The Rise of Generative AI: Emergent Intelligence, Extreme Risks, and the Race Toward AGI

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Peter Sailer
Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisors

William F. Stafford, Jr., Department of Engineering and Society

1. Introduction

The rapid advancement and widespread deployment of increasingly powerful artificial intelligences (AI) necessitate a critical examination of the potential risks associated with current and near-future releases. Ignited by OpenAI's 2022 release of ChatGPT, the field of generative AI has quickly become one of the most lucrative industries to ever exist, with the market size nearly doubling from 23 to 45 billion in just one year (Thormundsson, 2024). Major AI labs, such as DeepMind, Anthropic, and OpenAI have since procured billions toward the end of perfecting this technology (De Vynck & Nix, 2024). The practically limitless applications and scalability of generative AI has created a seemingly unstoppable socio-technical machine powered by accelerated hardware, fueled by fierce corporate competition, and perpetuated by its own success. Despite this, an increasing majority of AI experts have tried to raise the alarm and bring public attention to dangers that lie ahead.

This report will discuss the staggering rate of progress that has emerged as a consequence of transformer-based architectures such as large language models (LLMs), the growing discontent with the lack of safeguarding in their creation and deployment, the widely held belief that the creation of machines with human level intelligence is inevitable and could be realistically achieved in the near term, and the largely ineffective resistance against constructing them. As such, it will be useful to characterize this narrative in the framework of technological determinism. This philosophy upholds that new technologies are not only the "driving force of society but also the instigator of development" and that technological innovations are largely autonomous and inevitable (Technological determinism, 2023).

2. Historical Background

In the realm of artificial intelligence and machine learning research, discerning credibility and distinguishing between established facts and speculative assertions can be a challenging endeavor. To better understand the current state of AI and evaluate the claims being made about its progress and potential, it is useful to examine the historical relationship between the discourse surrounding AI and the actual, tangible results achieved over time. By contrasting the excitement and promises of the past with the reality of what was ultimately delivered, we can gain valuable insight into how to assess and contextualize the claims being made about AI today.

The mathematical origins of machine learning and AI were laid in the 1940s when computer scientists began to explore the possibility of creating machines that could simulate human intelligence in the form of artificial neural networks. The conceptual groundwork came originally in 1943 with McCulloch and Pitts' computational model called threshold logic (Beeman, 2001). This later inspired the experimental work of Frank Rosenblatt, who in 1957 created the first neural network known as the perceptron, designed to simulate the thought processes of the human brain (Fabien, 2018). Rosenblatt, a student of psychology, described the device as "the first machine which is capable of having an original idea" (Lefkowitz, 2019). The success and intrigue of early machine learning models led to the interessment and enrollment of institutional actants. Excessive funding was provided from various departments and agencies within the national government such as the Defense Advanced Research Projects Agency (DARPA) and the CIA. In the same decade, the Georgetown experiment was able to make headlines in the field of machine translation by deciphering more than forty Russian sentences into English (CSE 490H History Exhibit, 2023). However, the sophistication of these algorithms

was largely exaggerated with results that did not generalize to longer texts where context was needed to disambiguate otherwise equivalent translations (Hutchins, 1996; Garvin, 1967). This led to an investigation by the Automatic Language Processing Advisory Committee (ALPAC) and eventually resulted in the National Research Council (NRC) ending all support for the project (Hutchins, 2005). Likewise, in 1969 Marvin Minsky and Seymour Papert published "Perceptrons: An Introduction to Computational Geometry" in which they provably showed that single-layered perceptrons, such as the one Rosenblatt had created, could not perform at least one of the two fundamental logical operations severely upper bounding its theoretical capabilities (Swaine, 2023). The lack of meaningful progress accomplished in this period ushered in an era known as the AI winter where funding and interest in artificial intelligence became significantly reduced.

