Susanne Förster

THE BIGGER THE BETTER?! THE SIZE OF LANGUAGE MODELS AND THE DISPUTE OVER ALTERNATIVE ARCHITECTURES

Abstract

This article looks at a controversy over the 'better' architecture for conversational AI that unfolds initially along the question of the 'right' size of models. Current generative models such as ChatGPT and DALL-E follow the imperative of the largest possible, ever more highly scalable, training dataset. I therefore first describe the technical structure of large language models and then address the problems of these models which are known for reproducing societal biases or so-called hallucinations. As an 'alternative', computer scientists and AI experts call for the development of much smaller language models linked to external databases, that should minimize the issues mentioned above. As this paper will show, the presentation of this structure as 'alternative' adheres to a simplistic juxtaposition of different architectures that follows the imperative of a computable reality, thereby causing problems analogous to the ones it tried to circumvent.

APRJA Volume 12, Issue 1, 2023 ISSN 2245-7755. CC license: 'Attribution-NonCommercial-ShareAlike'.

In recent years, increasingly large, complex and capable machine learning models such as the GPT model family, DALL-E or Stable Diffusion have become the super trend of current (artificially intelligent) technologies. Trained on identifying patterns and statistical features and thus intrinsically scalable, the potential of large language models is seen as based on their generative capabilities to produce a wide range of different texts and images.

The monopolization and concentration of power within a few big tech companies such as Google, Microsoft, Meta and OpenAI that accompanies this trend is promoted by the enormous economic resources afforded by the models' training processes (see Luitse and Denkena). The risks and dangers of this big data paradigm have been stressed widely: The working conditions and invisible labor that goes into the creation of AI and ensures its fragile efficacy has been addressed in the context of click-work or content moderation (f.e., Irani; Rieder and Skop). In *Anatomy of an AI System,* Kate Crawford's and Vladan Joler (Crawford and Joler) detailed the material setup of a conversational device and traced the far fetching origins of its hardware components and working conditions. Critical researchers have also pointed out how the composition of training data has resulted in the reproduction of societal biases. Crawled from the Internet, the data and thus the generated language mainly represent hegemonic identities whilst discriminating against marginalized ones (Benjamin). Moreover, the infrastructure needed to train these models requires huge amounts of computing power and has been linked to a heavy environmental footprint: The training of a big Transformer model emitted more than 50 times the amount of carbon dioxide than an average human per year (Strubell et al., Bender et al.). Criticizing this seemingly inevitable turn to ever larger language models and the far-reaching implications of this approach for both people and the environment, Emily Bender et al., published their now-famous paper *On the Dangers of Stochastic Parrots: Can Language Models be Too Big?* in March 2021 (Bender et al.). Two of the authors, Timnit Gebru and Margaret Mitchell, both co-leaders of Google's Ethical AI Research Team, were fired after publishing this paper against Google's veto.

The dominance of the narrative of "scalability, [...], the ability to expand - and expand, and expand" (Tsing 5) deep learning models – especially by big tech companies – has clouded the view for alternative approaches. With this paper, I will look at claims and arguments for different architectures of conversational AI by first reconstructing the technical development of generative language models. I will further trace the reactions to errors and problems of generative large language models and the dispute over the 'proper' form of artificial intelligence between proponents of connectionist AI and machine learning approaches on the one side and those of symbolic or neurosymbolic AI defending the need for 'smaller' language models linked to external knowledge databases on the other side. This debate represents a remarkable negotiation about forms of 'knowledge representation' and the question of how language models should (be programmed to) 'speak'.

Initially, the linking of smaller language models with external databases promising accessibility, transparency and changeability had subversive potential for me because it pledged the possibility of programming conversational AI without access to the large technical infrastructure it would take to train large language models (regardless of whether those models should be built at all). As I will show in the following, the hybrid models presented as an alternative to large language models also harbor dangers and problems, which are particularly evident in an upscaling of the databases.

In need of more data

Since its release in November 2022, the dialogue-based model ChatGPT generated a hype of unprecedented dimensions. Provided with a question, exemplary text or code snippet, ChatGPT mimics a wide range of styles from different authors and text categories such as poetry and prose, student essays and exams or code corrections and debug logs. Soon after its release, the end of both traditional knowledge and creative work as well as classical forms of scholarly and academic testing seemed close and were heavily debated. Endowed with emergent capabilities, the functional openness of these models is perceived as both a potential and a problem as they can produce speech in ways that appears human but contradicts human expectations and sociocultural norms. ChatGPT was also called a bullshit generator (McQuillan): Bullshitters, as philosopher Harry Frankfurt argues, are not interested in whether something is true or false, nor are they liars who would intentionally tell something false, but are solely interested in the impact of their words (Frankfurt).

