Biocomputing as a Solution for Future Sustainable AI Developments

A White Paper Advocating the Development of Biocomputing for Venture Capitalists, AI Researchers, and Chip Manufacturers

Introduction: Artificial Intelligence and Climate Change

The world has shifted into the age of artificial intelligence – an age that began with the release of OpenAI's ChatGPT3 in 2022. Since then, there has been an explosive frenzy for better, larger, and more capable models. Currently, hundreds of advanced AI models have been trained – with popular models including:

- GPT3, GPT4, GPT4 Vision, GPT4 Turbo, Sora, DALL-E by OpenAI
- PaLM, Bard, Gemini Ultra, Pro, and Nano by Google
- CodeLLaMa, LLaMA and LLaMA 2 by Meta
- Claude 2.0, 2.1, and 3.0 by Anthropic
- Falcon 180B and Falcon 40B by Technology Innovation Institute

Over time, it is clear that more and more models will be developed. There is a clear dramatic increase in the number of AI models trained from 2014 to 2024, as shown in Figure 2, compared to Figure 1, which shows the number of AI models trained from 1950 to 2014[1].

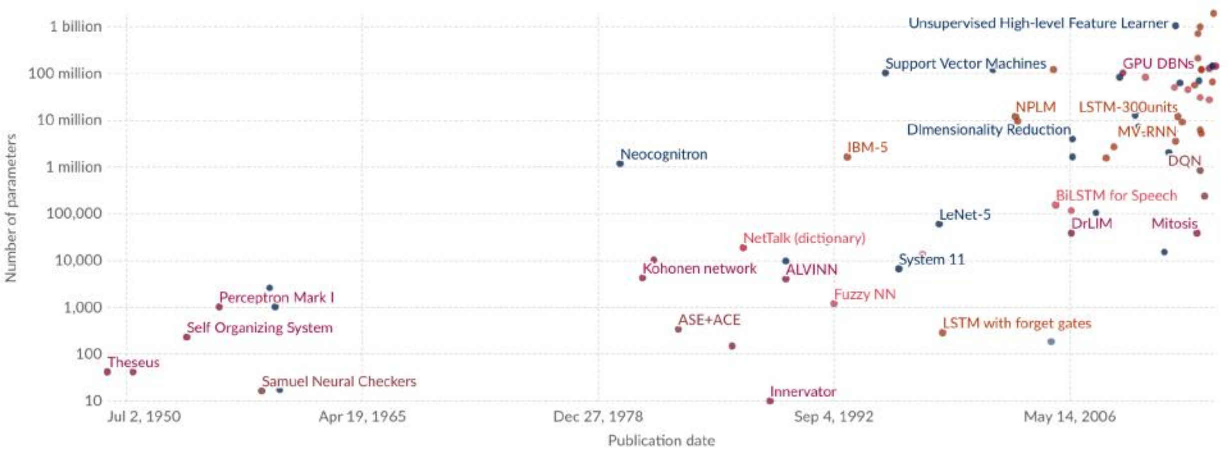

Figure 1. Number of AI Models Trained from 1950 to 2014.

