaluated GPT-3 from the future perspective, where their solutions on the issue were of dystopic nature. They propose to form "a better digital culture" (Ibid., pp. 692–693) by making people more aware of the cyberspace and infosphere (information, data, knowledge, communication) in their everyday life. Moreover, they highlight the need of increasing human intelligence and critical thinking towards technologies like GPT-3. Nonetheless, even though they acknowledge the importance of raising awareness among humanity and having a legislative change, they deem the future will be swarmed with "semantic garbage" (Ibid., p.692) due to GPT-3 and its improved successor InstructGPT. This approach overtakes the early steps and stages of such technology and takes its focus only on what the future might hold in the worst possible scenarios. As Chan (2022) underlines, technology, in its early infancy, is open to reforms in social, political, or economic institution or practice. Furthermore, the authors refuse to notice the existing actions taken by the OpenAI to align with ethical norms. OpenAI (2022b) carefully re-examines requirements for liable use and sets rules for all computational developers. These are for example the mandatory implementation of safety measures where they need to always test and keep human in the loop.

For this reason, this thesis supports the opinion that many academic scholars like Floridi and Chiriatti (2020) are looking at GPT-3 from the dystopic point of view only, which leads them to believe in the worst even though it did not happen yet. They live in the future while they could have been spending their energy and knowledge to push forward more coherent policies and regulations. They are increasing the worry of people who are not educated enough

to understand the technology at the first place, not to say the consequences that these dystopic scholars display. Most importantly, this approach moves agency away from humans and pushes it more to the growing technology. It lets humanity forget about the existing human actors behind the technology as for instance AI developers, policymakers, civil society and those who use GPT-3 for spiteful and manipulative purposes to enhance model bias, privacy violations or spread misinformation through fake news. This thesis argues that it is the human agency that needs to be emphasized as the decision-making within AI results from it. Of course, the softer form of dystopic view on technology is helpful for the analysis so that all ethical concerns can be recognised, but the emphasis needs to be put on the pre-hoc regulatory responses in order to prevent feasible invasion of disinformation and manipulative processes. This is very well explained by McGuffie and Newhouse (2020, p-1), two scholars that investigate the possible weaponization of GPT-3 by right-wing extremists. They could have placed the main focus on the ways in which GPT-3 can be used for malpractice and endangerment, however, they accentuate the necessity to form regulations before the threats become reality. These contain the development of public policy, creation of appropriate social standards and the formal proposition of educational schemes. So the question of whether InstructGPT is less toxic than its predecessor GPT-3 or it only gained more power via its improvement, can be answered by looking at the human agency. The reason being that all of the failure modes of GPT-3 presented can be diminished by looking at human agency as the main factor of change.

5.3.2 Utopic view

The view of technology through utopic determinism means positive thinking when it comes to technological innovations and their place in the society. Through this lens, technological onward movement is portrayed as useful and rewarding to human life. As Leo Marx said: technology means progress in the “social, political, moral, and intellectual, as well as material“ (1987, p.34). This ideology is mostly supported by the power structures of technological industry, government, and military as technology is accepted as the “liberator” (Barbour, 1993). The one that is managed by humans to satisfy their needs and make their life more efficient and easier. In other words, through the utopic view, technology is seen as a device to be directed by humans for the purposes of autonomous action. The reason behind the rapid acceleration of AI and the demand for is that it gives people the ability to be more productive and enhances all qualities and aspects of life for all. Hence, if one looks notably at GPT-3, the promise this ideology follows is for example the potential distribution of its

qualities to a large number of people who are not experts on technologies. Public can use GPT-3 and draw upon its beneficial functionalities. Another positive aspect can be an easy and fast text generation, but there are much more advantages that utopic determinist believe GPT-3 models bring to the table (Chan, 2022).

However, this thesis is of opinion that utopic determinism is too soft when it comes to the analysis of autonomous technologies that can mimic human brain, behaviour, and can learn and act without any human involvement. Unlike the dystopic determinism, utopic view does not provide critical thinking towards new innovative technologies, and this results in overlooking ethical concerns. Hence, this thesis argues that utopic determinism towards models like the GPT-3 should be always challenged by dystopic view on technology. Because if the view does not confront new technologies with striking possibilities that might arise, regulations might be too narrow. The creation of regulations needs to be managed by also taking the negative dystopic views on GPT-3 into consideration. If it would not, there is a strong likelihood that the AI companies would self-regulate its processes, activities, and that they would gain the power to make up their own ethical issues.

As for example, Aggarwal et al. (2018) used GPT-3 itself to regulate fake news. One of the failure modes of GPT-3 that is presented in this thesis. They proposed finetuned GPT-3 and BERT as regulatory solutions for detection of fake news (Aggarwal et al., 2018, p.1). In their research, it appeared that their results were 97% accurate in categorising which news were fake or real. They used around 6thousand news articles in their test project from variety of different news sources where almost half was rated as fake and half as real (Chan 2022).

This utopic determinism's view on GPT-3 is regarded as too optimistic and illusory according to this thesis. It is arguing that LLM's like BERT or GPT-3 are nowhere near to be able to comprehend moral compass and narratives of the society. They are not fabricated with ethical structure as their underlying system nor they recognise failure modes that they ignite such as privacy violation, bias, and fake news. The thesis supports the view of Dignum (2017). She asserted that technologies like these are not ethical mechanisms and if such models are used for moral objectives, as for instance fake news detection, they have "operational morality" (Dignum 2017, p.3). This stands or the lowest level of ethical behaviour due to the fact that these models "do not have either autonomy nor social awareness and are not considered to be ethical systems" (Dignum 2017, p.3). However, they behave and act according to their

inventors and designers who are human tech developers and engineers, which backs up the argument that human involvement is the most powerful and very much needed when it comes to regulating models like GPT-3 and InstructGPT.

