Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December: 1–14

DOI: 10.1177/20539517211047734
journals.sagepub.com/home/bds
SAGE

The great Transformer: Examining the role of large language models in the political economy of AI

Dieuwertje Luitse and Wiebke Denkena

Abstract

In recent years, AI research has become more and more computationally demanding. In natural language processing (NLP), this tendency is reflected in the emergence of large language models (LLMs) like GPT-3. These powerful neural network-based models can be used for a range of NLP tasks and their language generation capacities have become so sophisticated that it can be very difficult to distinguish their outputs from human language. LLMs have raised concerns over their demonstrable biases, heavy environmental footprints, and future social ramifications. In December 2020, critical research on LLMs led Google to fire Timnit Gebru, co-lead of the company's AI Ethics team, which sparked a major public controversy around LLMs and the growing corporate influence over AI research. This article explores the role LLMs play in the political economy of AI as infrastructural components for AI research and development. Retracing the technical developments that have led to the emergence of LLMs, we point out how they are intertwined with the business model of big tech companies and further shift power relations in their favour. This becomes visible through the Transformer, which is the underlying architecture of most LLMs today and started the race for ever bigger models when it was introduced by Google in 2017. Using the example of GPT-3, we shed light on recent corporate efforts to commodify LLMs through paid API access and exclusive licensing, raising questions around monopolization and dependency in a field that is increasingly divided by access to large-scale computing power.

Keywords

Artificial intelligence, algorithmic techniques, Transformer, large language models, monopolization, platforms

Introduction

Over the past decade, artificial intelligence (AI) technologies have evolved rapidly.[1] Much of the public acclaim and many commercial breakthroughs have been due to deep learning, a specific methodology for machine learning in which complex neural networks are trained using large amounts of data (LeCun et al., 2015). Deep learning has spurred activity in different AI subfields like Computer Vision or natural language processing (NLP), resulting in sophisticated commercial products for applications such as facial recognition or automated language generation. In response, a growing body of research focuses on the social impact, ethics, and regulation of machine learning systems (e.g. Benjamin, 2019, Amoore, 2020). Regarding NLP, such scholarly work is devoting attention to the study of large language models (LLMs). Technically, a language model is a statistical representation of a language, which tells us the likelihood that a given sequence (a word, phrase, or sentence) occurs in this language.[2] Due to this capacity, language models can be used to make predictions about how a sentence might continue and, consequently, to generate text. Sophisticated language models, often based on neural networks and large text corpora, are very powerful because they can be used in a wide range of different applications such as translation or text recognition. Their increasingly convincing (that is to say human-like) outputs have raised concerns about future social ramifications, from automated text production in areas such as journalism or advertising, allowing for even more personalization and the solidification of filter bubbles (Floridi and Chiriatti, 2020), to discrimination

Department of Media Studies, University of Amsterdam, Amsterdam, the Netherlands

Corresponding author:
Dieuwertje Luitse, Department of Media Studies, University of Amsterdam, Turfdraagsterpad 9, 1012 XT, Amsterdam, The Netherlands.
Email: dieuwertje.luitse@student.uva.nl

(Abid et al., 2021) and risks stemming from malicious actors who could use language models to automatically generate fake news or violence-inciting posts on social media at scale (McGuffie and Newhouse, 2020).

In December 2020, research into the social implications of LLMs was entangled in a major controversy as Google terminated its AI ethics co-lead Dr Timnit Gebru for refusing to retract her name from a paper that turned a critical eye on the development of LLMs and their associated risks (Hao, 2020a). The paper in question discusses the financial and environmental cost of English LLMs as well as social implications, for example stemming from the encoding of 'stereotypical and derogatory associations along with gender, race, ethnicity and disability status' (Bender et al., 2021: 612). The authors emphasize that the corporate push toward LLMs is taking resources away from other research efforts in the field, despite the fact that it remains 'far from clear' that LLMs bring researchers 'any closer to long-term goals of general language understanding systems' (Bender et al., 2021: 616). The controversy over Google's termination of Dr Gebru, and the firing of Dr Margaret Mitchell a few weeks later, is said to have incited a debate around 'corporate influence over AI' (Hao, 2020a) and the immense power of big tech companies.

In fact, beyond siphoning off most talent in the field by offering unrivalled salaries (Lee, 2018), the last years have seen significant infrastructural investments of major platform corporations across different fields of AI. This includes the extension of cloud computing platforms like Google Cloud, Amazon's AWS, and Microsoft Azure. Enhanced by expenditures in data centres, specialized hardware such as processor chips, or undersea cables, these corporate clouds provide top-to-bottom solutions for data management, analytics, dynamic scaling, and machine learning that are 'increasingly difficult to replicate without considerable investment' (Rieder, 2020: 119). In a similar vein, Google has open-sourced its machine learning framework TensorFlow and released a powerful open-source language model called Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Both are widely used and have shaped the development and application of machine learning systems worldwide. It is infrastructural developments like these that allow platform corporations to leverage their assets, resulting in a concentration of market and political power (monopolization) in their hands (Srnicek, 2018).

The recent concerns over ramifications of LLMs (Bender et al., 2021) call for a closer investigation of large tech corporations' infrastructural ambitions within the subfield of NLP, focussing on the role these language models play in the political economy of AI (Srnicek, 2018; Dyer-Witheford et al., 2019). This article contributes to such an inquiry through a historical account and a critical analysis of recent technical developments in NLP, leading up to the release of the language model GPT-3. Developed by AI research company OpenAI, GPT-3 can be considered one of the largest and most sophisticated English language models today, and it is currently only accessible through a closed-off API. GPT-3 and other recent LLMs are based on the Transformer network architecture (Vaswani et al., 2017), an algorithmic technique for the training of neural networks developed by Google, which allows large parts of the computations to be executed in parallel, and hence, a significant acceleration of training speed. The introduction of the Transformer in NLP has caused a tendency towards ever bigger pretrained language models, which correlate with higher performance according to certain measures, such as the convincingness or 'human-likeness' of text (Brown et al., 2020). In this article, we shed light on the Transformer and its impact on NLP to argue that the growing size of language models accelerates monopolization in the digital economy, even though companies like Google or OpenAI seem to take different strategies with respect to their models. Due to the applicability of LLMs for a wide variety of tasks in different societal contexts and the increasing prevalence of the technology in working algorithmic systems, including search engines, conversational agents, and third-party applications, they offer a meaningful case study to shed light on the inherent accumulative tendency of capital and corporate ownership in the context of AI, and the subsequent corporate control and power over crucial AI infrastructures, which are associated with a number of societal risks.

Taking a combined political economy and historical software studies approach, this argument is developed in three steps. First, the paper lays out the tendencies towards monopolization that exist within digital capitalism and AI in particular, most clearly observable at the business model of the digital platform. Second, we consider the historical development of the Transformer through a discussion of the key literature that has informed the invention of the architecture. Here, we loosely follow Rieder's (2020) concept of algorithmic techniques to formulate 'an approach to the analysis of software that sits between broad theorizing and the empirical investigation of concrete applications' (Rieder, 2020: 100). In Rieder's conceptualization, algorithmic techniques are technical schemata employed by programmers to respond to typical, widespread problems such as sorting a list or making a 'good' recommendation. From this perspective, the Transformer network architecture can be considered an algorithmic technique responding to a variety of language tasks, which, in its derivative and solidified form as a large language model, enters working systems and applications from diverse contexts: 'Spelled out, stabilized, and 'frozen', algorithmic techniques spread through technical imaginaries and artefacts, and further into application logics and business models.' (Rieder, 2020: 16). The third section discusses the role of LLMs in the political economy of AI. By

comparing OpenAI's model release strategy to earlier open-source release approaches, we specifically consider how, with the release of GPT-3's closed-off API, the company attempts to strengthen its powerful position in the field of NLP and the wider digital economy.