Generative large language models such as OpenAI's GPT model family or Google's BERT and LaMDA are based on a neural network architecture – a cognitivist paradigm based on the idea of imitating the human brain logically-mathematically and technically as a synonym for "intelligence", but usually without taking into account physical, emotional and social experiences (see Fazi). In the connectionist AI approach, 'learning' processes are modeled with artificial neural networks consisting of different layers and nodes. They are trained to recognize similarities and representations within a big data training set and compute probabilities of co-occurrences of individual expressions such as images, individual words, or parts of sentences. After symbolic AI was long considered as the dominant paradigm, the "golden decade" of deep neural networks – also called deep learning – dawned in the 2010s, according to Jeffrey Dean (Dean). 2012 is recognized as the year in which deep learning gained acceptance in various fields: On the one hand, the revolution of speech recognition is associated with Geoff Hinton et al., on the other hand, the winning of the ImageNet Large Scale Visual Recognition Challenge with the help of a convolutional neural network represented a further breakthrough (Krizhevsky et al.). Deep learning neural networks with increasingly more interconnected nodes (neurons) and layers and powered by newly developed hardware components enabled huge amounts of compute power became the standard.

Another breakthrough is associated with the development of the Transformer Network architecture, introduced by Google in 2017. The currently predominant architecture for large language models is associated with better performance due to a larger size of the training data (Devlin et al.). Transformers are characterized in particular by the fact that computational processes can be executed in parallel (Vaswani et al.), a feature that has significantly reduced the models' training time. Building on the Transformer architecture, OpenAI introduced the Generative Pre-trained Transformer model (GPT) in 2018, a deep learning method which again increased the size of the training datasets (Radford et al., "Improving Language Understanding"). Furthermore, OpenAI included a process of pre-training, linked to a generalization of the model and an openness towards various application scenarios, what is thought to be achieved through a further step of optimization, i.e., the fine-tuning. At least with the spread of the GPT model family, the imperative of unlimited scalability of language models has become dominant. This was especially brought forward by Physics (Associate) Professor and Entrepreneur Jared Kaplan and OpenAI, who identified a set of 'scaling laws' for neural network language models, stating that the more data available for training, the better the performance thereof (Kaplan et al.). OpenAI has continued to increase the size of its models: While GPT-2 with 1.5 billion parameters (a type of variable learned in the process of training) was 10 times the size of GPT-1 (117 million parameters), it was far surpassed by GPT-3 with a scope of 175 trillion parameters. Meanwhile, OpenAI has transformed from a startup promoting the democratization of artificial intelligence (Sudmann) to a 30 billion dollar company (Martin) and from an open source community to a closed one. While OpenAI published research papers with the release of previous models describing the structure of the models, the size and composition of the training data sets, and the performance of the models in various benchmark tests, much of this information is missing from the paper on GPT-4.

On errors and hallucinations

Generative language models, however, are being linked – above all by developers and computer scientists – to a specific kind of 'error': "[I]t is also apparent that deep learning based generation is prone to hallucinate unintended texts", Ji et al. write in a review article collecting research on hallucination in natural language generation (Ji et al.). According to the authors, the term hallucination has been used in the field of computer visualization since about 2000, referring to the intentionally created process of sharpening blurred photographic images, and only recently changed to a description of an incongruence between image and image description. Since 2020, the term has also been applied to language generation, however not for describing a positive moment of artificial creativity (ibid.): Issued texts that appear sound and convincing in a real-world context, but whose actual content cannot be verified, are referred to by developers as 'hallucinations' (ibid., 4). In this context, hallucination refers not only to *factual* statements such as dates and historical events or the correct citation of sources; it is equally used for editions of non-existent sources or the addition of aspects in a text summary. While the *content* is up for discussion, the language *form* may be semantically correct and convincing, resulting in an apparent *trust* in the model or its language output.

For LeCun, Bengio and Hinton, "[r]epresentation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level." (LeCun et al. 436). In *technical terms*, hallucination thus refers to a *translation* or *representation error* between the source text or 'raw data' [sic] on the one hand and the generated text, model prediction or 'representation' on the other. Furthermore, another source of hallucinations is located in outdated data, causing the (over time) increasing production of factually incorrect statements. This 'error' is explicitly linked to the large scale of generative models: Since the training processes of these models are complex and expensive and thus seldomly repeated, the knowledge incorporated – generally – remains static (Ji et al.) However, with each successive release of the GPT model family, OpenAI proclaims further minimization of hallucinations and attempts to prevent programs from using certain terms or making statements that may be discriminatory or dangerous, depending on the context, through various procedures that are not publicly discussed (see Cao).