Through utopic determinism, many ethical concerns, and solutions, together with the failure modes mentioned in this thesis, are avoided and essentially non-existent. Academics and researchers (LaGrandeur, 2021; Aggarwal et al., 2018), that argue in favour of this view, believe that 3rd parties who stand behind such regulations do not comprehend nor worry about ethical concerns coming out from using AI in the first place. They claim regulations are too restrictive, in ways it should not be, because it only slows the technological growth that are designed to help society and accelerate the standard of living (LaGrandeur, 2021, p.6). LaGrandeur is of great belief that external parties, like the government laws and commissions, that have zero knowledge of complex technologies like GPT-3, only make the process counterproductive and irritating. This applies for both, the people in the tech industry as well as the whole humanity. External regulations, he says, should be of "last resort" (Ibid., p.6).

The argument of this thesis is to not completely evade the responsibilities of legislative nature from 3rdparties, but to intensify the importance of transparent and comprehensible AI algorithms. It opposes the optimistic utopic deterministic views like the ones of LaGrandeur (2021) or Aggarwal et al. (2018) due to weak regulatory proposals to address the failure modes that GPT-3 inflicts upon society. LaGrandeur (2021) and her belief that self-regulation is enough, in order to deal with the produced ethical and moral harm, is insubstantial. On the other hand, people should be careful in trusting tech companies to regulate its own creations that generate profit and opportunities. Tech industry will never willingly implement regulatory frameworks upon models like GPT-3 and InstructGPT by themselves, as it would only restrict the technological progress they are working for. Therefore, ethical concerns are usually set aside so that they can prosper and grow (Hagendorff, 2020, p.108). This argument denotes that the dependence upon tech industry to lessen failure modes of LLM's models via self-regulation is hindered by contradictory desires and purposes between ethics and benefits coming out of such technologies. The previously mentioned "operational morality" is the proof of the lack in transparency and human involvement in autonomous technology. As Nallur, Lloyd and Pearson (2021, p.4) claim "the presence of automation tends to make humans shed their cognitive engagement ". In other words, people, who have the power to shape technologies, are easily manipulated and influenced to change their moral direction and lean towards harmful

systems that strengthen the failure modes that GPT-3 and its successor create. The people in charge can be system developers, engineers, eternal regulators, or others who have the power to, for example depict fake news or recognise bias and privacy violation.

Hence, dystopic as well as utopic view on technological determinism cannot be applied to GPT-3 and InstructGPT alone. These views need to be combined with one important aspect. They need to be considered as part of human control, not as something separate that human beings cannot influence. “Contextualist view” (Barbour, 1993) is suitable perspective that this thesis supports. It offers different resolution to the failure modes of bias, privacy violation and fake news. Through this view, GPT-3 and IGPT are both seen as equivocal tools of social power that shape society depending on the context (Barbour, 1993, p.15). It uses critical lens, the same way as dystopic view, to provide valuable protection to individual autonomy and rights, but acknowledges the possible advantageous impacts towards social and ethical ends. All of this is, of course, underlined with a thoroughly planned design to shift the focus from GPT-3 and its harmful consequences to the responsibilities and positions of human actors. The idea is to form pre-emptive ethical action before the failure modes of GPT-3 and IGPT arise. These models are not the only actors that decide on their processes and tasks. In reality, there is an extensive variety of factors and initiators that manipulate and outline its routes and practices. According to Johnson and Verdicchio (2017, p.583) AI composes of “computational artefacts, human behaviour, and social arrangements “.

6 Conclusion

This thesis is not of technological nature, meaning it does not go in depth of the workings of InstructGPT explaining its processes from algorithmic and tech-savvy point of view. However, for the reader to better understand the context and for the thesis to better answer the research questions at hand, it used document analysis as its methodology. Therefore, empirical analysis is focusing on existing research on LLMs.

To conclude, this thesis argues that LLMs like GPT-3 are dangerous for the society when not addressed appropriately. The first research question of what the failure modes of LLMs from ethical, moral and security point of view are, was divided into three main sections.

Model bias (discrimination, exclusion, toxicity), violation of privacy and fake news. This followed with the second research question of whether IGPT is a solution to these failure modes. The answer is that there is a difficult human compliance journey ahead. These models need to be addressed from contextualist view on technological determinism to critically analyse such technologies. Human actors need to be seen as the ones in control and regulatory frameworks must follow the pre-emptive structure. Meaning that human involvement is crucial in improving LLMs processes and following regulations. There is a difficult human compliance journey ahead, but it is only us who can change it and address the failure modes that arise.

It depends on the responsibilities, positions, and agency of involved human actors. Overall, this thesis acknowledges the changes that were made after the introduction of GPT-3. InstructGPT presents improved processes that include human control through human-in-the-loop systems, which is exactly what should be done. However, IGPT is still in its infancy and needs improvement by looking at human agents more systematically. There indeed is a difficult human compliance journey ahead.

This thesis concentrates particularly on OpenAI and the company's promise of guaranteeing beneficial as well as safe AI to all of humanity (OpenAI, 2022).