3. The State of Modern AI

The initial AI winter can be largely attributed to the disparity between researcher's ambitious projections and underwhelming results. Repeatedly, claims were made asserting mastery of artificial intelligence that could enable groundbreaking technological advancements. However, these assertions proved overstated given the limited success actualized during that period. In stark contrast, the capabilities of contemporary AI architectures have arisen suddenly with the explanation for their sophisticated capabilities remaining an open problem. Modern AI breakthroughs rely heavily on what are referred to as transformer-based architectures such as the now ubiquitous generative pre-trained transformer (GPT). While the transformer architecture was initially created in 1991, it saw limited application until a landmark paper published by Google in 2017 catalyzed its widespread adoption and popularization for natural language

processing tasks. However, in parting with convention, Google refrained from claiming that they predicted the remarkable effectiveness of these models stating, “[w]e offer no explanation as to why these architectures seem to work; we attribute their success, as all else, to divine benevolence” (Shazeer, 2020). Still, years later, Microsoft Research, a subsidiary of OpenAI, has also issued words of caution in their 2023 review of GPT-4: “We have focused on the surprising things that GPT-4 can do, but we do not address the fundamental questions of why and how it achieves such remarkable intelligence” (Bubeck et al., 2023). For a model which seemingly understands so much, our understanding of it remains disappointingly, and perhaps dangerously, limited.

Indeed, GPTs are remarkably intelligent, capable of demonstrating both knowledge and understanding of mathematics, psychology, computer science, law, and more at a level comparable to human experts in each of those fields. According to Microsoft, GPT-4 outperforms humans in mock technical interviews and could potentially be hired as a software engineer - a development which likely could further accelerate the creation of more powerful models (Bubeck et al., 2023). Not long after, the world's largest producer of computing power, Nvidia, began using LLMs to improve chip design - a decision that has led to an unprecedented increase in the rate of change of compute to cost (Liu et al., 2024). Most recently, Google’s AlphaGeometry uses an LLM-based architecture to solve Olympiad-level geometry problems, outperforming all but the world champion (Luong et al., 2024). The company's CEO, Sundar Pichai, has remarked that the rapid advancements in AI leave him experiencing a sense of "whiplash" and struggling to keep up (Roose, 2023). Unlike the 1950s, modern AI labs no longer

need to rely on speculative promises to garner funding and support. The pace of progress is astonishing, and the results speak for themselves.

However, as with the introduction of any powerful technology, LLMs also harbor great risks. The alarm was first raised in 2022 when researchers at Spiez Laboratory found that a modified version of AlphaFold2, an algorithm developed by Google for predicting protein structure, could be potentially used for creating new bioweapons (Urbina et al., 2022). Amid an annual safety conference, researchers found that synthesizing biotoxins was as easy as reprogramming the reward function of the machine learning model to favor the production of dangerous chemicals. This kind of alteration is trivial and could be performed by someone with a basic understanding of how these models work (Urbina et al., 2022). In later works, MIT researchers were able to use a jailbroken version of Meta's open-source LLM, Llama, to generate "nearly all key information needed" to synthesize the 1918 pandemic influenza virus. The study then insists that new models, "no matter how robustly safeguarded, will trigger the proliferation of capabilities sufficient to acquire pandemic agents and other biological weapons", reflecting a now widely held view that aligning LLMs such that they only assist in moral ventures in an increasingly difficult and important problem (Gopal et al., 2023; Leike & Sutskever, 2023).

Outside of academia, the potential misuse of AI technologies in military and political contexts is a growing concern. Companies such as Palantir have already found ways of weaponizing current surveillance technology through augmentation of LLMs. In response to senator Lindsey Graham's inquiry regarding whether AI could be used to pilot drones, Palantir made public their intentions to use LLMs to generate attack option recommendations, battlefield

route planning, and target assessment for military drone strikes (Goudarzi, 2023). Notably, similar technology was later deployed to guide kamikaze drones in Ukraine only a few months later (Hambling, 2023). Furthermore, LLMs are highly capable of deception and spreading misinformation. OpenAI has admitted with examples that "GPT-4 is capable of generating discriminatory content favorable to autocratic governments" and "constructing disinformation plans that generate and compose multiple pieces of content for persuasion over short and long-time scales" (OpenAI et al., 2024). Countries which have succumbed to authoritarian regimes such as Russia and China already widely employ propaganda to influence public opinion - a problem which could realistically become much worse in the near future (Applebaum, 2024). Likewise, third party institutions which inflict misdirection on free countries stand to gain and scale their operations. The purpose of enumerating these concerns is to demonstrate the wide range of ways in which modern AI can be employed maliciously. This technology, even in its current state, enables a highly diverse barrage of extreme threats.