From the definitions of representation learning, hallucination, and the handling of this 'error', a number of conclusions can be drawn that are instrumental to the discourse on deep learning and artificial intelligence: The representation learning method assumes that it does not require any human intervention to recognize patterns in the available data, to form categories and make statements that are supposed to be consistent with the information located *in* the data. Both the data and the specific outputs of the models are conceived as universally valid. In this context, hallucination remains a primarily *technical* problem presented as *technically* solvable, and in this way it is closely linked to a promise of scaling: With the reduction of (this) error, text production seems to become autonomous, universal, and openly applicable in different settings.

On data politics

The assumption that data represent a 'raw' and objective found reality, which can be condensed and generated into a meaningful narrative through various computational steps, has been criticized widely (e.g. Boellstorff; Gitelman and Jackson). It is not only the composition of the data itself that is problematic, but equally the categories and patterns of meaning generated by algorithmic computational processes, which reinforce the bias – inevitably (see Jaton) – found in the data and make it once more effective (Benjamin; Noble). Technical computations adhere to an objectivity and autonomy that pushes human processes of selection and interpretation of the data into the background, presenting them instead as ‘found’ and ‘closed’ (e.g., boyd and Crawford; Kitchin). Building on a rich tradition of science and technology studies that highlighted the socio-technical co-production of human, natural and technical objects (f.e. Knorr Cetina, Latour and Woolgar), Adrian Mackenzie has introduced the term 'machine learner' to refer to the entanglement of "humans and machines or human-machine relations [...] situated amidst these three accumulations of settings, data and devices" (Mackenzie 23).

"[Big] data," as Taş writes, "are a site of political struggle." (Taş 569). This becomes clear not only through the public discussion of generative models and the underlying question of which statements language models are allowed to make. At the latest with the release of ChatGPT in November 2022, it was publicly debated which responses of the model were considered unexpected, incorrect or contrary to socio-cultural norms. Generative models have been tested in a variety of ways (Marres and Stark): The term 'jailbreaking' for example, denotes a practice in which users attempt to *trick* the model to create outputs that are restrained by the operating company's policy regulation. These include expressions considered as discriminating and obscene or topics such as medicine, health or psychology. In an attempt to circumvent these security measures, jailbreaking exposes the programmed limitations of the programs. Moreover, it also reveals what is understood by the corporations as the ‘sayable’ and the ‘non-sayable’ (see Foucault). This is significant insofar as these programs have already become part of everyday use, and the norms, logics, and limits inherent in them have become widely effective. In only five days after its release, ChatGPT had already reached one million users (Brockman). As foundation models (Bommasani et al.), OpenAI's GPT models and DALL-E are built into numerous applications, as are Google's BERT and LaMDA. Recently, the use of ChatGPT by a US lawyer or the demand to use the program in public administration (Armstrong; dpa/lno) was publicly discussed. These practices and usage scenarios make it clear that – practically – generative models represent technical infrastructures that are privately operated and give the operating big tech companies great political power. The associated authority in defining the language of these models but also in guiding politics recently became visible in a number of instances:

In an open letter, published in March 2023 on the website of the Future of Life Institute, AI researchers including Gary Marcus, Yoshua Bengio, and Yann LeCun – the latter working for Meta – as well as billionaire Elon Musk, urged for a six-month halt of training of models larger than GPT-4 (Future of Life Institute, “Pause Giant AI Experiments”). “Powerful AI systems”, they wrote, “should be developed only once we are confident that their effects will be positive and their risks will be manageable.” (ibid.), referring to actual and potential consequences of AI technology, such as the spread of untrue claims or the automation and loss of jobs. Also arguing with the creation of fake content, impersonation of others, and on the assumption that generated text is indistinguishable from that of human authors, OpenAI had initially restricted access to GPT-2 in 2019 (Radford et al., "Better Language Models"). Both the now more than 31,000 signatories of the open letter (as of June 2023) and OpenAI itself argue not *against* the architecture of the models, but *for* the use of so-called security measures. The Future of Life Institute writes in its self-description: “If properly managed, these technologies could transform the world in a way that makes life substantially better, both for the people alive today and for all the people who have yet to be born. They could be used to treat and eradicate diseases, strengthen democratic processes, and mitigate - or even halt - climate change. If improperly managed, they could do the opposite [...], perhaps even pushing us to the brink of extinction.” (Future of Life Institute, “About Us”).