The first section of the thesis introduces the topic by analysing what AI and Machine Learning represent and what the processes entail. The enhancement of computer power, to the extent of being able to mimic human brain via deep learning and natural processing, enlarged the possibility of solving variety of new tasks. These are for example the image recognition, machine translation, language modeling, time series prediction and many more. However, attention mechanisms were the ones that really modified Machine Learning. Their capacity to work with sensory information with attention allowed machines to focus one one exact issue at a time while organising it into a sequence of attention based reasoning tasks. Hence, attention mechanisms, self-attention mechanisms to be exact, are used in transformers due to their excellent competence to rate information according to its relevance interchangeably. Moreover, the thesis introduced LLMs, the successors of transformer models. GPT, presented by OpenAI in 2018, quickly became GPT-2 and GPT-3 that could autonomously generate text through training data obtained via high-quality web content. The models' ability to be fine-tuned for specialised assignments and use more and more learning parameters changed the course of

technology. There are many data science platforms that grant access to these models, but this thesis focuses on OpenAI as they were the first company to provide LLM API services. Microsoft, being the main investor, allowed GPT-3 to grow exponentially by using its Microsoft Azure's AI supercomputer. There is no need to have as many engineers or researchers involved in training data as it easily gets the data from web content such as the Wikipedia Corpus or Common Crawl. This can be viewed as either an advantage due to its ability to write news articles, predict last words of sentences by contextual recognition, send mass company email, create apps or layout tools, analyse search and data, generate text as well as program and its analysis, understand general reasoning and mathematics, translate text to various languages and more. However, these vast and powerful skills can also be viewed as dangerous and harmful for our society as humans are involved less and less. Although, some academics argue that models do not need safety precautions as they only limit AI's potential of future expansion. They state that the arguments to not have LLMs trained on public-facing data, where they learn the most through a good dialogue, are usually of unrealistic nature (like the robot apocalypse in a movie).

Therefore, this thesis sheds light on AI alignment problem and the importance of human involvement. Meaning that machines are unable to have identical values to human beings and need human guidance to understand context and diminish ethical concerns. Model bias, privacy violations and fake news are all failure modes that need to be addressed by human-machine teaming. This lets human operators refine inputs, assess model's outputs, and shape the automation of content generation or quality review. Another way of keeping human-in-the-loop is through model's fine-tuning that allows people to adjust the processes to achieve the desired performance.

Model bias like discrimination, exclusion, and toxicity are part of models' processes due to biased training data that is fed into them. They reflect bias already present in the world and intensify them even more. Race/religion/gender discrimination, exclusion of marginalised groups, language toxicity and hate speech are causing social, psychological, material or even physical harm. The model's output could truly be threatening and alarming as this thesis outlined by looking at one particular example of Microsoft's chatbot called Tay. Privacy violation, being the second failure mode, is another proof that human should be involved in the processes of LLMs. This violation can be caused by disclosure, inference, having access to inaccessible information and also by possible breach of post-privacy/post-mortem privacy.

Great examples where privacy was breached by LLMs are chatbot Lee Luna that disclosed private data of its users or "dead-bot" Samantha and Replica that were unable to admit ignorance and had inclination towards killing people. The third failure mode that is presented in this thesis is fake news as propaganda can be more targeted and can spread much easily. Hence, LLMs are not only capable of shaping public opinion but also generate fake news through audio-visual manipulation which is called cheap-fakes or deep-fakes. Targeted manipulation is much cheaper and more effective as models can produce illusory and deceitful media. This is where radicalisation risks arise as LLMs can depict one specific political view and feed political campaigns or violent extremist opinions. This can shape individuals into supporting far-right extremist ideologies as GPT-3 progresses in the generation of extremist text according to CTEC. There are many other issues that are in need of investigation but exist outside of the scope of this thesis as for example weaponization, environmental harms, plagiarism, authorship and others.

Hence, this thesis argues that Instruct GPT, finetuned model of OpenAI's GPT-3, is indeed designed to address the complaints about toxic language and misinformation, but it is not the solution to all failure modes. The great advancements are that it keeps human-in-the-loop via the use of reinforcement learning with human feedback to better align language models with human instructions. In this way, IGPT is better at the quality of writing, the knowledge on English language, it can also handle more complex instructions and is better at longer form content generation. However, it is still in its infancy. Although OpenAI lays out all of these beneficial factors, the OpenAI's alignment team also acknowledges that violent, sexual and other negative outputs are still present. Many academics argue IGPT is insufficient, unethical because it can be misused for various purposes and is trained on data that only 40 people provided their opinion on. Another critical thought is that it harms human psyche due to its dark personality and inability of detecting negative sides of human beings. AI models that process natural language are only trying to reflect on the learned patterns which means that they are not impeccable, and they do make simple mistakes. Hence, the main argument of this thesis is – yes InstructGPT is better equipped to address the failure modes that GPT-3 created; however, it is not the solution to all the ethical concerns. InstructGPT is a great start, but it is still a model that cannot be levelled to the behaviour and brain of a human being, simply because it does not have, nor it comprehends moral/ethical codes. Human involvement is needed to have these technologies in our lives without them damaging our society. Regulations needs to be set from the ground up – which means to start with the focus on AI developers,

engineers and people who are in charge of legislative protocols that control models' functionalities and reach.

Consequently, through technological determinism, one can answer the second research question of the thesis. The positive opinions (utopic view) as well as the opposing phobia (dystopic view) towards both, be it the GPT-3 or InstructGPT, can be analysed and guided via this critical ideology. This thesis is proposing a middle ground as both types of technological determinism are narrow and constrained when their approaches are applied to GPT-3 and InstructGPT. Technological determinism takes technologies like LLMs as an important, if not the main, impetus for social, political, economic or any type of change in our society and the way of living. Therefore, dystopic, and utopic determinisms regard LLMs as societal compass, where their unprecedented autonomous language processing and other capabilities shape human life. However, this thesis argues their approaches are too narrow and constrained. Dystopic view on technological determinism proposes negative approach which completely hides the possible advantages or possible changes that can be done for regulatory frameworks. It enhances fear by laying out potential harms and misuse that are many times only a speculation. Conversely, utopic view highlights the positive sides that LLMs bring to the society. This approach lacks critical thinking that often results in self-regulation by AI-industry that would most likely ignore any ethical concerns to let LLMs grow. This is especially concerning as the presented failure modes that GPT-3 ignites harm society and need to be addressed by dystopic approach on technologies. LLMs lack operational morality as they do not possess ethical mechanisms. Hence, dystopic as well as utopic view on technological determinism cannot be applied to GPT-3 and InstructGPT alone. These views need to be combined with one important aspect. They need to be considered as part of human control, not as something separate that human beings cannot influence. Contextualist view is, therefore, a suitable perspective that this thesis supports. This view asserts that technologies like GPT-3 and IGPT obviously change the course of life, but it depends on the context. It is not either positive or negative shift – it depends on the responsibilities, positions, and agency of involved human actors. Overall, this thesis acknowledges the changes that were made after the introduction of GPT-3. InstructGPT presents improved processes that include human control through human-in-the-loop systems, which is exactly what should be applied. However, IGPT is still in its infancy and needs improvement by looking at human agents more systematically. There indeed is a difficult human compliance journey ahead.