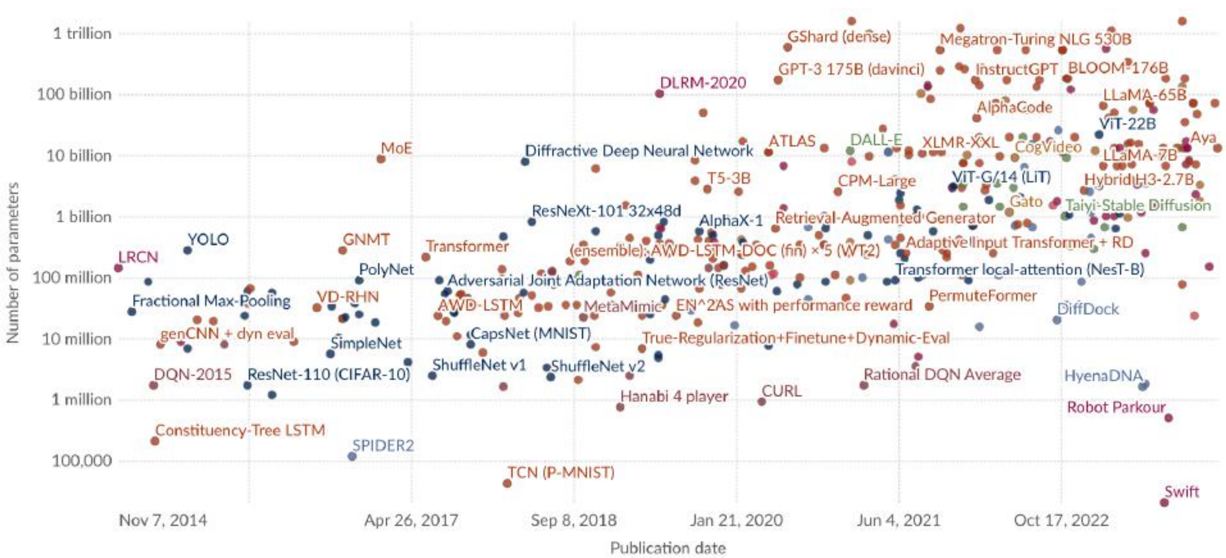

Figure 2. Number of AI Models Trained from 2014 to 2024.

—

1. Parameters in Notable Artificial Intelligence

The Age of Artificial Intelligence has arrived, and it's here to stay. However, the acceleration of AI developments has a hidden consequence: the acceleration of global climate change.

Already, the rise of AI in the past 2 years alone have led to a significant growth in data center utilization. Across all AI models from 2012 to 2022, the amount of computing power (provided by data centers) required grew by about 1 million percent, according to Princeton Professor Varma[2]. In fact, the amount of compute required by current AI systems and models seems to double every 6 months[3], while OpenAI's own research suggests that the compute required has been doubling every 3.4 months since 2012[4]. And to access more compute, more energy is required, resulting in increased carbon dioxide emissions that drastically exacerbate the climate crisis. For example, increased AI-related data center usage increased Amazon's carbon emissions by 15% in 2023[5].