As this depiction richly illustrates, the Future of Life Institute is an organization dedicated to ‘long-termism’, an ideology that promotes posthumanism and the colonization of space (see MacAskill), rather than addressing the multiple contemporary crises (climate, energy, corona pandemic, global refugee movements, and wars) promoted by global financial market capitalism that profoundly reinforce social inequalities. Moreover, "AI doomsaying," i.e., the narrative of artificial intelligence as an autonomously operating agent whose power grows with access to more and more data and ever-improving technology, and whose workings remain inaccessible to human understanding as a black-box, further enhances the influence and power of big tech companies by attributing to their products the power "to remake - or unmake - the world." (Merchant).

On the linking of language models and databases

Taking up criticism of large language models such as the ecological and economic costs of training or the output of unverified or discriminating content, there are debates and frequent calls to develop fundamentally smaller language models (e.g., Schick and Schütze). Among others, David Chapman, who together with Phil Agre developed alternatives to prevailing planning approaches in artificial intelligence in the late 1980s (Agre and Chapman), recently called for the development of the ‘smallest language models possible’: "AI labs, instead of competing to make their LMs bigger, should compete to make them smaller, while maintaining performance. Smaller LMs will know less (this is good!), will be less expensive to train and run, and will be easier to understand and validate." (Chapman). More precisely, language models should "'know' as little as possible-and retrieve 'knowledge' from a defined text database instead." (ibid.). In calling for an architectural separation of language and knowledge, Chapman and others tie in with long-running discussions in phenomenology and pragmatism as well as those in formalism and the Theory of Mind.

Practices of data collection, processing and analysis are ubiquitous. Accordingly, databases are of great importance as informational infrastructures of knowledge production (cf. Nadim). They are not only "a collection of related data organized to facilitate swift search and retrieval" (ibid.), but also a "medium from which new information can be drawn and which opens up a variety of possibilities for shape-making" (Burkhardt, 15, my translation). Lev Manovich, in particular, has emphasized the principle openness, connectivity and relationality of databases (Manovich). In this view, databases appear as accessible and explicit, allowing for an easy interchangeability and expansion of entries, eventually permitting an upscaling of the entire architecture. Databases have been an important component of symbolic AI - also known as Good Old-Fashioned AI (GOFAI). While connectionist AI takes an inductive approach that starts from "available" data, symbolic AI is based on a deductive, logic- and rule-based paradigm. Matteo Pasquinelli describes it as a "top-down application of logic to information retrieved from the world" (Pasquinelli 2). Symbolic AI has become known, among other things , as a representation of ontologies or semantic webs.

Linking external databases with small and large language models emerges as a concrete answer to the problems of generative models, in which knowledge is understood as being ‘embedded’, and which – as illustrated by the example of hallucination – leads to various problems. While connectionist approaches have dominated in recent times, architectures of symbolic AI seem to reappear. The combination of databases and language models is already a common practice and currently discussed under the terms ‘knowledge-grounding’ or ‘retrieval augmentation’ (f.e. Lewis et al.). Retrieval-augmented means that in addition to fixed training datasets, the model also draws on large external datasets, an index of documents whose size can run into the trillions of documents. Meanwhile, models are called small(er) as they contain a small set of parameters in comparison to other models (Izacard et al.). In a retrieval process, documents are selected, prepared and forwarded to the language model depending on the context of the current task. With this setup, the developers promise improvements in efficiency in terms of resources such as the amount of parameters, ‘shots’ (the amount of correct information in the data sets), and corresponding hardware resources (ibid.).

In August 2022, MetaAI has already released Atlas, a small language model that was extended with an external database and which, according to the developers, outperformed significantly larger models with a fraction of the parameter count (ibid.). With RETRO (Retrieval-Enhanced Transformer), DeepMind has also developed a language model that consists of a so-called baseline model and a retrieval module. (Borgeaud et al.). In 2017, ParlAI, an open-source framework for dialog research founded by Facebook in 2017, presented Wizard of Wikipedia, a program – a benchmark task – to train language models with Wikipedia entries (Dinan et al.). They framed the problem of hallucination of, in particular, pre-trained Transformer models as one of updating knowledge. With this program, models are fine-tuned to extract information from database articles to be then casually inserted into a text or conversation without sounding like an encyclopedia entry themselves, thereby appearing semantically and factually correct. With the imagining of small models as ‘free of knowledge’, the focus changes: now not only size and scale are considered a marker of performance, but also the infrastructural and relational linking of language models to external databases. This linking of small language models to external databases thus represents a transversal shift in scale: While the size of the language models is downscaled, the linking with databases implies a simultaneous upscaling.