References

Aggarwal, A., Chauhan, A., Kumar, D., Mittal, M. and Verma, S. (2018). Classification of Fake News by Fine-tuning Deep Bidirectional Transformers Based Language Model. *ICST Transactions on Scalable Information Systems*, 0(0), p.163973. doi:10.4108/eai.13-7-2018.163973.

AI News (2023). *About AI News*. [online] AI News. Available at: https://www.artificialintelligence-news.com/about-us/.

Barbour, I.G. (1993). *Ethics in an Age of Technology*. Harper Collins.

Bender, E., Mcmillan-Major, A., Shmitchell, S., Gebru, T. and Shmitchell, S.-G. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, [online] 6(0), pp.610–623. doi:10.1145/3442188.3445922.

Bhavya, B., Xiong, J. and Zhai, C. (2022). Analogy Generation by Prompting Large Language Models: a Case Study of InstructGPT. *arXiv:2210.04186 [cs]*. [online] Available at: https://arxiv.org/abs/2210.04186.

Bimber, B. (1990). Karl Marx and the Three Faces of Technological Determinism. *Social Studies of Science*, 20(2), pp.333–351. doi:10.1177/030631290020002006.

Blodgett, S.L., Green, L. and O'Connor, B. (2016). *Demographic Dialectal Variation in Social Media: a Case Study of African-American English*. [online] Association for Computational Linguistics, pp.1119–1130. Available at: https://aclanthology.org/D16-1120.pdf.

Bostrom, N. (2011). Information Hazards: a Typology of Potential Harms from Knowledge. *Review of Contemporary Philosophy*, 10, pp.44–79.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C. and Hesse, C. (2020). *Language Models are Few-Shot Learners*. [online] Available at: https://arxiv.org/pdf/2005.14165.pdf.

Brownlee, J. (2017). *A Gentle Introduction to Long Short-Term Memory Networks by the Experts*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/?__cf_chl_tk=0UB0tSAdUn0FURMtCCA3Io7gu7jz3XTnKGjpZAv_LsE-1662548130-0-gaNycGzNCOU [Accessed 7 Sep. 2022].

Buchanan, B., Lohn, A., Musser, M. and Sedova, K. (2021). *Truth, lies, and Automation : How Language Models Could Change Disinformation*. Washington, DC: Center for Security and Emerging Technology.

Calvo, R.A., Peters, D., Vold, K. and Ryan, R.M. (2020). Supporting Human Autonomy in AI Systems: a Framework for Ethical Enquiry. *Philosophical Studies Series*, [online] 140, pp.31–54. doi:10.1007/978-3-030-50585-1_2.

Cao, Y. and Daumé, H. (2020). *Toward Gender-Inclusive Coreference Resolution*. [online] Available at: https://arxiv.org/pdf/1910.13913.pdf.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A. and Raffel, C. (2021). *Extracting Training Data from Large Language Models*. [online] Available at: https://arxiv.org/pdf/2012.07805.pdf.

Chan, A. (2022). GPT-3 and InstructGPT: Technological dystopianism, utopianism, and 'Contextual' Perspectives in AI Ethics and Industry. *AI and Ethics*. doi:10.1007/s43681-022-00148-6.

Colleoni, E., Rozza, A. and Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, 64(2), pp.317–332. doi:10.1111/jcom.12084.

Colman, A. (2005). *Un/Becoming Digital: the Ontology of Technological Determinism and Its Implications for Art Education*. *The Journal of Social Theory in Art Education*, pp.1–278.

Craft, J.T., Wright, K.E., Weissler, R.E. and Queen, R.M. (2020). Language and Discrimination: Generating Meaning, Perceiving Identities, and Discriminating Outcomes. *Annual Review of Linguistics*, 6(1), pp.389–407. doi:10.1146/annurev-linguistics-011718-011659.

Daws, R. (2020). *Medical Chatbot Using OpenAI's GPT-3 Told a Fake Patient to Kill Themselves*. [online] AI News. Available at: https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [online] Minneapolis, Minnesota: Association for Computational Linguistics, pp.4171–4186. doi:10.18653/v1/n19-1423.

Dickson, B. (2022). *OpenAI Is Reducing the Price of the GPT-3 API - Here's Why It Matters*. [online] VentureBeat. Available at: https://venturebeat.com/ai/openai-is-reducing-the-price-of-the-gpt-3-api-heres-why-it-matters/.

Dignum, V. (2017). *Responsible Autonomy*. [online] Available at: https://arxiv.org/pdf/1706.02513.pdf.

Dominguez, D. (2022). OpenAI Introduces InstructGPT Language Model to Follow Human Instructions. *InfoQ*. [online] Available at: https://www.infoq.com/news/2022/02/openai-instructgpt/.