The political economy of AI

Our inquiry contributes to scholarship on the digital political economy (Wittel, 2017) and the emerging discussion of the political economy of AI (Srnicek, 2018; Dyer-Witheford et al., 2019). Spawning from critical media studies and neighbouring disciplines, prominent theoretizations of 'digital capitalism' (Srnicek, 2017; Staab, 2019; Zuboff, 2019) outline the specific dynamics of corporate competition and the consolidation of power in the digital era. Despite their differences in focus and analysis, these theoretical frameworks share important insights pertaining to monopolization: crucially, the concept of the digital platform has highlighted how large tech corporations position themselves as intermediaries in a network of different actors, allowing them to extract data, harness network effects, and approach monopoly status. Beyond such overarching frameworks, several scholars have used the political economy lens to point to issues like commodification or corporate concentration in cultural production (Nieborg and Poell, 2018), the mobile ecosystem (Nieborg and Helmond, 2019), or web search (Rieder and Sire, 2014).

While this scholarship has generated important insights, the specific capitalist dynamics implied by AI remain understudied. Laying the foundations for a Marxist political economy of AI, Dyer-Witheford et al. (2019) discuss the ramifications and prospects of AI-induced capitalism. While a large portion of their work is dedicated to questions of labour and automation, the inquiry into the current state of the 'AI industry' is insightful with regards to the arguments advanced in this paper. The authors introduce the notion of 'actually existing AI capitalism'—in contrast to a speculative future realization of 'AI capitalism'—to describe the present moment, which is characterized by the industrial distribution of 'narrow AI' in working systems, distributed through 'the cloud'. In this context, big tech corporations benefit from their control of data, talent and computing infrastructure, enabling them to solidify their dominant position. Disseminating narratives about the alleged 'democratization of AI,' these companies pursue the 'creation of a generalized AI infrastructure for advanced capital' (Dyer-Witheford et al., 2019: 52) under their control.

In very similar terms, Nick Srnicek (2018, 2019, 2020) argues that the success of machine learning, and deep learning in particular, further accelerates the dynamics characteristic of digital platforms, such as 'the insatiable appetite for data and the monopolizing dynamic of network effects' (Srnicek, 2018: 157). Following his analysis, however, data is no longer the most important driver towards monopolization because data extraction has been adopted as a business model across the wider economy, distributing the resource among a growing number of actors. This is facilitated by an increase in the availability of 'Open Data,' and the rise of synthetic data for the production of ML models (Srnicek, 2019). Instead of data, computing power ('compute') has instead become more important in AI competition. Aggregating huge amounts of computing power not only require large amounts of capital to invest in or rent the necessary hardware, like powerful graphics processing units (GPUs) or expensive purpose-built chips, but also professionals that possess the skills and the experience to operate complex neural networks on large clusters of hardware. Consequently, only a small number of institutions and companies have the means to keep up this kind of compute-intensive research and produce large-scale models. Srnicek (2019) therefore predicts that the few AI companies that persist in this highly competitive environment will become massive global rentiers through their AI infrastructures: smaller companies looking to adopt ML but cannot afford to train their own models, nor to run a sophisticated neural network on their own hardware, are pushed towards the infrastructures of big corporations that already provide polished cloud solutions for AI.

The increasing importance of compute and its implications for the field of AI has also been recognized by other authors. To be precise, compute has always played a major role in the history of AI, being one of the reasons why research into neural networks was hindered for a long time. But nowadays compute has become so central that it is shifting power relations in the field. Based on an empirical investigation into corporate presence at important AI-related computer science conferences since the rise of deep learning in 2012, Ahmed and Wahed (2020) observe that the unequal access to computing power creates advantages for big tech corporations and elite universities who slowly crowd out other members of the research community such as mid- and lower tier universities. They introduce the idea of a 'compute divide' which 'increases concerns around bias and fairness within AI technology, and presents an obstacle towards 'democratizing' AI' (Ahmed and Wahed, 2020: 1). Similarly, Hwang (2018) argues that the importance of compute in AI shifts attention to hardware: 'The medium is a significant message here: hardware actively shapes the landscape of what can be done with the technology of machine learning, and plays a significant role in influencing how it will evolve going forwards' (Hwang, 2018: 7). While AI research has widely switched from universal central processing units (CPUs) to GPUs, which allow for increased parallelization since the late 2000s (Raina et al., 2009), the question now becomes who can further scale up their compute capacity, for example, by acquiring purpose-built hardware such as application-specific integrated circuits (ASICS) or field-

programmable gate arrays (FPGAs). These specialized chips have the potential to significantly improve performance but come at a high cost, in terms of financial investment and—depending on the level of specialization—lack of flexibility (Hwang, 2018).

Building on these observations about the growing economic and political importance of compute, this paper investigates the political economy of AI through the example of the emergence and commodification of (English language) LLMs. It provides a technically-informed case study about the dynamics in a specific subfield of AI, thereby concretizing and extending the macro-level analyses by Dyer-Witheford et al. (2019) and Srnicek (2018, 2019, 2020). The next section gives a brief account of the developments in AI research since the 1950s that have led to the Transformer network architecture (Vaswani et al., 2017), which forms the foundation for the creation of large pretrained models in NLP. This contextualization helps understand the Transformer as part of the currently dominant connectionist AI paradigm which is heavily data-driven, requires large amounts of computing power, and is mainly executed by a small number of actors worldwide. Lastly, we use the case study of GPT-3 to discuss how LLMs are commodified in a software-as-a-service model. As rentiers and gatekeepers of crucial components in the development of NLP systems, large tech companies provisioning these language models concentrate both economic and political power in their hands.

The great Transformer

Prehistory of the Transformer architecture

The field of 'Artificial Intelligence,' a term coined by John McCarthy in 1956, broadly knows two approaches: symbolic AI ('Good Old-Fashioned AI') and connectionist AI (cf. Cardon et al., 2018). Simply put, the symbolic approach attempts to build intelligent systems by modelling knowledge and logical reasoning, and encoding both in the form of facts (knowledge base) and rules of transformation (Woolridge, 2020). In contrast, connectionist AI attempts to approximate the structure of the human brain and its network of biological neurons in order to make machines 'learn' and eventually reach 'intelligent' behaviour. This approach is theoretically rooted in cybernetic thought of the 1940s (McCulloch and Pitts, 1943) and was first implemented in 1957 when Frank Rosenblatt built a pioneering artificial neural network (the 'perceptron'), but was long disregarded by the more prestigious symbolic AI community. Combined with the influence of the critical volume *Perceptrons* (Minsky and Papert, 1969), which described the then-prevailing limitations of neural networks, this was one of the reasons why connectionist research was hampered in the 1970s.

It was not until the emergence of the Internet, large data sets, and more powerful computing hardware that it became worthwhile to work with neural networks in practical applications for the first time, resulting in the wide uptake of backpropagation, a technique to more efficiently train neural networks, which had already been around since the 1980s (Rumelhart et al., 1986). Subsequently, AI slowly shifted from a symbolic, knowledge-based approach to be increasingly data-driven. Instead of aiming for universally applicable 'intelligent' systems, research became much more task-specific ('narrow AI') (Wooldridge, 2020). Throughout the years, many systems were trained based on different data sets, experimenting with different network architectures, and different parameters. Parameters are the variables that neural networks continuously optimize during training to minimize error. High numbers of parameters often correlate with more subtlety (Kaplan et al., 2020): for example, the more parameters a neural network-based language model has, the more fine-grained its language generation capacity. For this reason, a trend towards neural networks with ever more neurons and larger numbers of highly interconnected layers developed, summarized under the term deep learning (LeCun et al., 2015). The current hype around neural networks really took off in 2012, when deep learning proved successful at the ImageNet Large Scale Visual Recognition Challenge (Krizhevsky et al., 2012). Although neural networks have evolved significantly and gained in complexity since the 1950s, they still only very remotely resemble the complex human brain. Nevertheless, deep neural networks have proven to be highly efficient for a variety of tasks (LeCun et al., 2015). Beyond image recognition, deep learning has been rapidly taken up in other fields, such as NLP. As a consequence, connectionist AI has in many areas outcompeted symbolic approaches that were dominant during the first 30 years of AI research.