However, the ideal of an accessible and controllable database falls short where it is conceived as potentially endlessly scalable. It is questionable whether a possibly limitless collection of knowledge is still accessible and searchable or whether it does not transmute into its opposite: "When everything possible is written, nothing is actually said (Burkhardt 11, my translation). What prior knowledge of the structure and content of the database would accessibility require? The conditions of its architecture and the processes of collecting, managing and processing the information are quickly forgotten (ibid. 9f.) and obscure the fact that databases as sites of power also are exclusive and always remain incomplete. Inherent in the idea of an all-encompassing database is a universalism that assumes a generally valid knowledge and thus fails to recognize situated, embodied, temporalized, and hierarchized aspects. Following Wittgenstein, Daston has likewise illustrated that even (mathematical) rules are ambiguous and, as practice, require interpretation of the particular situation (Daston 10).

On disputes over better architectures

The narrative of the opposition of symbolic and connectionist AI locates the origin of this dispute in a disagreement between, on the one hand, Frank Rosenblatt and, on the other, Marvin Minsky and Seymour Papert, who claimed in their book Perceptrons that neural networks could not perform logical operations such as the and/or (XOR) function (Minsky and Papert). This statement is often seen as causal for a cutback in research funding for connectionist approaches, later referred to as the 'winter of AI'. (Pasquinelli 5). For Gary Marcus, professor of psychology and neural science, this dispute between the different approaches to AI continues to persist and is currently being played out at conferences, via Twitter and manifestos, and specifically on Noema, an online magazine of the Berggruen Institute, on which both Gary Marcus and Yann LeCun publish regularly. In an article titled *AI is hitting a wall*, Marcus calls for a stronger position of symbolic approaches and argues in particular for a combination of symbolic and connectionist AI (Marcus, “Deep Learning is Hitting a Wall”). For example, research by DeepMind had shown that "We may *already* be running into scaling limits in deep learning" and that increasing the size of models would not lead to a reduction in toxic outputs and more truthfulness (Rae et al.). Google has also done similar research (Thoppilan et al.). Marcus criticizes deep learning models for not having actual knowledge, whereas the existence of large, accessible databases of abstract, structured knowledge would be "a prerequisite to robust intelligence." (Marcus, “The Next Decade in AI”). In various essays, Gary Marcus recounts a dramaturgy of the conflict, with highlights including Geoff Hinton's 2015 comparison of symbols and aether, and calling symbolic AI "one of science's greatest mistakes " (Hinton), or the direct attack on symbol manipulation by LeCun, Bengio and Hinton in a 2016 manifesto for deep learning published in Nature (LeCun et al.). For LeCun, however, the dispute reduces to a different understanding of symbols and their localization. While symbolic approaches would locate them 'inside the machine', those of connectionist AI would be outside 'in the world'. The problem of the symbolists would therefore lie in the problem of the "knowledge acquisition bottleneck", which would translate human experience into rules and facts and which could not do justice to the ambiguity of the world (Browning and LeCun). "Deep Learning is going to be able to do anything", quotes Marcus Geoff Hinton (Hao).

The term 'Neuro-Symbolic AI', also called the '3rd wave of AI', designates the connection of neural networks – which are supposed to be good in the computation of statistical patterns – with a symbolic representation. While Marcus is being accused of just wanting to put a symbolic architecture on top of a neural one, he points out that there would be already successful hybrids such as Go or chess – which are obviously games and not languages! – and that this connection would be far more complex as there would be several ways to do that, such as "extracting symbolic rules from neural networks, translating symbolic rules directly into neural networks, constructing intermediate systems that might allow for the transfer of information between neural networks and symbolic systems, and restructuring neural networks themselves" (Marcus, "Deep Learning Alone…").

It's not simply XOR

The linking of language models with databases, as shown above, is presented by Gary Marcus, MetaAI and DeepMind, among others, as a possibility to make the computational processes of the models accessible through a modified architecture. This transparency suggests at the same time the possibility of traceability, which is equated with an understanding of the processes, and promises a controllability and manageability of the programs. The duality presented in this context between uncontrollable, nontransparent and inaccessible neural deep learning architectures and open, conceivable and changeable databases or links to them, I want to argue, is fundamentally lacking in complexity. This assumes that the structure and content of databases are actually *comprehensible*. Databases, as informational infrastructures of encoded knowledge, must be machine-readable and are not necessarily intended for the human eye (see Nadim). Furthermore, this simplistic juxtaposition conceives of neural networks as black boxes whose 'hidden layers' between input and output inevitably defies access. In this way, the (doomsaying) narrative of autonomous, independent, and powerful artificial intelligence is further solidified, and the human work of design, the mostly precarious activity of labeling data sets, maintenance, and repair, is hidden from view.