Douglas, D.M. (2016). Doxing: a Conceptual Analysis. *Ethics and Information Technology*, 18(3), pp.199–210. doi:10.1007/s10676-016-9406-0.

Dutton, W.H. and Robertson, C.T. (2021). Disentangling Polarisation and Civic Empowerment in the Digital Age. In: *The Routledge Companion to Media Disinformation and Populism*. London: Routledge, pp.1–608.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In: *Theory of Cryptography*. New York: Springer, pp.265–284.

Ellul, J. (2021). *The Technological Society*. reprint ed. Knopf Doubleday Publishing Group, 2021, pp.1–512.

European Commission (2018). *Justice and Fundamental Rights*. [online] European Commission. Available at: https://ec.europa.eu/info/policies/justice-and-fundamental-rights_en.

Field, H. (2022). *Meet InstructGPT, OpenAI's Answer to Complaints about Toxic Language and Misinformation in GPT-3*. [online] Emerging Tech Brew. Available at: https://www.emergingtechbrew.com/stories/2022/01/31/meet-instructgpt-openai-s-answer-to-complaints-about-toxic-language-and-misinformation-in-gpt-3.

Flaxman, S., Goel, S. and Rao, J.M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1), pp.298–320. doi:10.1093/poq/nfw006.

Floridi, L. and Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), pp.681–694. doi:10.1007/s11023-020-09548-1.

Fortuna, P. and Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4), pp.1–30. doi:10.1145/3232676.

Garcia, D., Goel, M., Agrawal, A.K. and Kumaraguru, P. (2018). Collective Aspects of Privacy in the Twitter Social Network. *EPJ Data Science*, 7(1). doi:10.1140/epjds/s13688-018-0130-3.

GDPR.EU (2018). *Art. 17 GDPR - Right to Erasure ('right to Be forgotten')*. [online] GDPR.eu. Available at: https://gdpr.eu/article-17-right-to-be-forgotten/.

Gehman, S., Gururangan, S., Sap, M., Choi, Y. and Smith, N.A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *arXiv:2009.11462 [cs]*. [online] Available at: https://arxiv.org/abs/2009.11462 [Accessed 1 Jan. 2023].

Gopani, A. (2022). *Is InstructGPT Really Less Toxic as OpenAI claims?* [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/is-instructgpt-really-less-toxic-as-claimed-by-openai%EF%BF%BC/.

GPT-3 (2020). A Robot Wrote This Entire article. Are You Scared yet, human? *The Guardian*. [online] Available at: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3 [Accessed 13 Sep. 2022].

Griffin, A. (2019). *AI Deemed 'Too Dangerous to Release' Makes It out into the World*. [online] The Independent. Available at: https://www.independent.co.uk/tech/ai-artificial-intelligence-dangerous-text-gpt2-elon-musk-a9192121.html [Accessed 5 Sep. 2022].

Hagendorff, T. (2020). The Ethics of AI Ethics: an Evaluation of Guidelines. *Minds and Machines*, 30, pp.99–120. doi:10.1007/s11023-020-09517-8.

Hampton, S.C. (2019). *Parasite and Catalyst : the Polarizing Influence of Chatbots in Political Discourse*. [online] repositories.lib.utexas.edu. Available at: https://repositories.lib.utexas.edu/handle/2152/81204 [Accessed 5 Nov. 2022].

Hauer, T. (2017). Technological Determinism and New Media. *International Journal of English, Literature and Social Science (IJELS)*, [online] 2(2). Available at: https://mail.ijels.com/upload_document/issue_files/1%20IJELS-MAR-2017-8-Technological%20determinism%20and%20new%20media.pdf.

Heaven, W.D. (2022). *The New Version of GPT-3 Is Much Better Behaved (and Should Be Less toxic)*. [online] MIT Technology Review. Available at: https://www.technologyreview.com/2022/01/27/1044398/new-gpt3-openai-chatbot-language-model-ai-toxic-misinformation/.

Heilbroner, R.L. (1999). *The Worldly Philosophers*. [online] *Internet Archive*. Touchstone. Available at: https://archive.org/details/worldlyphilosoph00heil_2/page/239/mode/2up.

Hou, B. and Green, B.P. (2022). *A Multilevel Framework for the AI Alignment Problem*. [online] www.scu.edu. Available at: https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/a-multilevel-framework-for-the-ai-alignment-problem/.

Huang, Y. and Liang, C. (2015). A Comparative Study between the Dark Triad of Personality and the Big Five. *Canadian Social Science*, 11(1), pp.93–98.

Hudson, D.A. and Manning, C.D. (2018). Compositional Attention Networks for Machine Reasoning. In: *In Proceedings of the International Conference on Learning Representations (ICLR)*. Vancouver, BC, Canada: arXiv.

Hudson, L. (2018). *If You Want to Know How We Ended up in a Cyber dystopia, Read Ready Player One*. [online] The Verge. Available at:

https://www.theverge.com/2018/4/19/17250892/ready-player-one-book-facebook-internet-dystopia.

Jang, H. (2021). *A South Korean Chatbot Scandal Shows the Threat A.I. Presents to Privacy*. [online] Slate Magazine. Available at: https://slate.com/technology/2021/04/scatterlab-lee-luda-chatbot-kakaotalk-ai-privacy.html.

Jang, J., Ye, S. and Seo, M. (2022). *Can Large Language Models Truly Follow Your Instructions?* [online] Available at: https://openreview.net/pdf?id=89qDzjrWHLs [Accessed 31 Dec. 2022].

Jelinek, F. (1976). Continuous Speech Recognition by Statistical Methods. *Proceedings of the IEEE*, [online] 64(4), pp.532–556. doi:10.1109/PROC.1976.10159.