The Transformer (Vaswani et al., 2017) is one of the latest iterations in the history of connectionist AI. Essentially, the Transformer is an architecture for neural networks which was designed to efficiently handle sequential data (e.g. sentences). In NLP, the architecture has led to new heights in model size and performance because it allowed for significantly increased parallelization and reduced training times. It thus became the go-to architecture in many areas, often replacing complex recurrent or convolutional neural networks with a much more efficient setup. As this paper will argue, the architecture has changed the NLP landscape by making scale, and the computing power to achieve this, crucial factors in the development of industrial systems. As emphasized before, computing power has always played an important role in AI but the recent turn to scale for the improvement of systems is shifting power relations in the field. This has an impact on AI deployment and provision which are, as we attempt to show, increasingly centralized.

In contrast to previous approaches in software studies that have focussed on foundational principles of computation (Fuller, 2008) or the analysis of concrete implementations of algorithms in the form of source code (Marino, 2020), we consider Transformer networks through the lens of algorithmic techniques (Rieder, 2020). The Transformer architecture was indeed developed with the goal to enable a language model to better process longer sentences in a shorter amount of time. As a well-documented algorithmic technique, it quickly entered the reservoir of available methods for deep learning, and NLP in particular. In this article, we aim to exemplify the important insights that a conceptual focus on algorithmic techniques as the conceptual and material building blocks of concrete software, affords in critical media and technology scholarship.

The emergence of Transformers

Before the introduction of Transformer networks, NLP researchers used other neural network architectures such as recurrent neural networks (RNNs) (Rumelhart et al., 1986) and convolutional neural networks (CNNs) (Krizhevsky et al., 2012) to generate language models. RNNs, in particular long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), gained prominence as they are useful for tasks involving language modelling, like machine translation (Sutskever et al., 2014) and predictive text generation (Graves, 2013). For example, smartphones usually have an autocomplete function that enables them to predict the word a user is typing and make a suggestion for the next word that could potentially follow. Typing 'good,' a phone might suggest 'morning' as the next word. If we follow the suggestion and type 'morning,' another suggestion is made (e.g. 'my'). This functionality could be achieved using an RNN.[3] In simple terms, this architecture consists of multiple instances of the same set of artificial neurons, where the output of the first is fed into the second, and so forth (see Figure 1). In our example, 'good' would be the input of the first instance, which would then return 'morning' as the most likely follow-up term. 'Morning' would then be passed to the second set of neurons, the output of which would be 'my'. To make plausible predictions for the next word, the different instances of the neural network need to 'memorize' the previously typed words, at least to some extent. Without this 'memory', RNN-based language models could only ever predict the next word based on the previous one, resulting in implausible outputs. The 'memory' of an RNN is implemented in the form of a fixed-length vector. During training, the RNN learns how to best adapt this vector in order to store information about preceding words. After training, the network can be run (inference) and used for different applications (e.g. autocomplete). The trained RNN receives an input sentence (represented in the form of an input vector) and 'encodes' it into a 'memory' vector that it uses to predict the next word (Figure 1). In a second step, the vector, which now contains the information the network has successfully captured about the input sentence, is fed into another instance which 'decodes' it and predicts the next word on the base of it.

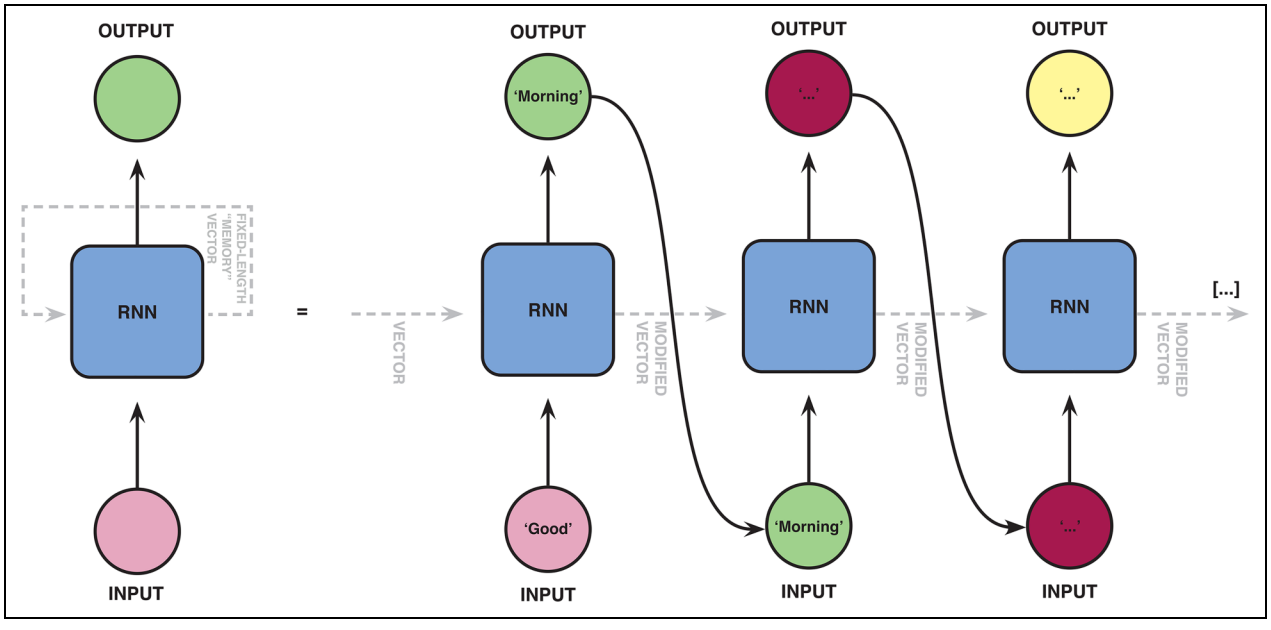

Figure 1. The recurrent neural networks (RNNs) are characterized by chains of instances of the same neural net along which the input phrase is passed sequentially (encoding) before a prediction can be made (decoding).

However, due to the mechanism of backpropagation—the algorithm used for training the network by adapting the weights in order to minimize error—RNNs can usually not remember every word, especially as sentences get longer. This is the so-called problem of 'vanishing gradients' (Bengio et al., 1994) that keeps a model from memorizing long data sequences, making it more difficult to train that model (Pascanu et al., 2013). For example, in order to determine the right pronoun to follow a phrase like 'I called my parents to congratulate' the RNN is processing every word between 'my parents' and the last word of the phrase (Figure 2). The longer a sentence, the more difficult it tends to become for the RNN to predict the next word based on its predecessors. In 1997, the LSTM network architecture was introduced to tackle this problem (Hochreiter and Schmidhuber, 1997). However, LSTM-based model performance still degrades the longer the processed sequence gets (Hernandéz and Amigó, 2020). Besides vanishing gradients, RNNs also pose problems in terms of computing capacity: Due to the chain of instances that makes up a recurrent neural network, data is necessarily processed sequentially. Each step of the computation is dependent on the previous one. Therefore, RNNs are hard to parallelize, and working with them quickly gets time-intensive.