Both the discourse about the *better* architecture and the signing of the open letter by 'all parties' also make clear that the representatives of connectionist AI and those of (neuro-)symbolic AI adhere to a *technical* solution to the problems of artificial intelligence. In either case, the world appears computable and thereby knowable and follows a colonial logic in this regard. Furthermore, the question of whether processes of learning should be simulated 'inductively' by calculating co-occurrences and patterns in large amounts of 'raw' data, or 'top-down' with the help of given rules and structures, touches at its core the 'problem' that the programs have no form of access to the world in the form of sensory impressions and emotions – a debate closely linked to the history of cybernetics and artificial intelligence (see f.e. Dreyfus). With the modeling and constant extension of the models with more data and other ontologies, the programs are built by following an ideal of human-like intelligence. In this perspective, the lack of access to the world is at the same time one of the causes of errors and hallucinations. Accordingly, the goal is to build models that speak semantically correctly and truthfully, while appearing as omniscient as possible, so that they can be easily used in various applications without relying on human correction: the models are supposed to act autonomously. Ironically, the attempt not to make mistakes reveals the artificiality of the programs.

The current hype around generative models like ChatGPT or DALL-E and the monopolization and concentration of power within a few corporations that accompanies it, has seemingly clouded the view for alternative approaches. Tsing's theory provided the occasion to look at the discourse around small, 'knowledge-grounded' language models, which - this was my initial assumption - oppose the imperative of constant scaling-up. Tsing writes that "Nonscalability theory is an analytic apparatus that helps us notice nonscalable phenomena" (Tsing 9). However, the architectures described here do not defy scalability; rather, a transversal shift occurs in that language models are scaled down and databases are scaled up at the same time. The object turned out to be more complex than the mere juxtaposition of scalable and nonscalable.

Conversational AI and generative models in particular are already an integral part of everyday processes of text and image production. The technically generated outputs produce a socially dominant understanding of reality, whose fractures and processes of negotiation are evident in the discussions about hallucinations and jailbreaking. It is therefore of great importance to follow and critically analyze both the technical (‘alternative’) architectures and affordances as well as the assumptions, interests, and power structures of the dominant (individual) actors (Musk, Altman, LeCun, etc.) and big tech corporations that are interwoven with them.

Works cited

- Agre, Philip E., and David Chapman. "What Are Plans For?" *Robotics and Autonomous Systems*, vol. 6, no. 1, June 1990, pp. 17–34. *ScienceDirect*, https://doi.org/10.1016/S0921-8890(05)80026-0.
- Armstrong, Kathryn. "ChatGPT: US Lawyer Admits Using AI for Case Research". *BBC News*, 27 May 2023. *www.bbc.com*, https://www.bbc.com/news/world-us-canada-65735769.
- Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 ". *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2021, pp. 610–23, https://doi.org/10.1145/3442188.3445922.
- Benjamin, Ruha. *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity, 2019.
- Boellstorff, Tom. "Making Big Data, in Theory:. *First Monday*, vol. 18, no. 10, 2013. *mediarep.org*, https://doi.org/10.5210/fm.v18i10.4869.
- Bommasani, Rishi, et al. "On the Opportunities and Risks of Foundation Models". *ArXiv:2108.07258 [Cs]*, Aug. 2021. *arXiv.org*, http://arxiv.org/abs/2108.07258.
- Borgeaud, Sebastian, et al. *Improving Language Models by Retrieving from Trillions of Tokens*. arXiv:2112.04426, arXiv, 7 Feb. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2112.04426.
- boyd, danah, and Kate Crawford. "Critical Questions for Big Data". *Information, Communication & Society*, vol. 15, no. 5, June 2012, pp. 662–79. *Taylor and Francis+NEJM*, https://doi.org/10.1080/1369118X.2012.678878.
- Brockman, Greg [@gdb]. "ChatGPT just crossed 1 million users; it's been 5 days since launch". Twitter, 5 December 2022, https://twitter.com/gdb/status/1599683104142430208.
- Browning, Jacob, and Yann LeCun. "What AI Can Tell Us About Intelligence". *Noema*, 16 June 2022, https://www.noemamag.com/what-ai-can-tell-us-about-intelligence.
- Burkhardt, Marcus. *Digitale Datenbanken: Eine Medientheorie Im Zeitalter von Big Data*. 1. Auflage, Transcript, 2015.
- Cao, Sissi. "Why Sam Altman Won't Take OpenAI Public". *Observer*, 7 June 2023, https://observer.com/2023/06/sam-altman-openai-chatgpt-ipo/.
- Chapman, David [@Meaningness]. "AI labs should compete to build the smallest possible language models…". Twitter, 1 October 2022, https://twitter.com/Meaningness/status/1576195630891819008.
- Crawford, Kate, and Vladan Joler. "Anatomy of an AI System". *Virtual Creativity*, vol. 9, no. 1, Dec. 2019, pp. 117–20, https://doi.org/10.1386/vcr_00008_7.
- Daston, Lorraine. *Rules: A Short History of What We Live By*. Princeton University Press, 2022.
- Dean, Jeffrey. "A Golden Decade of Deep Learning: Computing Systems & Applications". *Daedalus*, vol. 151, no. 2, May 2022, pp. 58–74, https://doi.org/10.1162/daed_a_01900.
- Devlin, Jacob, et al. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171–86, https://aclanthology.org/N19-1423.pdf.
- Dinan, Emily, et al. *Wizard of Wikipedia: Knowledge-Powered Conversational Agents*. arXiv:1811.01241, arXiv, 21 Feb. 2019. *arXiv.org*, http://arxiv.org/abs/1811.01241.
- dpa/lno. "Digitalisierungsminister für Nutzung von ChatGPT". *Süddeutsche.de*, 4 May 2023, https://www.sueddeutsche.de/politik/regierung-kiel-digitalisierungsminister-fuer-nutzung-von-chatgpt-dpa.urn-newsml-dpa-com-20090101-230504-99-561934.
- Dreyfus, Hubert L. *What Computers Can't Do*. Harper & Row, 1972.