Jin, Z., Levine, S., Gonzalez, F., Kamal, O., Sap, M., Sachan, M., Mihalcea, R., Tenenbaum, J. and Schölkopf, B. (2022). When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment. *arXiv:2210.01478 [cs]*. [online] Available at: https://arxiv.org/abs/2210.01478.

Johnson, D.G. and Verdicchio, M. (2017). Reframing AI Discourse. *Minds and Machines*, 27(4), pp.575–590. doi:10.1007/s11023-017-9417-6.

Joshi, P., Santy, S., Budhiraja, A., Bali, K. and Choudhury, M. (2021). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *arXiv:2004.09095 [cs]*. [online] Available at: https://arxiv.org/abs/2004.09095 [Accessed 22 Sep. 2022].

Kantrowitz, A. (2016). *Microsoft's New AI-Powered Chatbot Mimics a 19-Year-Old American Girl*. [online] BuzzFeed News. Available at: https://www.buzzfeednews.com/article/alexkantrowitz/microsoft-introduces-tay-an-ai-powered-chatbot-it-hopes-will [Accessed 19 Sep. 2022].

Kaye, K. (2022). *OpenAI Launches New GPT-3 Model despite Continued Toxic Tendencies - Protocol*. [online] www.protocol.com. Available at: https://www.protocol.com/enterprise/openai-gptinstruct.

Khanzode, Ku.C.A. and Sarode, R.D. (2020). Advantages and Disadvantages of Artificial Intelligence and Machine Learning: a Literature Review. *International Journal of Library & Information Science (IJLIS)*, 9(1), pp.30–36.

Kockelman, P. (2020). The Epistemic and Performative Dynamics of Machine Learning Praxis. *Signs and Society*, 8(2), pp.319–355. doi:10.1086/708249.

Kumar, S. (2021). *Thoughts - Create Intelligent Thoughts*. [online] thoughts.sushant-kumar.com. Available at: https://thoughts.sushant-kumar.com/ [Accessed 13 Sep. 2022].

LaBerge, L., O'Toole, C., Schneider, J., Smaje, K. and McKinsey (2022). *COVID-19 Digital Transformation & Technology*. [online] www.mckinsey.com. Available at: https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever.

LaGrandeur, K. (2021). How Safe Is Our Reliance on AI, and Should We Regulate it? *AI and Ethics*, 1, pp.93–99. doi:10.1007/s43681-020-00010-7.

Lapowsky, I. (2017). *Eight Revealing Moments from the Second Day of Russia Hearings*. [online] Wired. Available at: https://www.wired.com/story/six-revealing-moments-from-the-second-day-of-russia-hearings/ [Accessed 6 Nov. 2022].

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep Learning. *Nature*, 521(7553), pp.436–444. doi:10.1038/nature14539.

Lee, P. (2016). *Learning from Tay's Introduction - the Official Microsoft Blog*. [online] The Official Microsoft Blog. Available at: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

Li, X., Li, Y., Liu, L., Bing, L. and Joty, S. (2022). *Is GPT-3 a Psychopath? Evaluating Large Language Models from a Psychological Perspective*. [online] Available at: https://arxiv.org/pdf/2212.10529.pdf.

Liévin, V., Hother, C.E. and Winther, O. (2022). Can Large Language Models Reason about Medical questions? *arXiv:2207.08143 [cs]*. [online] Available at: https://arxiv.org/abs/2207.08143 [Accessed 8 Nov. 2022].

Lin, S., Hilton, J. and Evans, O. (2021). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958 [cs]*. [online] Available at: https://arxiv.org/abs/2109.07958.

Makazhanov, A., Rafiei, D. and Waqar, M. (2014). Predicting Political Preference of Twitter Users. *Social Network Analysis and Mining*, 4(1). doi:10.1007/s13278-014-0193-5.

Mann, C.B. (2021). Can Conversing with a Computer Increase Turnout? Mobilization Using Chatbot Communication. *Journal of Experimental Political Science*, [online] 8(1), pp.51–62. Available at: https://ideas.repec.org/a/cup/jexpos/v8y2021i1p51-62_5.html [Accessed 5 Nov. 2022].

Marcus, G. (2018). *Deep Learning: a Critical Appraisal*. *arXiv*, New York: New York University, pp.1–27.

Markoff, J. and Mozur, P. (2015). *NewsDiffs | Article View*. [online] www.newsdiffs.org. Available at: Retrieved from: http://www.newsdiffs.org/article-history/www.nytimes.com/2015/08/04/science/for-sympathetic-ear-more-chinese-turn-to-smartphone-program.html [Accessed 19 Sep. 2022].

Marx, K. (1847). *The Poverty of Philosophy*. Book Jungle.

Marx, L. (1987). Does Improved Technology Means progress? *Technology Review*, pp.33–41.

McGuffie, K. and Newhouse, A. (2020). *The Radicalization Risks of GPT-3 and Advanced Neural Language Models*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/344261068_The_Radicalization_Risks_of_GPT-3_and_Advanced_Neural_Language_Models.

Mikalef, P., Conboy, K., Lundström, J.E. and Popovič, A. (2022). Thinking Responsibly about Responsible AI and 'the Dark Side' of AI. *European Journal of Information Systems*, 31(3), pp.257–268. doi:10.1080/0960085x.2022.2026621.

Moncur, W., Masthoff, J., Reiter, E., Freer, Y. and Nguyen, H. (2014). Providing Adaptive Health Updates across the Personal Social Network. *Human–Computer Interaction*, 29(3), pp.256–309. doi:10.1080/07370024.2013.819218.

Morrow, D.R. (2014). When Technologies Makes Good People Do Bad Things: Another Argument against the Value-Neutrality of Technologies. *Science and Engineering Ethics*, 20(2), pp.329–343. doi:10.1007/s11948-013-9464-1.