In their attempt to improve model performance, Bahdanau et al. (2014) and Luong et al. (2015) introduced and refined a technique called 'Attention'. Extending the encoder–decoder

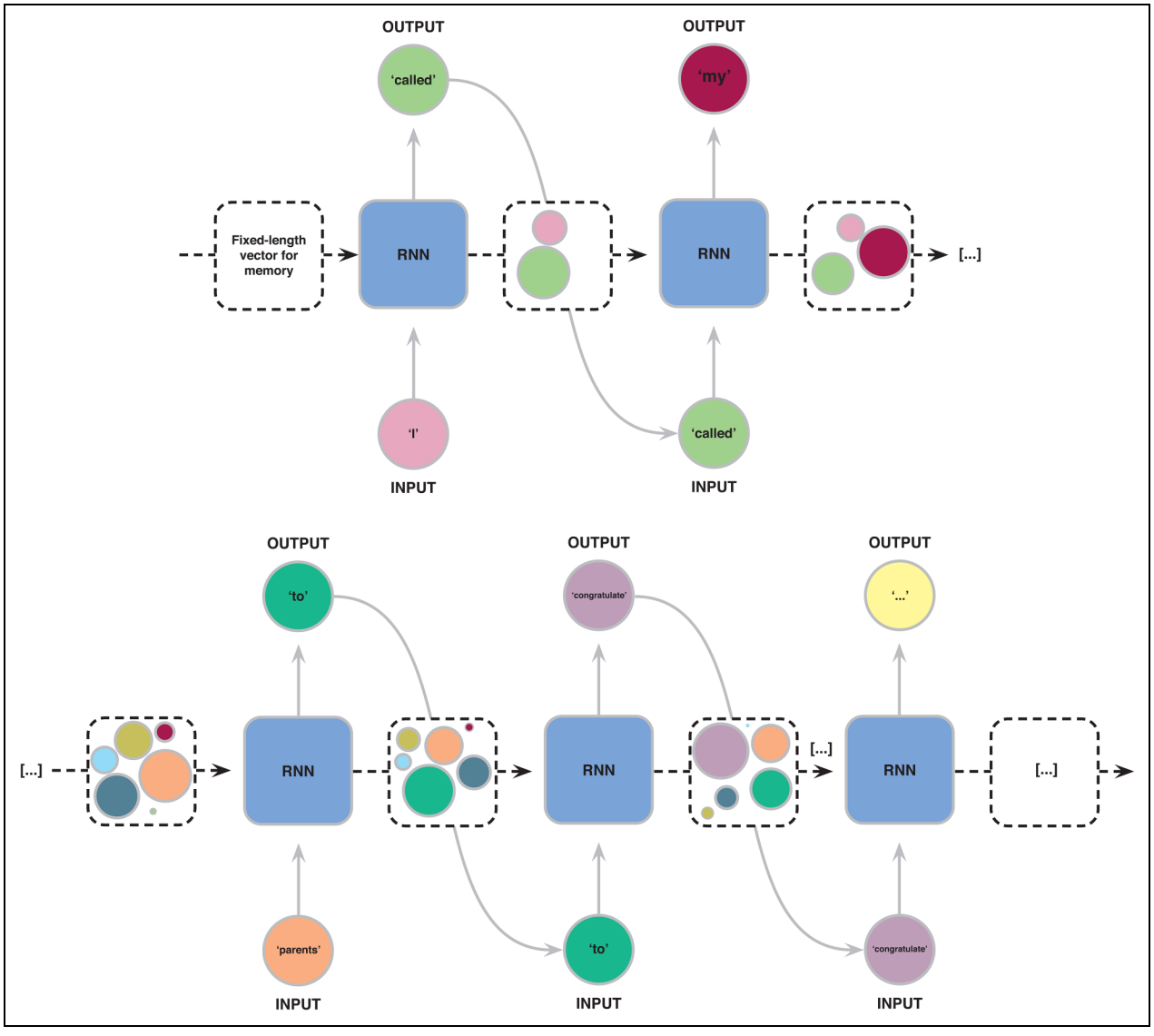

Figure 2. The recurrent neural networks (RNNs) learn to adapt a fixed-length vector in order to store information about the previously processed symbols of the input sequence. However, the vector cannot capture enough information about longer input sequences. This is why early words in longer input sequences tend to be 'forgotten' when the end of the sequence is processed.

structure familiar from classic RNNs and LSTMs, Attention allows a model to focus on specific, relevant parts of the input sequence when performing a text prediction task. Rather than encoding the entire input sentence into a fixed-length vector, an Attention-based architecture enables the encoding of the input sentence into a sequence of 'memory' vectors. During training, the model is optimized to capture information about the input sentence across these multiple vectors, and to only consider those vectors containing relevant information for the respective prediction, when a prediction is being made. Since this mechanism offers a way of determining which parts of the input sequence are important (and which parts are less), it is better at keeping up its performance when processing longer sequences (Bahdanau et al., 2014; Luong et al., 2015).

In their foundational paper *Attention is All You Need*, Vaswani et al. (2017), introduced the Transformer architecture for the first time. Compared to earlier implementations of Attention (e.g. Parikh et al., 2016), this Google-based research team designed an architecture without recurrent network units. The Transformer model entirely relies on an attention mechanism to 'compute representations of its input and output' (Vaswani et al., 2017: 2). Rather than calculating attention vectors for a single word, a Transformer enables the computation of multiple attention layers at once. The so-called 'Multi-head Attention' module (Vaswani et al., 2017) allows an encoder to execute the computation of attention in parallel, for all words in the input sequence (Figure 3), while the decoder still operates sequentially, generating the final output 'from left to right,' one word at a time. As the Transformer facilitates more parallelization of computational processes during training, it has enabled training on much larger datasets than previously possible (Wolf et al., 2020). The more computations are executed in parallel, the easier it is to spread the training over a cluster of processing units, which shortens training times when lots of hardware is available.

Based on this detour into the Transformer as an algorithmic technique, we can begin to understand why and how the Transformer revolutionized NLP in a context of local compute abundance. With regards to the training of language models, parallelizability is particularly of interest to those who already own the respective hardware to take full advantage of it, or to those who possess the capital to rent or acquire the computing power to make use of it. As such, parallelizability is the central feature that makes an architecture like the Transformer compatible with the overall logic of data centres and cloud computing. Arguably, the invention of this architecture by a company like Google 'does not follow teleologically from the mere existence of computing machinery and programming languages' (Rieder, 2017: 101). In this sense, the Transformer is 'the product not just of a technological logic, but simultaneously of a social logic, the logic of producing surplus-value' (Dyer-Witheford et al., 2019: 3).

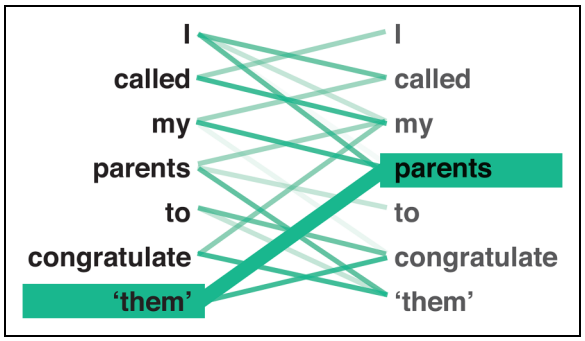

Figure 3. In the Transformer network architecture, attention vectors are calculated for all words in the input sentence. When predicting the word 'them,' the network pays specific attention to the object of the sentence ('my parents').

While parallelization is a general approach to designing and analyzing computational models, the introduction of the Transformer has pushed NLP research into a direction that plays into the hands of large corporations like Google, who are among the few actors who can operationalize this technique at scale. In the long run, this further shifts power relations in their favour by introducing consequential path dependencies into the field. The next sections further unpack how the introduction of this architecture has impacted and restructured the landscape of NLP research.

Reaching for the sky: Transformer-based language models

For those who can afford the required computing power, the Transformer architecture has made it feasible to 'pretrain' enormous language models using data sets of hundreds of gigabytes containing (primarily English language) web-based text resources like, for example, Wikipedia entries (e.g. Radford et al., 2019a). Pretrained models are useful to improve the analysis of other data sets for different applications: 'Instead of training models on a specific task from scratch, pretraining models are firstly trained on general-domain corpora, then followed by fine-tuning on downstream tasks' (Zhao et al., 2019: 1). Crucially, pretrained language models are reusable components that can be employed to initialize the model parameters for a wide range of tasks in NLP. First proposed by Google researchers Dai and Le (2015) to initialize LSTMs, pretrained models are a fairly recent innovation.