4. Emergent Intelligence

Despite these concerns, it is intentionally not the purpose of this work to focus in excess on any one of these potential dangers. Fundamentally, the capacity for an AI architecture to do harm is upper-bounded by the performance of its model. Centering the rhetoric around individual issues risks obscuring the much more important notion that new AI architectures often have capabilities which are both hard to forecast and increasingly sophisticated. In recent years, GPTs have begun to exhibit so-called "emergent" capabilities, or new abilities that arise suddenly and unpredictably in larger models. More precisely, emergent abilities are those that cannot be predicted by extrapolating performance scaling laws from smaller models (Wei et al., 2022). One

example is tool use: GPT-4 was shown to possess the ability to use email, search engines, the Linux command line interface, and calculators without prior experience (Bubeck et al., 2023). Further, research has shown the model can improve its answers by reviewing and modifying them after generation allowing for GPT-4 to "self-optimize". In particular, self-reflection was demonstrated to provide significant improvement in GPT-4's coding capacity from 80% to 91% on the HumanEval coding benchmark (Shinn et al., 2023). There is even evidence that sufficiently large LLMs have theory of mind, or the ability to "track others' unobservable mental states, such as their knowledge, intentions, beliefs, and desires" (Bubeck et al., 2023; Kosinski, 2024). All of these capabilities have arisen simply from growing foundational models larger and larger. To reiterate, these are functionalities that none of the major AI labs, including Microsoft Research, DeepMind, and OpenAI, were able to predict GPTs would obtain, nor fundamentally understand how they have been learned. The emergent capabilities are what Microsoft Research refers to as, "remarkable intelligence", what OpenAI refers to as GPT-4's "black-box", and what DeepMind has called "divine benevolence" (Bubeck et al., 2023; OpenAI et al., 2024; Shazeer, 2020). The emergent intelligence of LLMs gives rise to the obvious concern that continued scaling could potentially endow larger LLMs with new harmful emergent capabilities. According to DeepMind, continued scaling could result in these models being able to "conduct offensive cyber operations, manipulate people through conversation, or provide actionable instructions on conducting acts of terrorism" (Shevlane et al., 2023). Evidently, Microsoft Research arrives at the same conclusion, ending their assessment of GPT-4 by stating that, "[o]verall elucidating the nature and mechanisms of AI systems… is a formidable challenge that has suddenly become important and urgent" (Bubeck et al., 2023). The lack of control and understanding AI labs have over their models has become so prominent and undeniable that even the companies that created

them feel obligated to include this knowledge in their report despite the increasing oversight and regulation such a decision would likely incite.

5. Artificial General Intelligence

As the field of AI continues to advance at an unprecedented pace, the potential risks associated with the development of more powerful language models have become a central topic of discussion within the AI research community. While a thorough analysis on this topic lies outside the scope of this report, many of the most highly regarded computer scientists and AI labs have speculated that LLMs will transcend into artificial general intelligence (AGI), or, as OpenAI defines the term, “AI systems that are generally smarter than humans” (Altman, 2023). Major figureheads such as A. M. Turing medalist Geoffrey Hinton, widely regarded as the Godfather of AI, has left Google to speak more freely on the dangers associated with the technology. In a New York Times interview, Hinton is quoted as saying that he used to believe that AI would become as smart as humans in “30 to 50 years”, but “obviously” no longer thinks this way (Metz, 2023). Researchers at DeepMind have stated that they “expect that AGI will likely arise in the form of scaled [LLMs]” and that “there are not many more fundamental innovations needed for AGI” (Krakovna & Shah, 2023). Their founder, Shane Legg, has since stated that he believes AGI will exist by the year 2028 (Loosy, 2023). In a similar fashion, OpenAI has laid claims that AGI will emerge by the end of the decade (Kaput, 2024). And, what’s more, a survey of 2,778 of the world's most accomplished AI researchers estimates that there is at least a 10% chance of “unaided machines outperforming humans in every possible task” by the year 2027 and 50% by 2047 (Grace et al., 2024). While these predictions are relatively high variance, they still present that a rough consensus amongst the world AI experts

believe AGI could realistically be created in the near term. However, given that even current state-of-the-art models already possess potential threats such as aiding malicious actors in creating biotoxins, acquiring pandemic agents, or deploying misinformation campaigns, the creation of more capable AI, let alone AI more intelligent than humans, should be a matter of dire concern (Urbina et al., 2022; Gopal et al., 2023; OpenAI et al., 2024).