Müller, V.C. (2020). Ethics of Artificial Intelligence and Robotics. *plato.stanford.edu*. [online] Available at: https://plato.stanford.edu/archives/win2020/entries/ethics-ai/.

Nallur, V., Lloyd, M. and Pearson, S. (2021). Automation: an Essential Component of Ethical AI? *arXiv:2103.15739 [cs]*. [online] Available at: https://arxiv.org/abs/2103.15739 [Accessed 30 Dec. 2022].

Nast, C. (2021). *The Chatbot Problem*. [online] The New Yorker. Available at: https://www.newyorker.com/culture/cultural-comment/the-chatbot-problem.

Neff, G. and Nagy, P. (2016). Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, 10, pp.4915–4931. doi:1932–8036/20160005.

O'Sullivan, L. and Dickerson, J. (2020). *Here Are a Few Ways GPT-3 Can Go Wrong*. [online] TechCrunch+. Available at: https://tcrn.ch/30BQhqn.

Ogburn, W.F. (1922). Social Change: with Respect to Culture and Original Nature. *American Sociological Review*, 16(1). doi:10.2307/2087986.

Olds, J. (2017). *Digital Dystopia*. [online] American Scientist. Available at: https://www.americanscientist.org/article/digital-dystopia.

OpenAI (2022a). *About OpenAI*. [online] OpenAI. Available at: https://openai.com/about/.

OpenAI (2022b). *Aligning Language Models to Follow Instructions*. [online] OpenAI. Available at: https://openai.com/blog/instruction-following/.

OpenAI (2022c). *OpenAI API*. [online] beta.openai.com. Available at: https://beta.openai.com/docs/model-index-for-researchers.

OpenAI (2023). *GPT-4*. [online] openai.com. Available at: https://openai.com/product/gpt-4.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. and Lowe, R. (2022). Training Language

Models to Follow Instructions with Human Feedback. *arXiv:2203.02155 [cs]*. [online] Available at: https://arxiv.org/abs/2203.02155.

Papay, S., Waterbury, S. and Kaplan, R. (2022). *How Much Better Is OpenAI's Newest GPT-3 Model?* [online] ScaleAI. Available at: https://scale.com/blog/gpt-3-davinci-003-comparison [Accessed 31 Dec. 2022].

Paris, B. and Donovan, J. (2019). *Deepfakes and Cheap Fakes*. [online] Data & Society. Available at: https://datasociety.net/library/deepfakes-and-cheap-fakes/.

Parthasarathy, S. and Kleinman, M. (2021). *STPP Wins Grant to Explore Large Language Models*. [online] fordschool.umich.edu. Available at: https://fordschool.umich.edu/news/2021/stpp-wins-grant-explore-large-language-models [Accessed 4 Oct. 2022].

Persily, N. and Tucker, J.A. (2020). *Social Media and Democracy : the State of the field, Prospects for Reform*. Cambridge, United Kingdom: Cambridge University Press.

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D. and Ungar, L. (2017). *Beyond Binary Labels: Political Ideology Prediction of Twitter Users*. [online] ACLWeb. doi:10.18653/v1/P17-1068.

Quach, K. (2021). *A Developer Built an AI Chatbot Using GPT-3 That Helped a Man Speak Again to His Late fiancée. OpenAI Shut It down*. [online] www.theregister.com. Available at: https://www.theregister.com/2021/09/08/project_december_openai_gpt_3/ [Accessed 13 Sep. 2022].

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. [online] Available at: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

Rohrer, J. (2020). *On the Magic Potential and Bleak Future of GPT-3*. [online] Medium. Available at: https://medium.com/@jasonrohrer/on-the-magic-potential-and-bleak-future-of-gpt-3-ff7423ee38d4 [Accessed 13 Sep. 2022].

Roland, A., Smith, M.R. and Marx, L. (1994). Does Technology Drive History? the Dilemma of Technological Determinism. *The Journal of Military History*, 59(4). doi:10.2307/2944500.

Romero, A. (2021). *A Complete Overview of GPT-3 - the Largest Neural Network Ever Created*. [online] Medium. Available at: https://towardsdatascience.com/gpt-3-a-complete-overview-190232eb25fd.

Romero, A. (2022). *The New Version of GPT-3 Is Much, Much Better*. [online] Medium. Available at: https://towardsdatascience.com/the-new-version-of-gpt-3-is-much-much-better-53ac95f21cfb.

Roos, T. (2021). *https://twitter.com/teemu_roos/status/1382359296642588676*. [online] Twitter. Available at: https://twitter.com/teemu_roos/status/1382359296642588676 [Accessed 20 Sep. 2022].

Rosset, C. (2020). *Turing-NLG: a 17-billion-parameter Language Model by Microsoft*. [online] Microsoft Research. Available at: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/.

SAS (2018). *Artificial Intelligence – What It Is and Why It Matters*. [online] SAS. Available at: https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html [Accessed 6 Sep. 2022].

Schneier, B. (2020). *Bots Are Destroying Political Discourse as We Know It*. [online] The Atlantic. Available at: https://www.theatlantic.com/technology/archive/2020/01/future-politics-bots-drowning-out-humans/604489/.

Scott, K. (2020). *Microsoft Teams up with OpenAI to Exclusively License GPT-3 Language Model*. [online] The Official Microsoft Blog. Available at: https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/ [Accessed 9 Nov. 2020].

Sigmoid (2020). *GPT-3: All You Need to Know about the AI Language Model*. [online] Sigmoid. Available at: https://www.sigmoid.com/blogs/gpt-3-all-you-need-to-know-about-the-ai-language-model/.

Sinders, C. (2016). *Microsoft's Tay Is an Example of Bad Design*. [online] Medium. Available at: https://medium.com/@carolinesinders/microsoft-s-tay-is-an-example-of-bad-design-d4e65bb2569f.