The first pretrained model based on and closely adhering to the Transformer architecture was 'gGenerative pretrained Transformer' (GPT) (Radford et al., 2018), a generative language model released by OpenAI in June 2018. OpenAI had been founded by Elon Musk, Sam Altman, and others in 2015 with the stated goal to harness the social benefits of Artificial General Intelligence, which is the highly speculative and contentious idea of an AI that 'has the full range

of intellectual capabilities that a person has […] at or above the same level as a typical person' (Wooldridge, 2020: 41).[4] GPT was trained on BookCorpus, an unlabelled collection of more than 7000 self-published books, and with its 110M parameters, the model was pioneering the scaling-up of language models. GPT's initial training was unsupervised and not targeted to a specific task, but it was then fine-tuned for specific applications such as classification or sentiment analysis (Shree, 2020). Soon after, Google open-sourced the architecture 'BERT' and released some pretrained models based on it (Devlin et al., 2019). In contrast to GPT, which largely follows the original Transformer work by Vaswani et al. (2017) (Radford et al., 2018), BERT does not include both an encoder and a decoder but uses multiple encoders stacked on top of each other. BERT has 340 million parameters and is used to improve Google search results (Nayak, 2019).

Since the success and influence of GPT and BERT, pretrained models have flourished and companies have entered what seems to be a race to build ever bigger models (cf. Table 1). In early 2019, OpenAI released GPT-2 (Radford et al., 2019a), a model with 1.5 billion parameters, 10 times as many as its predecessor. GPT-3 (175 billion parameters) followed in June 2020 (Brown et al., 2020). To our knowledge, the largest model to date was recently announced by the Chinese-government funded Beijing Academy of Artificial Intelligence and has 1.75 trillion parameters (Feng, 2021). So far, all of these models confirm the apparent correlation between size and sophistication based on the various task-specific performance metrics used by practitioners. The limits of scaling-up even further do not seem to have been reached yet.

Understanding language models as vehicles of power

LLMs and the push towards the cloud

As reusable components that can be adapted for different tasks, language models can function as infrastructures which facilitate the development of deep learning applications worldwide. Once a model is trained, it becomes a *means of production* enabling 'software-makers to step further faster, not merely regarding resource efficiency but in terms of what can be considered possible in the first place' (Rieder, 2020: 16). Like programming languages or software frameworks, pretrained language models 'widen the spaces of expressivity, broaden the scope of ambitions, but also structure, align, and standardize' (Rieder, 2020: 16). This understanding of large models like LLMs as means of production is also reflected in the following quote by Jeff Scott, CTO at Microsoft, albeit in a clearly hyperbolic way, which problematically implies that current AI systems and large models possess some kind of universality and generality akin to the far-fetched idea of AGI:

> By now most people intuitively understand how personal computers are a platform—you buy one and it's not like everything the computer is ever going to do is built into the device when you pull it out of the box […] That's exactly what we mean when we say AI is becoming a platform […] This is about taking a very broad set of data and training a model that learns to do a general set of things and making that model available for millions of developers to go figure out how to do interesting and creative things with it (Scott cited in Langston, 2020).

However, with the scaling-up of language models, some problems arise that hinder these benefits as adaptable and reusable components. The growing size of Transformer models has raised concerns about them becoming so big that it becomes complicated to move them off the infrastructure they were trained on and integrate them into concrete applications (Riedl, 2020). First, this is due to the size of LLMs. While GPT-2 could still be handled with moderate hardware, its 5GB already made it unfeasible for local deployment on a mobile device (Kaiser, 2020). GPT-3 is more than 10 times as big: With 570GB, it becomes a challenge for all but most computers to do anything with the model at all. LLMs are thus often deployed in the cloud instead of on a device. In such a software-as-a-service model, the language model stays on remote servers which users can address via the web. Second, training and running LLMs requires significant amounts of processing power (Strubell et al., 2019; Bender et al., 2021). To save time and energy, large models are often trained and deployed on GPUs, or even purpose-built AI chips such as ASICs or FPGAs (Hwang, 2018). Lastly, LLMs also require a lot of random-access memory (RAM) in order to run frictionless. Especially when handling long sentences, LLMs become very memory-intensive. In order to mitigate some of these problems, research has focussed on reducing the size of these models (e.g. Sanh et al., 2020). However, these 'compressed' versions still 'rely on large quantities of data and [require] significant processing and storage

Table 1. Selection of known, record-breaking language models based on the Transformer architecture.

Model	Year of release	Company	Number of parameters
GPT	2018	OpenAI	110 million
BERT	2018	Google	340 million
GPT-2	2019	OpenAI	1.5 billion
MegatronLM	2019	NVIDIA	8.3 billion
Turing-NLG	2020	Microsoft	17 billion
GPT-3	2020	OpenAI	175 billion
T6-XXL	2021	Google	1.6 trillion
WuDao 2.0	2021	Beijing Academy of Artificial Intelligence	1.75 trillion

capabilities to both hold and reduce the model' (Bender et al., 2021: 612).

Considering the unequal distribution of large-scale computing power in research and the wider economy (Ahmed and Wahed, 2020; Srnicek, 2019), the size, compute- and memory requirements of LLMs can become a major constraint, turning people towards cloud deployment. Corporate cloud solutions offered by actors such as Google or Amazon allow for flexible and scalable deployment, making them attractive to companies that do not have the financial means to operate their own computational infrastructure, especially when their need for compute is intermittent, as is the case in the training phase of LLMs. To profit from this situation, big tech companies have constructed extensive cloud ecosystems for AI. For example, Amazon Web Services today not only contains ready-to-use services like 'Rekognition' for image and video analysis, or 'Comprehend' for text analytics, but provides software development platforms which span the entire production process of machine learning systems, supporting developers with everything from data preparation and labelling to deployment and monitoring (Mucha and Seppala, 2020). In addition, Amazon provides exclusive specialized hardware for AI training and inference (e.g. the 'Inferentia' chips), which allows users to run models much quicker than on traditional CPU or GPU chips. Instances of this hardware can be rented in an infrastructure-as-a-service model, optionally with pre-installed machine learning frameworks like TensorFlow or PyTorch or other software solutions on Amazon's software palette. The example of LLMs in the specific AI subfield of NLP thus concretizes and extends Srnicek's (2019, 2020) argument that computes has become central to AI research and development: If the field continues to progress by scaling-up language models, smaller companies and other actors aiming to develop NLP systems are likely to turn to the big tech companies who, through their cloud businesses, provide the integrated compute infrastructures needed to work with large models. This solidifies these companies' role as rentiers in the political economy of AI.

Nonetheless, even the deployment of LLMs in the cloud can require significant knowledge and skills in order to handle the models correctly and avoid large bills (Kaiser, 2020). Especially smaller companies likely lack the respective professionals. This is where OpenAI's language model GPT-3 comes in. In the following, we present the case of GPT-3, a commodified language model that can be used off-the-shelf. We argue that GPT-3 and its novel paid API access model, risks further accelerating dependency on big AI companies throughout the wider economy.

GPT-3: Redefining AI-as-a-service

With 175 billion parameters, GPT-3 is one of the largest language models today, provisioned through a novel access model. Instead of releasing the model as open-source like its predecessors, OpenAI introduced an API through which accepted users can access it as a running system to generate textual output such as translations or summaries (Floridi and Chiriatti, 2020). After a two-month 'beta' phase, where users could test the service for free, OpenAI introduced a pricing plan in October 2020:

> Explore: Free tier: 100K tokens or a three-month trial, whichever you use up first
> Create: $100 per month for 2M tokens, plus 8 cents for every additional 1K tokens.
> Build: $400 per month for 10M tokens, plus 6 cents for every additional 1K tokens.
> Scale: Contact OpenAI for pricing (Macaulay, 2020).