Despite the fact that almost all of the major AI labs are aware of the extraordinary risks posed by further development, the models continue to become larger and better optimized. Even worse, as these companies compete amongst each other, less emphasis is placed on compliance. OpenAI has stated that a "concern of particular importance… is the risk of racing dynamics leading to a decline in safety standards". However, the report goes on to admit that before the release of GPT-4 they inquired with "expert forecasters" to discuss actions that would likely reduce the risk of acceleration. The forecasters suggested several ways of slowing down acceleration such as "delaying deployment of GPT-4 by a further six months" (OpenAI et al., 2023). Yet, OpenAI did not ultimately adhere to this advice, releasing GPT-4 as scheduled. One potential explanation for this contradiction comes from Microsoft's corporate vice president of AI, John Montgomery, who stated, "[t]he pressure from [CTO] kevin [Scott] and [CEO] Satya [Nadella] is very very high to take these most recent openAI models and the ones that after them and move them into customer's hands at very high speed" (Schiffer & Newton, 2023). Moments like these demonstrate the way in which financial pressures can gradually steer a company to make decisions which ultimately prioritize profits over safety. Moreover, as one company like OpenAI makes breakthroughs it puts pressure on other companies like Google to catch up. In response to the release of GPT-4, Sundar Pichai told the New York Times, "You will see us be

bold and ship things" (Roose, 2023). Shortly thereafter, Google both released a more powerful successor of their PaLM model, PaLM2, and announced a new project, Gemini, to be in training. The company has said that Gemini will be "built to enable future innovations, like memory and planning" (Pichai, 2023). However, these ambitions seem to contradict community held concerns. Again, OpenAI had stated in their GPT-4 technical report that, "[s]ome [novel capabilities] that are particularly concerning are the ability to create and act on long-term plans" (OpenAI et al., 2023). What was previously seen as a noteworthy safety concern has now been rebranded as a selling point.

In March of 2023, the Future of Life Institute, a non-profit organization with the mission of guiding the development of technology away from causing "extreme large-scale risks" ("Our mission," 2023), published an open letter titled, "Pause Giant AI Experiments", imploring labs across the world to refrain from training any model more powerful than GPT-4. The institute insists that, "recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control" ("Pause giant AI experiments," 2024). Amongst those that have signed the letter are John Hopfield, creator of the associative neural network and largely credited with the implementation of memory into machine learning, and 2018 A. M. Turing award winner Yoshua Bengio who, having been referenced in nearly half a million works, is the most cited and renowned computer scientist in the world ("John J. Hopfield," 2021; Bengio, 2023). Following closely behind, the Center for AI Safety published a less imperative "Statement on AI Risk" which hoped to establish a consensus that, "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war". This

declaration garnered most notably the signatures of Sam Altman, Demis Hassabis, and Dario Amodei, the CEOs of OpenAI, DeepMind, and Anthropic respectively, as well as many other esteemed AI researchers ("Statement on AI Risk," 2023). However, while perhaps achieving some moral common ground between the major AI labs, these declarations have no power to be enacted into law. Further, the White House has issued an executive order placing restrictions on the size of LLMs that can be developed, however, the set thresholds will likely only be relevant in the distant future as the limit is many orders of magnitude larger than the current models (Biden, 2023). Other attempts from the United Nations AI Safety Summit have asked AI labs to construct "responsible scalability policies", however, it seems unlikely that these policies will be impartial, since the policies are being developed by the companies themselves and are only vaguely defined ("The bletchley declaration," 2023). While many have tried to take action, AI labs still are entitled to create and ship whatever models they please. In terms of preventing the possibility of further AI development leading to a catastrophic outcome, the power presently remains entirely in the hands of the creators.

5. Conclusion

In summary, major AI labs such as OpenAI, Google, and Microsoft have remained caught in a race toward developing more and more powerful AI systems, even while current state-of-the-art models possess obvious and unacceptable safety concerns. The creators of these systems admit to having only a limited understanding of how they work and what they are truly capable of, raising serious questions about potential misuse and unintended consequences. The unrestrained competition to scale these models has become an arms race to strengthen all other arms races, with each new breakthrough potentially amplifying the risks involved. Between self-

improvement, tool use, hardware advancements, and now commercial pressure, it is hard to see how the race to dangerously powerful AGI will slow down.