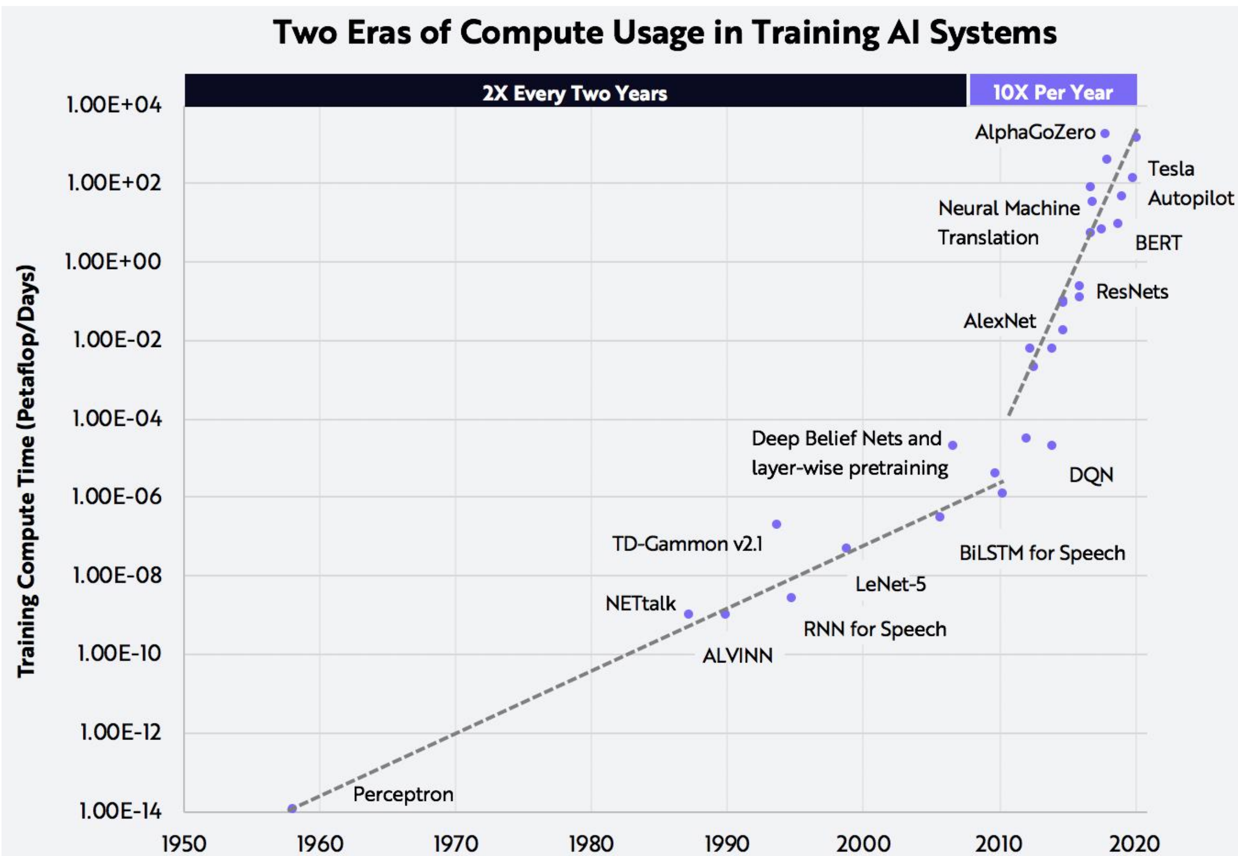

Figure 3. Exponential Growth in Compute Usage for AI Training.

According to the International Energy Agency, the proportion of global electricity consumption attributable to data centers, along with their associated carbon emissions, tripled from 2% to 6% between 2018 and 2020[6].

—

2. New Chip Built for AI Workloads
3. Computational Power and AI
4. AI and Compute
5. Carbon Impact of Artificial Intelligence
6. Electricity 2024 – Analysis

Should this growth pattern persist, projections indicate that by 2030, data centers could account for between 8% and 21% of worldwide energy consumption, precipitating a crisis in global energy supply. In a comprehensive analysis published in Joule, Dr. Vries projected that, by 2027, the annual electricity consumption of AI related technologies could reach at least 85 terawatt-hours[7]. For context, the total annual electricity consumption of the Netherlands is roughly 122 terawatt-hours[8].

Findings from the University of Massachusetts Amherst also highlight AI as a considerable contributor to carbon emissions. To train a moderately-sized language model – typically 40 billion parameters – approximately 300,000 kilograms of carbon dioxide is released[9]. Yet, OpenAI's GPT-4 and Google's Gemini Ultra models were launched with over 1,760 billion and 1,560 billion parameters, respectively. More recently, the release of Claude's Opus model in 2024 introduced a model with over 2,000 billion parameters, underscoring the escalating environmental impact of advanced AI technologies.

Put simply, AI is compute bound. And the problem is, for a linear gain in AI performance, an exponential amount of parameters and thus compute is required[10]. And that means exponentially more carbon emissions and energy consumption[11]. As of now, the field of AI remains unsustainable. **Thus, AI training, deployment, and research needs to be made significantly more energy efficient to align with climate change goals.**

This whitepaper aims to address how AI will exponentially exacerbate the global climate crisis, point to current primary solutions funded by venture capital or researched by AI and chip companies to reduce AI's environmental impact, and highlight biocomputing as an alternative solution that should attract greater research and funding focus.

—

7. Growing Energy Footprint
8. Electricity Generation in the Netherlands
9. Energy and Policy Considerations
10. High Cost of AI Compute
11. Carbon Impact of Artificial Intelligence

Problem: Current Silicon Hardware Solutions to Create Sustainable AI is Unfeasible

As AI grows more elaborate and data-intensive, two things begin to scale up exponentially: the need for more memory storage and the need for more energy. Based on OpenAI's GPT models, the number of parameters used by each successive model increases by roughly 10x every 2-3 years, indicated in Table 1.