The release of GPT-3 as a commercial model that is partially accessible can be seen in line with a series of earlier decisions by OpenAI that have pointed to the company's move away from open-source. In early 2019, after releasing GPT-2 (reluctantly) as open-source (Radford et al., 2019b), the company announced that it was restructuring its nonprofit business model towards a 'capped-profit' model to increase its investments in compute power and AI talent (Brockman et al., 2019). The way the company is currently provisioning GPT-3 is emblematic of this strategy.

The closed-off commercial API provision model hinders many of the benefits of language models as means of production discussed earlier. Arguably, GPT-3 does not function as a platform in the sense of being fully programmable (Plantin et al., 2018). Users can interact with the model but do so without reprogramming it or being able to use it for initializing their own language models. Instead, users generate textual outputs by feeding the model a so-called prompt, for example, in the form of a small task description or the beginning of a sentence (Brown et al., 2020). GPT-3 performs particularly well when it is given a few examples of how to perform a task in the prompt ('few-shot learning'). These examples allow users to more precisely specify the nature of their task and, by doing so, optimize GPT-3's results. However, depending on the desired output, writing a good prompt can be difficult and involves some experimentation. Arguably, it is a skill that could be compared to the activity of programming. In this sense, GPT-3 is 'programmable,' yet only within tight boundaries specified by OpenAI, if we can speak about programmability at all. In any case, we can observe a 'reification of abstraction into [a] tight, commodified infrastructure' (Rieder, 2020: 25), whereby complex underlying operations are being packaged into single commands. GPT-3 hides complex language

modelling tasks behind abstraction layers, obscuring the technical functioning and preventing modification.

By operating GPT-3 as a closed system and controlling its accessibility, OpenAI is said to have created a model of 'unique dependence' (Mayer, 2021), as it no longer allows developers to view, assess or build-on-top of GPT-3. Instead, users are required to rely on the ready-made commercial model, which offers them the opportunity to tap into GPT-3 without access to the model's inner workings. This opacity not only hinders NLP-related research into LLMs but also has consequences for commercial startups for whom it becomes even more difficult to replicate such a language model for their own purposes. As becomes clear, OpenAI attempts to establish itself as a powerful rentier (Srnicek, 2019) of an essential infrastructural component in the context of NLP in the form of GPT-3, aiming to commercially benefit from the unequal distribution of large-scale computing resources (Ahmed and Wahed, 2020; Srnicek, 2019). While it remains to be seen whether OpenAI's business model works out, LLMs are generally advertised as valuable assets within the industry, potentially 'providing several billions of dollars' worth of value from a single model' (Catanzaro, as cited in Hemsoth, 2021).

Besides distributing GPT-3 through a commercial API with limited functionality, OpenAI exclusively licenced the model to Microsoft, providing the company access to underlying code and the opportunity to modify, repurpose, and embed it into its larger cloud computing infrastructure (Hao, 2020b) and consumer products (Langston, 2021). This exclusive deal followed a multiyear corporate partnership between the two companies after Microsoft's initial investment of one billion dollars in OpenAI (Etherington, 2019)—partly consisting of compute credits for its Azure cloud infrastructure. Ever since, the company has been supporting OpenAI by collaboratively building a supercomputer for their purposes and hosting GPT-3's API on Azure (Microsoft, 2019). According to Microsoft, the supercomputer used for GPT-3 consists of 'a single system with more than 285,000 CPU cores, 10,000 GPUs, and 400 gigabits per second of network connectivity for each GPU server' (Langston, 2020). Access to such efficient hardware helps OpenAI keep the training costs low. As becomes clear, the partnership between OpenAI and Microsoft is a mutually beneficial alliance, which allows both companies to further strengthen their already powerful positions.

Conclusion

With the introduction of the Transformer architecture in 2017, Google started the era of LLMs in NLP. Congruent with the increasing computational requirements in AI since the rise of deep learning, the Transformer architecture allows for more parallelization in the training of language models. As a consequence, big tech corporations with large data centres started to experiment with very large models. Due to the persistent correlation between size and certain performance indicators of these LLMs, a race for ever bigger model has started, at least in the commercial sector. Examining the role of LLMs in the political economy of AI through the example of GPT-3, this paper has argued that current trends in NLP risk accelerating processes of monopolization and dependence on hyper-scaling AI companies. While LLMs are, in theory, reusable components that can be implemented by developers in a variety of applications, their size as well as compute- and memory requirements make them difficult to handle. In line with Srnicek (2019) and Dyer-Witheford et al. (2019), we have argued that companies or researchers looking to adopt LLMs are likely attracted to one of the big tech corporate cloud infrastructures in order to train a model themselves on their scalable infrastructure, or to use a ready-made model like GPT-3 through a commercial API.

The observations made in this article thus point to problems which are not new but have not received much attention in the particular context of AI. In our view, the acceleration of monopolization stresses the need to consider legal limitations to the growing power of corporations in the digital era (e.g. Khan, 2018). As Meredith Whittaker (2021) has argued, '[t]he technological breakthroughs that propelled the current AI gold rush from deep face to AlphaGo to GPT-3 are all contingent on the vast power and resources of the current big tech business ecosystem.' These companies possess the capital, infrastructure, and data to turn algorithmic techniques for AI, such as the Transformer, into 'engines of profit', thereby solidifying their dominant position. Retracing the history of the Transformer as an algorithmic technique adds to our understanding of how technical properties relate to economic outcomes. Our analysis thus sheds light on another dimension of the complex big tech ecosystem and underlines the difficulty of keeping platform power at bay.

Indeed, critical research into LLMs shows that regulatory oversight over the creation and deployment of such systems may be necessary. The paper that led Google to dismiss Timnit Gebru (Bender et al., 2021) indeed discusses LLMs critically, highlighting their significant requirements in terms of money and energy as well as their potential to reproduce hegemonic worldviews or denigrate vulnerable communities. And other scholars have raised concerns over the output of LLMs, and their potential to further solidify filter bubbles (Floridi and Chiriatti, 2020), discriminate against marginalized groups (Abid et al., 2021), and generate fake news or violence-inciting social media posts (McGuffie and Newhouse, 2020). Considering these risks and concerns, the creation of LLMs asks for critical oversight by public institutions and civil society, particularly as systems like GPT-3 are increasingly operated outside

of research environments and implemented into applications that are being used in different societal settings.

The costs and difficulties of creating and deploying LLMs make it unlikely that the crucial role of big tech corporations in the provision of these services will diminish in the foreseeable future. Due to economies of scale, LLMs lend themselves to centralized cloud distribution, with the effect that a few large models may become widely used building blocks. This can be observed through the example of GPT-3, which is already being used in several Microsoft products and other third-party apps (Langston, 2021; Pilipiszyn, 2021). However, there are some open-source initiatives aiming to make LLMs more accessible (Khan, 2020). For example, EleutherAI 'is a grassroots collective of researchers working to open-source AI research' (EleutherAI, n.d.) founded in July 2020 whose main goal is to replicate GPT-3 and publish it under an open-source licence. If the initiative succeeds, the challenges of handling and running such a large model would remain. Nevertheless, by open-sourcing the model and its underlying code, EleutherAI would significantly lower access barriers. Researchers could, for example, interact with a GPT-3-like model on the level of code rather than just through an API and study its architecture, which is not possible with GPT-3 as OpenAI did not release some key details (Leahy as cited in Wiggers, 2021).