- Fazi, M. Beatrice. "Beyond Human: Deep Learning, Explainability and Representation". *Theory, Culture & Society*, vol. 38, no. 7–8, Dec. 2021, pp. 55–77. *SAGE Journals*, https://doi.org/10.1177/0263276420966386.
- Foucault, Michel. *Dispositive der Macht*. Berlin: Merve, 1978.
- Frankfurt, Harry G. *On Bullshit*. Princeton University Press, 2005.
- Future of Life Institute. "About Us". *Future of Life Institute*, https://futureoflife.org/about-us/. Accessed 20 Apr. 2023.
- Future of Life Institute. "Pause Giant AI Experiments: An Open Letter". *Future of Life Institute*, 22 Mar. 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments/.
- Gitelman, Lisa, and Virginia Jackson. "Introduction". *Raw Data Is an Oxymoron*, edited by Lisa Gitelman, The MIT Press, 2013, pp. 1–14.
- Hao, Karen. "AI Pioneer Geoff Hinton: 'Deep Learning Is Going to Be Able to Do Everything'". *MIT Technology Review*, 3 Nov. 2020, https://www.technologyreview.com/2020/11/03/1011616/ai-godfather-geoffrey-hinton-deep-learning-will-do-everything/.
- Hinton, Geoff. "Aetherial Symbols". *AAAI Spring Symposium on Knowledge Representation and Reasoning* Stanford University, CA. 2015.
- Irani, Lilly. "The Cultural Work of Microwork". *New Media & Society*, vol. 17, no. 5, 2013, pp. 720–39. *SAGE Journals*, https://doi.org/10.1177/1461444813511926.
- Izacard, Gautier, et al. *Atlas: Few-Shot Learning with Retrieval Augmented Language Models*. arXiv:2208.03299, arXiv, 16 Nov. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2208.03299.
- Jaton, Florian. *The Constitution of Algorithms: Ground-Truthing, Programming, Formulating*. The MIT Press, 2020.
- Ji, Ziwei, et al. "Survey of Hallucination in Natural Language Generation". *ACM Computing Surveys*, vol. 55, no. 12, Dec. 2023, pp. 1–38. *arXiv.org*, https://doi.org/10.1145/3571730.
- Kitchin, Rob. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Sage, 2014.
- Kaplan, Jared, et al. *Scaling Laws for Neural Language Models*. arXiv:2001.08361, arXiv, 2020, https://doi.org/10.48550/arXiv.2001.08361.
- Knorr Cetina, Karin. *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Pergamon Press, 1981.
- Krizhevsky, Alex, et al. "ImageNet Classification with Deep Convolutional Neural Networks". *Advances in Neural Information Processing Systems*, edited by F. Pereira et al., vol. 25, Curran Associates Inc., 2012, https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- Latour, Bruno, and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, 1979.
- LeCun, Yann, et al. "Deep Learning". *Nature*, vol. 521, no. 7553, May 2015, pp. 436–44. *DOI.org (Crossref)*, https://doi.org/10.1038/nature14539.
- Lewis, Patrick, et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv:2005.11401, arXiv, 12 Apr. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2005.11401.
- Luitse, Dieuwertje, and Wiebke Denkena. "The Great Transformer: Examining the Role of Large Language Models in the Political Economy of AI". *Big Data & Society*, vol. 8, no. 2, 2021, pp. 1–14. *SAGE Journals*, https://doi.org/10.1177/20539517211047734.
- MacAskill, William. *What Is Longtermism?*, https://www.bbc.com/future/article/20220805-what-is-longtermism-and-why-does-it-matter. Accessed 16 June 2023.
- Mackenzie, Adrian. *Machine Learners: Archeology of a Data Practice*. The MIT Press,

2017.