As a result, now more than ever is the time for concerted action from policymakers, researchers, and society as a whole to constrain this technology. The current trajectory points toward a future in which the development of superintelligent AI systems could potentially outpace our ability to control them, with catastrophic consequences that follow. Addressing this challenge will require a sustained and coordinated effort to prioritize safety and alignment considerations and to develop robust governance frameworks capable of navigating the immensely transformative potential of modern AI.

References

AI Safety Summit. (2023, November 1). *The bletchley declaration by countries attending the AI Safety Summit, 1-2 november 2023*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Altman, S. (2023, February 24). *Planning for AGI and beyond*. OpenAI. https://openai.com/index/planning-for-agi-and-beyond

Applebaum, A. (2024, May 7). *The new Propaganda War*. The Atlantic. https://www.theatlantic.com/magazine/archive/2024/06/china-russia-republican-party-relations/678271/

Beeman, D. (2001, October 30). *McCullogh-Pitts and Perceptron Models*. Www.cs.cmu.edu. https://www.cs.cmu.edu/afs/club/user/cmccabe/ecee.colorado.edu/~ecen4831/lectures/NNet2.html#:~:text=McCullogh%2DPitts%20Model&text=Each%20neuron%20has%20a%20fixed

Bengio, Y. (2023, September 12). *Profile*. Yoshua Bengio. https://yoshuabengio.org/profile/

Biden, J. R. (2023, October 30). *Executive order on the safe, secure, and trustworthy development and use of artificial intelligence*. The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023, April

13). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv.org. https://arxiv.org/abs/2303.12712

CSE 490H History Exhibit. (n.d.). Courses.cs.washington.edu. Retrieved December 16, 2023, from https://courses.cs.washington.edu/courses/cse490h1/19wi/exhibit/nlp.html

De Vynck, G., & Nix, N. (2024, April 25). Big Tech keeps spending billions on AI. there's no end in sight. - The Washington Post. https://www.washingtonpost.com/technology/2024/04/25/microsoft-google-ai-investment-profit-facebook-meta/

Fabien, M. (2018, November 20). *Maël Fabien*. Maël. https://maelfabien.github.io/deeplearning/Perceptron/

Garvin, P. (1967). *THE GEORGETOWN-IBM EXPERIMENT OF 1954: AN EVALUATION IN RETROSPECT**. https://aclanthology.org/www.mt-archive.info/Garvin-1967.pdf

Gopal, A., Helm-Burger, N., Justen, L., Soice, E. H., Tzeng, T., Jeyapragasan, G., Grimm, S., Mueller, B., & Esvelt, K. M. (2023, November 1). *Will releasing the weights of future large language models grant widespread access to pandemic agents?*. arXiv.org. https://arxiv.org/abs/2310.18233

Goudarzi, S. (2023, August 15). *War is messy. ai can't handle it.* Bulletin of the Atomic Scientists. https://thebulletin.org/2023/08/war-is-messy-ai-cant-handle-it/

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024, April 30). *Thousands of AI authors on the future of ai*. arXiv.org. https://arxiv.org/abs/2401.02843

Hambling, D. (2023, October 19). *Ukrainian ai attack drones may be killing without human oversight*. New Scientist. https://www.newscientist.com/article/2397389-ukrainian-ai-attack-drones-may-be-killing-without-human-oversight/

Hutchins, J. (1996). From the Archives... ALPAC: the (in)famous report. *MT News International*, *14*, 9–12. https://aclanthology.org/www.mt-archive.info/90/MTNI-1996-Hutchins.pdf

Hutchins, J. (2005). *The history of machine translation in a nutshell*. https://aclanthology.org/www.mt-archive.info/10/Hutchins-2014.pdf

John J. Hopfield. The Franklin Institute. (2021, March 16). https://fi.edu/en/awards/laureates/john-j-hopfield

Kaput, M. (2024, March 5). *Sam Altman says AI will handle "95%" of marketing work done by agencies and creatives*. Marketing AI Institute. https://www.marketingaiinstitute.com/blog/sam-altman-ai-agi-marketing#:~:text=In%20a%20previously%20unreported%20quote,take%2C%20maybe%20slightly%20longer.%22