Year	Model	Parameters
2018	GPT1	117M
2019	GPT2	1.5B
2020	GPT3	175B
2023	GPT4	1760B

Table 1. Release Date of GPT Models and their Associated Parameter Count.

And across all companies developing AI models, it is clear that the trend of increasing AI model size is likely to continue, as shown in Figure 4[12].

Figure 4. Exponential Increase in Parameter Count for Recently Developed AI Models.

—

12. Parameters in Notable Artificial Intelligence

The current strategy pursued to address this issue involves the creation of more energy-efficient silicon hardware chips. The evolution of traditional silicon hardware has been historically driven by two foundational principles:

1. Moore's Law, which posits that the number of transistors on a microchip doubles approximately every two years, thereby increasing its processing power.
2. Dennard's Scaling Law, which suggests that as chips get smaller, they become more energy efficient.

These developments have resulted in the production of silicon chips that possess the capability to efficiently train and operate AI technologies.

But Moore's Law and Dennard's Scaling Law have slowed[13-15]. Each individual transistor is so small – smaller than a virus – that chip manufacturers are breaking the fundamental properties of physics[13]. And thus, improving traditional silicon hardware to make better, denser, and more efficient chips have also slowed dramatically in progress[13].

Taking into account that the number of transistors on a chip doubles every 3-4 years now coupled with the fact that current AI models require the doubling of compute every 3-6 months, silicon computing can no longer keep up with the rapid requirements of AI. To fix this, companies use massive amounts of chips to train AI models, which in return requires massive amounts of energy to operate.

Another problem with silicon chips is silicon itself. An estimate from the Semiconductor Research Corporation posits that if the current exponential growth in AI's data usage persists, relying on memory components made of silicon, global demand for silicon would soon exceed the annual worldwide production of the material, leading to shortages and bottlenecks in future silicon chip manufacturing[16].

Current silicon chips also cannot bypass the von Neumann bottleneck[13-15]. The core architecture of virtually every digital computer has followed a simple pattern first developed in the 1940s, known as the von Neumann model: store data in one place, compute the data in another, then shuttle information between the two places. Because these currently exist in two separate locations, electricity needs to travel a non-trivial distance to facilitate computation, which makes silicon chips energy and time inefficient, driving increased energy consumption and thus carbon emissions that accelerate global climate change.

A final challenge with silicon hardware is the general release of heat as a byproduct of computing. Current Nvidia GH200 AI chips use kilowatts of energy to operate. In a typical server rack, that means 21-24 kilowatts of thermal load needs to be dissipated – resulting in massive energy costs just to cool the silicon hardware – typically 40-50% of a data center's electricity consumption, according to Figure 5[15]. Moreover, AMD's latest AI accelerator chip has seen a power consumption increase from 560W to 760W, while Nvidia's upcoming chips are anticipated

—

13. Hidden Costs of AI
14. Brain Organoid Reservoir Computing
15. Thermal Management
16. Plan for Semiconductors

to require more than 1000W. This trend towards higher power consumption. suggests that chips will generate more heat, a pattern expected to persist. Consequently, it is estimated that global data centers will experience a 50% increase in energy consumption solely for the purpose of cooling the silicon[17].

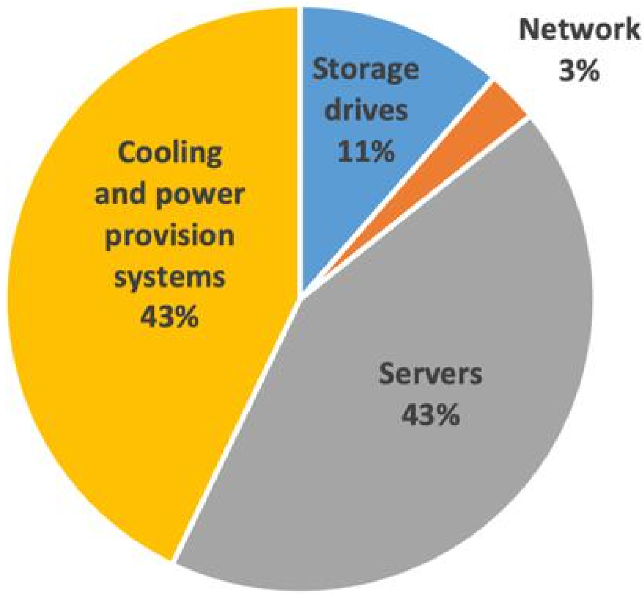

Figure 5. Division of Electricity Consumption in Data Centers.