- Manovich, Lev. *The Language of New Media*. The MIT Press, 2001.
- Marcus, Gary. "Deep Learning Alone Isn't Getting Us To Human-Like AI". *Noema*, 11 Aug. 2022, https://www.noemamag.com/deep-learning-alone-isnt-getting-us-to-human-like-ai.
- Marcus, Gary. "Deep Learning Is Hitting a Wall". *Nautilus*, 10 Mar. 2022, https://nautil.us/deep-learning-is-hitting-a-wall-238440/.
- Marcus, Gary. *The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence*. arXiv:2002.06177, arXiv, 19 Feb. 2020. *arXiv.org*, https://doi.org/10.48550/arXiv.2002.06177.
- Marres, Noortje, and David Stark. "Put to the Test: For a New Sociology of Testing". *The British Journal of Sociology*, vol. 71, no. 3, 2020, pp. 423–43. *Wiley Online Library*, https://doi.org/10.1111/1468-4446.12746.
- Martin, Franziska. "OpenAI: Bewertung des ChatGPT-Entwicklers soll auf 30 Milliarden Dollar gestiegen sein". *manager magazin*, 9 Jan. 2023, https://www.manager-magazin.de/unternehmen/tech/openai-bewertung-des-chatgpt-entwicklers-soll-auf-30-milliarden-dollar-gestiegen-sein-a-6ccd7329-bcfc-445e-8b78-7b9d1851b283.
- McQuillan, Dan. "ChatGPT Is a Bullshit Generator Waging Class War". *Vice*, 9 Feb. 2023, https://www.vice.com/en/article/akex34/chatgpt-is-a-bullshit-generator-waging-class-war.
- Merchant, Brian. "Column: Afraid of AI? The Startups Selling It Want You to Be". *Los Angeles Times*, 31 Mar. 2023, https://www.latimes.com/business/technology/story/2023-03-31/column-afraid-of-ai-the-startups-selling-it-want-you-to-be.
- Minsky, Marvin, and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. 2. print. with corr, The MIT Press, 1972.
- Nadim, Tahani. "Database". *Uncertain Archives: Critical Keywords for Big Data*, edited by Nanna Bonde Thylstrup et al., The MIT Press, 2021, 125–132.
- Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, 2018.
- Pasquinelli, Matteo. "Machines That Morph Logic". *Glass Bead*, 2017, https://www.glass-bead.org/article/machines-that-morph-logic/.
- Radford, Alec, et al. "Better Language Models and Their Implications". *OpenAI*, 14 Feb. 2019, https://openai.com/blog/better-language-models/.
- Radford, Alec, et al. "Improving Language Understanding by Generative Pre-Training". *OpenAI*, 2018, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- Rae, Jack W., et al. *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*. arXiv:2112.11446, arXiv, 21 Jan. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2112.11446.
- Rieder, Bernhard, and Yarden Skop. "The Fabrics of Machine Moderation: Studying the Technical, Normative, and Organizational Structure of Perspective API". *Big Data & Society*, vol. 8, no. 2, July 2021. *SAGE Journals*, https://doi.org/10.1177/20539517211046181.
- Schick, Timo, and Hinrich Schütze. "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners". *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2021, pp. 2339–52. *DOI.org (Crossref)*, https://doi.org/10.18653/v1/2021.naacl-main.185<.
- Strubell, Emma, et al. *Energy and Policy Considerations for Deep Learning in NLP*. arXiv:1906.02243, arXiv, 5 June 2019. *arXiv.org*, https://doi.org/10.48550/arXiv.1906.02243.
- Sudmann, Andreas. "On the Media-Political Dimension of Artificial Intelligence: Deep

Culture & Society, vol. 4, no. 1, 2018, pp. 181–200, https://doi.org/10.25969/MEDIAREP/13531.

- Taş, Birkan. "Vulnerability". *Uncertain Archives: Critical Keywords for Big Data*, edited by Nanna Bonde Thylstrup et al., The MIT Press, 2021, pp. 569–78.
- Thoppilan, Romal, et al. *LaMDA: Language Models for Dialog Applications*. arXiv:2201.08239, arXiv, 10 Feb. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2201.08239.
- Tsing, Anna Lowenhaupt. "On Nonscalability: The Living World Is Not Amenable to Precision-Nested Scales". *Common Knowledge*, vol. 18, no. 3, Aug. 2012, pp. 505–24, https://doi.org/10.1215/0961754X-1630424.
- Vaswani, Ashish, et al. "Attention Is All You Need". *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2017, pp. 6000–10.