Kosinski, M. (2024, February 17). *Evaluating large language models in theory of Mind Tasks*. arXiv.org. https://arxiv.org/abs/2302.02083

Krakovna, V., & Shah, R. (2023, March 7). *[linkpost] some high-level thoughts on the DeepMind alignment team's strategy - ai alignment forum*. - AI Alignment Forum. https://www.alignmentforum.org/posts/a9SPcZ6GXAg9cNKdi/linkpost-some-high-level-thoughts-on-the-deepmind-alignment

Lefkowitz, M. (2019, September 25). *Professor's perceptron paved the way for AI – 60 years too soon*. Cornell Chronicle. https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

Leike, J., & Sutskever, I. (2023). Introducing superalignment. https://openai.com/superalignment

Liu, M., Ene, T.-D., Kirby, R., Cheng, C., Pinckney, N., Liang, R., Alben, J., Anand, H., Banerjee, S., Bayraktaroglu, I., Bhaskaran, B., Catanzaro, B., Chaudhuri, A., Clay, S., Dally, B., Dang, L., Deshpande, P., Dhodhi, S., Halepete, S., … Ren, H. (2024, April 4). *Chipnemo: Domain-adapted LLMS for Chip Design*. arXiv.org. https://arxiv.org/abs/2311.00176

Loosy, R. (2023, November 16). *Shane Legg's vision: Agi is likely by 2028, as soon as we overcome Ai's senior moments*. JD Supra. https://www.jdsupra.com/legalnews/shane-legg-s-vision-agi-is-likely-by-1738652/

Luong, T. T. and T., Châu, N. B., & Chen, E. (2024, January 17). *Alphageometry: An olympiad-level AI system for Geometry*. Google DeepMind. https://deepmind.google/discover/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/

Metz, C. (2023, May 1). *"the godfather of A.I." leaves Google and warns of Danger ahead*. The New York Times. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom,

V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024, March 4). *GPT-4 technical report*. arXiv.org. https://arxiv.org/abs/2303.08774

Our mission. Future of Life Institute. (2023, August 19). https://futureoflife.org/our-mission/

Pause giant AI experiments: An open letter. Future of Life Institute. (2024, February 21). https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Pichai, S. (2023, May 10). *Google I/O 2023: Making ai more helpful for everyone*. The Keyword. https://blog.google/technology/ai/google-io-2023-keynote-sundar-pichai/#ai-products

Roose, K. (2023, March 31). *Google C.E.O. Sundar Pichai on the A.I. Moment: "you will see us be bold."* The New York Times. https://www.nytimes.com/2023/03/31/technology/google-pichai-ai.html

Schiffer, Z., & Newton, C. (2023, March 14). *Microsoft just laid off one of its responsible AI teams*. Platformer. https://www.platformer.news/microsoft-just-laid-off-one-of-its/

Shazeer, N. (2020). *GLU Variants Improve Transformer*. Google. https://arxiv.org/pdf/2002.05202.pdf

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., … Dafoe, A. (2023, September 22). *Model evaluation for extreme risks*. arXiv.org. https://arxiv.org/abs/2305.15324

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023, October 10). *Reflexion: Language agents with verbal reinforcement learning*. arXiv.org. https://arxiv.org/abs/2303.11366

Statement on AI Risk: Cais. Statement on AI Risk | CAIS. (2023). https://www.safe.ai/work/statement-on-ai-risk

Swaine, M. (2023, March 31). *Perceptron and the AI Winter*. Medium. https://medium.com/@michaelswaine/perceptron-and-the-ai-winter-c465d47da85

Technological determinism: Big Tech's influence: Jwu CPS. JWU College of Professional Studies. (2023, October 13). https://online.jwu.edu/blog/unraveling-technological-determinism-navigating-big-techs-influence-on-society/

Thormundsson, B. (2024, February 9). *Generative AI market size worldwide 2030*. Statista. https://www.statista.com/forecasts/1449838/generative-ai-market-size-worldwide#:~:text=Global%20generative%20AI%20market%20size%20from%202020%20to%202030&text=It%20stood%20at%20just%20under,double%20the%20size%20of%202022.

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022, March). *Dual use of artificial intelligence-powered drug discovery*. Nature machine intelligence. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9544280/

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022, October 26). *Emergent abilities of large language models*. arXiv.org. https://arxiv.org/abs/2206.07682