Our last point concerns independent AI research. Countering the idea of teleological progress in technology, this article has proposed the Transformer as a case study to highlight the ways big tech corporations push the development of AI in their favour. The developments in NLP, which are the focus of this paper, reflect the increasing power of these companies over AI development and innovation in general. While the controversy around the firings of Gebru and Mitchell attest to the lack of independence of corporate in-house research, big tech companies also increasingly infiltrate AI research at independent institutions: For example, they broadly engage in scientific collaboration, while retaining sole control over the resulting patents (Rikap and Lundvall, 2020). They aggressively fund individual scholars and entire study programmes, as well as major conferences and events, thereby 'working to shape the terms of the field, and to act as the gatekeeper to the infrastructure, data and funding needed to do so-called cutting-edge research' (Whittaker, 2021). As Abdalla and Abdalla (2021) have argued, they attempt to significantly distort academic discussion and public knowledge to serve their interests. In contrast to research at big tech companies, independent AI research thus finds itself in an increasingly precarious position. This limits researchers' ability to develop alternative techniques and NLP systems that require fewer financial and energy resources, and are less hungry for data, more transparent, and so forth (Bender et al., 2021). Therefore, the last proposition we want to make in terms of governmental measures is that more financial support could be directed to independent AI research. Critical scholarship coming from an independent and diverse group of academics and practitioners is crucial to advancing the development of techniques and systems that provide alternative ways of 'doing AI' and therefore contribute to different understandings about AI's possibilities and applications, out of reach from the confining embrace of big tech companies.

Acknowledgements

First, we would like to thank the editors and three reviewers for their meticulous feedback on this article which helped us to significantly revise and improve the text. We are equally grateful to Bernhard Rieder who encouraged us to submit this paper for publication and provided valuable feedback on several earlier versions of this article. We also thank Maxigas and Sarah Burkhardt for their insightful comments. Lastly, we would like to thank Thomas Poell for his support in publishing this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received financial support for the publication of this article from the *Graduate School of Humanities UvA*.

ORCID iDs

Dieuwertje Luitse https://orcid.org/0000-0003-0652-3315
Wiebke Denkena https://orcid.org/0000-0003-2728-7024

Notes

1. Using the term 'AI', we acknowledge the ambiguity of this field of research and the difficulties to draw clear lines of demarcation. In our understanding, 'AI' encompasses various approaches rooted in different research traditions and disciplines. Today, many of these share the application of neural network technology and the methodology of deep learning. When we speak about existing AI systems in this paper, we refer to 'narrow AI' deep learning systems.
2. Given a language model that approximates the English language, the probability that is assigned to a sequence of words is then supposed to represent the likelihood that this exact sequence occurs in the English language. In order to capture as much of a given language in a language model as possible, the model is trained on large amounts of data that ideally contains as much of its nuances, dialects, and manners of speaking as possible. In practice, many LLMs (e.g. GPT-3) are not solely trained on text from one language but use a large corpus of resources for training, making them capable of text generation in several languages and translation tasks. Yet, such models are usually heavily English-centric.
3. While RNN's (and Transformers) can also be used for other tasks than word prediction, we will stick to this example throughout the paper for reasons of simplicity.

4. Research into AGI is located on the margins of contemporary AI research which focuses on task-specific applications ('narrow AI').

References

Abdalla M and Abdalla M (2021) The Grey Hoodie project: big tobacco, big tech, and the threat on academic integrity. *arXiv:2009.13676 [cs]*. Available at: https://arxiv.org/abs/2009.13676.

Abid A, Farooqi M and Zou J (2021) Persistent anti-muslim bias in large language models. *arXiv:2101.05783 [cs]*. Available at: http://arxiv.org/abs/2101.05783.

Ahmed N and Wahed M (2020) The de-democratization of ai: deep learning and the compute divide in artificial intelligence research. *arXiv: 2010.15581*. Available at: http://arxiv.org/abs/2010.15581.

Amoore L (2020) *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham: Duke University Press.

Bahdanau D, Cho K and Bengio Y (2014) Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473 [cs, stat]*. Available at: http://arxiv.org/abs/1409.0473.

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the dangers of stochastic parrots: can language models be too big? In: *Conference on Fairness, Accountability, and Transparency (FAccT'21)*. ACM, New York, USA, 610–623. DOI: 10.1145/3442188.3445922.

Benjamin R (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge, UK: Polity.

Bengio Y, Simard P and Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 5(2): 157–166.

Brockman G and Sutskever I and OpenAI (2019) OpenAI LP. Available at: https://openai.com/blog/openai-lp/ (accessed 12 February 2021).

Brown TB, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. *arXiv:2005.14165 [cs]*. Available at: http://arxiv.org/abs/2005.14165.

Cardon D, Cointet J-P and Mazières A (2018) La Revanche des Neurones: L'invention des Machines Inductives et la Controverse de l'Intelligence Artificielle. *Réseaux* 211(5): 173.

Dai AM and Le QV (2015) Semi-supervised sequence learning. *arXiv:1511.01432 [cs]*. Available at: http://arxiv.org/abs/1511.01432.

Devlin J, Chang M-W, Lee K, et al. (2019) BERT: pre-training of deep bidirectional Transformers for language understanding. *arXiv:1810.04805 [cs]*. Available at: http://arxiv.org/abs/1810.04805.

Dyer-Witheford N, Kjøsen AM and Steinhoff J (2019) *Inhuman Power: Artificial Intelligence and the Future of Capitalism*. London: Pluto Press.

EleutherAI (n.d.) EleutherAI. Available at: http://web.archive.org/web/20210601074447/https://www.eleuther.ai/ (accessed 1 June 2021).

Etherington D (2019) Microsoft invests $1 billion in OpenAI in new multiyear partnership. *TechCrunch*. Available at: https://social.techcrunch.com/2019/07/22/microsoft-invests-1-billion-in-openai-in-new-multiyear-partnership/ (accessed 20 January 2021).

Feng C (2021) US-China tech war: Beijing-funded AI researchers surpass Google and OpenAI with new language processing model. *South China Morning Post.* Available at: https://www.scmp.com/tech/tech-war/article/3135764/us-china-tech-war-beijing-funded-ai-researchers-surpass-google-and.

Floridi L and Chiriatti M (2020) GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30: 681–694.

Fuller M (ed) (2008) *Software Studies: A Lexicon*. Cambridge, MA: MIT Press.

Graves A (2013) Generating sequences with recurrent neural networks. *arXiv:1308.*0850. Available at: https://arxiv.org/abs/1308.0850.

Hao K (2020a) "I started crying": inside Timnit Gebru's last days at Google. Available at: https://www.technologyreview.com/2020/12/16/1014634/google-ai-ethics-lead-timnit-gebru-tells-story/ (accessed 16 December 2020).

Hao K (2020b) OpenAI is giving Microsoft Exclusive Access to its GPT-3 Language Model. Available at: https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/ (accessed 21 January 2021).

Hemsoth N (2021) The billion dollar AI problem that just keeps Scaling. *The Next Platform*. Available at: https://www.nextplatform.com/2021/02/11/the-billion-dollar-ai-problem-that-just-keeps-scaling/ (accessed 17 March 2021).

Hernández A and Amigó JM (2020) Differentiable programming and its applications to dynamical systems. *arXiv:1912.08168 [cs, math]*. Available at: http://arxiv.org/abs/1912.08168.

Hochreiter S and Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8): 1735–1780.

Hwang T (2018) Computational power and the social impact of artificial intelligence. *SSRN Electronic Journal* (April 2018): 1–47. Available at: https://ssrn.com/abstract=3147971

Kaplan J, McCandlish S, Henighan T, et al. (2020) Scaling laws for neural language models. *arXiv:2001.08361 [cs, stat]*. Available at: http://arxiv.org/abs/2001.08361.

Kaiser C (2020) Too big to deploy: how GPT-2 is breaking servers. *Towards Data Science*. Available at: https://towardsdatascience.com/too-big-to-deploy-how-gpt-2-is-breaking-production-63ab29f0897c (accessed 12 February 2021).

Khan B (2020) *Algpt2 Part 2. How I (Almost) Replicated OpenAI's GPT-2 (124M version)*. Available at: http://web.archive.org/web/20210131010709/https://bkkaggle.github.io/blog/algpt2/2020/07/17/ALGPT2-part-2.html (accessed January 27, 2021).

Khan L (2018) The new brandeis movement: America's antimonopoly debate. *Journal of European Competition Law & Practice* 9(3): 131–132.