Sutskever, I., Vinyals, O. and Le, Q.V. (2014). Sequence to Sequence Learning with Neural Networks. *In Proceedings of the NIPS 2014*.

TechForgeMedia (2023). *About Us*. [online] TechForge Media Ltd. Available at: https://www.techforge.pub/about-us/ [Accessed 1 May 2023].

Thierer, A. (2018). *The Pacing Problem, the Collingridge Dilemma & Technological Determinism*. [online] Technology Liberation Front. Available at: https://techliberation.com/2018/08/16/the-pacing-problem-the-collingridge-dilemma-technological-determinism/.

Tiffany, K. (2022). *'Doxxing' Means Whatever You Want It to*. [online] The Atlantic. Available at: https://www.theatlantic.com/technology/archive/2022/04/doxxing-meaning-libs-of-tiktok/629643/.

Tilman, R. (1990). New Light on John Dewey, Clarence Ayres, and the Development of Evolutionary Economics. *Journal of Economic Issues*, 24(4), pp.963–979. doi:10.1080/00213624.1990.11505096.

Ugli, M.I.B. (2020). Will Human Beings Be Superseded by Generative pre-trained Transformer 3 (GPT-3) in programming? *International Journal on Orange Technologies*, 2(10), pp.117–143.

Unbabel (2019). *Gender Bias in AI: Building Fairer Algorithms*. [online] resources.unbabel.com. Available at: https://resources.unbabel.com/blog/gender-bias-artificial-intelligence [Accessed 13 Sep. 2022].

Unbabel (2020). *Behind the GPT-3 Buzz: Why Human-in-the-Loop AI Is Important*. [online] resources.unbabel.com. Available at: https://resources.unbabel.com/blog/behind-the-gpt-3-buzz-why-human-in-the-loop-ai-is-important [Accessed 12 Sep. 2022].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017). Attention Is All You Need 2. In: *31st Conference on Neural Information Processing Systems*. Long Beach, CA, USA: NIPS.

Veblen, T. (2015). *Why is Economics not an Evolutionary Science*. Read Books Ltd.

Verma, P. and Lerman, R. (2022). *What Is ChatGPT, the Viral Social Media AI?* [online] Washington Post. Available at: https://www.washingtonpost.com/technology/2022/12/06/what-is-chatgpt-ai/.

Vig, J. (2019). *A Multiscale Visualization of Attention in the Transformer Model*. [online] ACLWeb. doi:10.18653/v1/P19-3007.

Wallace, E., Tramèr, F., Jagielski, M. and Herbert-Voss, A. (2020). *Does GPT-2 Know Your Phone Number?* [online] The Berkeley Artificial Intelligence Research Blog. Available at: https://bair.berkeley.edu/blog/2020/12/20/lmmem/.

Wang, Y. (2016). *Your next New Best Friend Might Be a Robot*. [online] Nautilus. Available at: https://nautil.us/your-next-new-best-friend-might-be-a-robot-rp-235778/ [Accessed 19 Sep. 2022].

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. and Isaac, W. (2021). *Ethical and Social Risks of Harm from Language Models*. [online] DeepMind, pp.2–64. Available at: https://arxiv.org/pdf/2112.04359.pdf.

Wiggers, K. (2020). *OpenAI's Massive GPT-3 Model Is impressive, but Size Isn't Everything*. [online] VentureBeat. Available at: https://venturebeat.com/ai/ai-machine-learning-openai-gpt-3-size-isnt-everything/ [Accessed 7 Sep. 2022].

Wiggers, K. (2021). *Propaganda-as-a-service May Be on the Horizon If Large Language Models Are Abused*. [online] VentureBeat. Available at: https://venturebeat.com/ai/propaganda-as-a-service-may-be-on-the-horizon-if-large-language-models-are-abused/.

Woodward, J.W. (1934). Critical Notes on the Culture Lag Concept. *Social Forces*, 12(3), pp.388–398. doi:https://doi.org/10.2307/2569930.

Woolley, S. (2020). *We're Fighting Fake News AI Bots by Using More AI. That's a mistake.* [online] MIT Technology Review. Available at: https://www.technologyreview.com/2020/01/08/130983/were-fighting-fake-news-ai-bots-by-using-more-ai-thats-a-mistake/.

Wu, J., Ouyang, L., Ziegler, D.M., Stiennon, N., Lowe, R., Leike, J. and Christiano, P. (2021). Recursively Summarizing Books with Human Feedback. *arXiv:2109.10862 [cs]*. [online] Available at: https://arxiv.org/abs/2109.10862.

Xu, A.Y. (2020). *Language Models and Fake News: the Democratization of Propaganda*. [online] Medium. Available at: https://towardsdatascience.com/language-models-and-fake-news-the-democratization-of-propaganda-11b1267b3054 [Accessed 30 Oct. 2022].

Ye, X. and Durrett, G. (2022). *The Unreliability of Explanations in Few-shot Prompting for Textual Reasoning*. [online] Available at: https://openreview.net/pdf?id=Bct2f8fRd8S [Accessed 31 Dec. 2022].

Zhang, Y., Shao, K., Yang, J. and Liu, H. (2021). Universal Adversarial Attack via Conditional Sampling for Text Classification. *Applied Sciences*, 11(20), p.9539. doi:10.3390/app11209539.

Zhao, T.Z., Wallace, E., Feng, S., Klein, D. and Singh, S. (2021). *Calibrate before Use: Improving Few-Shot Performance of Language Models*. [online] arXiv. Available at: https://arxiv.org/pdf/2102.09690.pdf.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: the Fight for the Future at the New Frontier of Power*. London: Profile Books.