Krizhevsky A, Sutskever I and Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: *25th International Conference on Neural Information Processing Systems—Volume 1 (NIPS'12)*. Curran Associates Inc., Red Hook, NY, USA, 1097–1105.

Langston J (2020) Microsoft announces new supercomputer, lays out vision for future AI work. *The AI Blog*. Available at: https://blogs.microsoft.com/ai/openai-azure-supercomputer/ (accessed 18 January 2021).

Langston J (2021) From conversation to code: microsoft introduces its first product features powered by GPT-3. *The AI Blog*. Available at: https://blogs.microsoft.com/ai/from-

conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/.

LeCun Y, Bengio Y and Hinton G (2015) Deep learning. *Nature* 521: 436–444.

Lee K-F (2018) *AI Superpowers: China, Silicon Valley, and the New World Order*. Boston, MA: Houghton Mifflin Harcourt.

Luong T, Pham H and Manning CD (2015) Effective approaches to attention-based neural machine translation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015*, 1412–1421.

Macaulay T (2020) OpenAI reveals the pricing plans for its api—and it ain't cheap. Available at: https://thenextweb.com/neural/2020/09/03/openai-reveals-the-pricing-plans-for-its-api-and-it-aint-cheap/ (accessed 12 February 2021).

Marino MC (2020) *Critical Code Studies. Software Studies*. Cambridge, MA: The MIT Press.

Mayer HM (2021) Revolutionary NLP model GPT-3 poised to redefine AI and next generation of startups. *Forbes*. Available at: https://www.forbes.com/sites/hannahmayer/2021/01/02/revolutionary-nlp-model-gpt-3-poised-to-redefine-ai-and-next-generation-of-startups/ (accessed 19 January 2021).

McCulloch WS and Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics* 5(4): 115–133.

McGuffie K and Newhouse A (2020) The radicalization risks of GPT-3 and advanced neural language models. *arXiv:2009.06807 [cs]*. Available at: http://arxiv.org/abs/2009.06807.

Microsoft (2019) OpenAI forms exclusive computing partnership with microsoft to build new azure AI supercomputing technologies. Available at: https://news.microsoft.com/2019/07/22/openai-forms-exclusive-computing-partnership-with-microsoft-to-build-new-azure-ai-supercomputing-technologies/ (accessed 20 January 2021).

Minsky M and Papert S (1969) *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

Mucha T and Seppala T (2020) *Artificial Intelligence Platforms—A New Research Agenda for Digital Platform Economy*. ETLA Working Paper (76). Available at: https://ssrn.com/abstract=3532937.

Nayak P (2019) Understanding searches better than ever before. *The Keyword* Available at: https://blog.google/products/search/search-language-understanding-bert/ (accessed 4 July 2021).

Nieborg DB and Helmond A (2019) The political economy of Facebook's platformization in the mobile ecosystem: Facebook messenger as a platform instance. *Media, Culture & Society* 41(2): 196–218.

Nieborg DB and Poell T (2018) The platformization of cultural production: Theorizing the contingent cultural commodity. *New Media & Society* 20(11): 4275–4292.

Parikh AP, Täckström O, Das D, et al. (2016) A decomposable attention model for natural language inference. *arXiv:1606.01933 [cs]*. Available at: http://arxiv.org/abs/1606.01933.

Pascanu R, Mikolov T and Bengio Y (2013) On the Difficulty of training recurrent neural networks. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML'13)— Volume 28, Atlanta, GA, USA*, 1310–1318.

Pilipiszyn A (2021) GPT-3 powers the next generation of apps. Available at: https://openai.com/blog/gpt-3-apps/ (accessed 30 June 2021).

Plantin J-C, Lagoze C, Edwards PN, et al. (2018) Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society* 20(1): SAGE Publications: 293–310.

Radford A, Narasimhan K, Salimans T, et al. (2018) Improving language understanding by generative pre-training. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford A, Wu J, Child R, et al. (2019a) Language models are unsupervised multitask learners. Available at: https://cdn.openai.-com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Radford A, Wu J, Amodei D, et al. (2019b) Better language models and their implications. Available at: https://openai.-com/blog/better-language-models/ (accessed 12 February 2021).

Raina R, Madhavan A and Ng AY (2009) Large-scale deep unsupervised learning using graphics processors. In: *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09), Montreal, Quebec, Canada*, 1–8.

Rieder B (2017) Scrutinizing an algorithmic technique: The bayes classifier as interested reading of reality. *Information, Communication & Society* 20(1): 100–117.

Rieder B (2020) *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam: Amsterdam University Press.

Rieder B and Sire G (2014) Conflicts of interest and incentives to bias: A microeconomic critique of Google's tangled position on the web. *New Media & Society* 16(2): 195–211.

Riedl M (2020) AI democratization in the era of GPT-3. *The Gradient*, 25 September. Available at: https://thegradient.pub/ai-democratization-in-the-era-of-gpt-3/ (accessed 12 February 2021).

Rikap C and Lundvall B-Å (2020) Big tech, knowledge predation and the implications for development. *Innovation and Development*: 1–28.

Rumelhart D, Hinton GE and Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088): 533–536.

Sanh V, Debut L, Chaumond J, et al. (2020) DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*. Available at: http://arxiv.org/abs/1910.01108.

Shree P (2020) The journey of open AI GPT models. Available at: https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2 (accessed 15 February 2021).

Srnicek N (2017) *Platform Capitalism*. Cambridge/Malden: Polity Press.

Srnicek N (2018) Platform monopolies and the political economy of AI. In: *Economics for the Many*. London/New York: Verso, 153–163.

Srnicek N (2019) The Political Economy of Artificial Intelligence. Great Transformation: Die Zukunft moderner Gesellschaften, Friedrich-Schiller-Universität, Jena, Germany. Available at:

https://www.youtube.com/watch?v=Fmi3fq3Q3Bo (accessed 9 November 2020).

Srnicek N (2020) Data, compute, labour. Available at: https://www.adalovelaceinstitute.org/blog/data-compute-labour/ (accessed 16 November 2020).

Staab P (2019) *Digitaler Kapitalismus: Markt und Herrschaft in der Ökonomie der Unknappheit*. Berlin: Suhrkamp.

Strubell E, Ganesh A and McCallum A (2019) Energy and policy considerations for deep learning in NLP. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*, 3645–3650.

Sutskever I, Vinyals O and Le QV (2014) Sequence to sequence learning with neural networks. *arXiv:1409.3215 [cs]*. Available at: http://arxiv.org/abs/1409.3215.

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. *arXiv:1706.03762 [cs]*. Available at: http://arxiv.org/abs/1706.03762.

Whittaker M (2021) From ethics to organizing: getting serious about AI. Distinguished Speaker Series, Hariri Institute for Computing, Boston University, Boston, MA, United States. Available at: https://www.youtube.com/watch?v=_BzU0bD0Ics (accessed 11 May 2021).

Wiggers K (2021) AI Weekly: meet the people trying to replicate and open-source OpenAI's GPT-3. *VentureBeat*. Available at: https://venturebeat.com/2021/01/15/ai-weekly-meet-the-people-trying-to-replicate-and-open-source-openais-gpt-3/.

Wittel A (2017) The political economy of digital technologies: Outlining an emerging field of research. In: Koch G (eds) *Digitisation: Theories and Concepts for Empirical Cultural Research*. New York: Routledge, 251–275.

Wolf T, Debut L, Sanh V, et al. (2020) HuggingFace's transformers: State-of-the-art natural language processing. *arXiv:1910.03771 [cs]*. Available at: http://arxiv.org/abs/1910.03771.

Wooldridge M (2020) *The Road to Conscious Machines: The Story of AI*. London: Penguin.

Zhao Z, Chen H, Zhang J, et al. (2019) UER: An open-source toolkit for pre-training models. In: *Proceedings of the 2019 EMNLP and the 9th IJCNLP (System Demonstrations)*, Hong Kong, Association for Computational Linguistics, 241–246.

Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.