Given the mounting challenges faced by the current silicon-based computing paradigm, it is imperative for deeptech investors and chip manufacturing companies to recognize that continuing along this trajectory is unsustainable for future AI developments. The exponential growth in AI's complexity and resource requirements demands a radical shift in the global approach to computing. To sustain the rapid advancement of AI technologies and their applications, investors and companies must explore and invest in alternative computing architectures that promise greater efficiency, scalability, and environmental sustainability. A new paradigm of computing that can meet the ambitious demands of tomorrow's AI is required.

—

17. Shocking Amount of Electricity

Solution: The Convergence of Silicon and Biology

The next amazing revolution is going to be digital biology. For the first time in human history, biology has the opportunity to be engineering, not science. With advances in cell culturing techniques, bioengineering technologies, and AI, a biologics-based computing paradigm is poised to emerge as a viable, efficient, scalable, and sustainable solution to the requirements of AI technologies[18].

The human brain is a complex, three-dimensional massively parallel network of over 200 billion cells, linked by hundreds of trillions of synapses, and expends only 20 watts while current silicon hardware consumes about 8 million watts to power a device of similar computational strength[19].

A biologics-based computing model offers profound advantages over traditional silicon-based systems. One benefit of biologics-based computing is its lack of heat byproduct generation. Unlike silicon chips that generate significant heat and require extensive cooling mechanisms, biological neurons operate without producing substantial thermal energy. This characteristic eliminates the need for the thousands of watts typically consumed in cooling processes, dramatically reducing the overall energy and carbon footprint of computing operations[18-20].

Moreover, through four billion years of evolution, neurons have evolved to become highly efficient in both performance and energy use. The unique architecture and operational mechanisms of neurons allow for more effective processing and storage of information with minimal energy requirements. Neurons also bypass the limitations posed by the von Neumann architecture prevalent in current silicon systems. Neurons serve dual functions as both processors and memory units, enabling faster data processing without the need to shuttle information back and forth between separate components. This integration facilitates a more efficient computation process, which is necessary for the complex and data-intensive tasks involved in AI training[20].

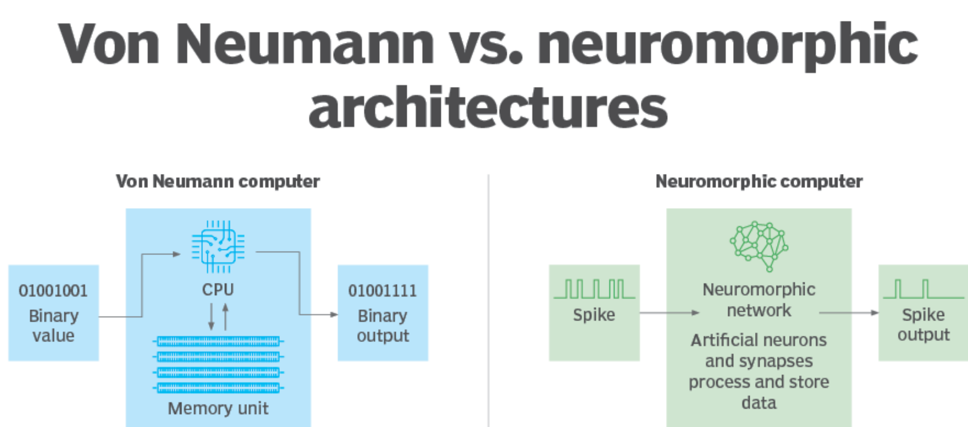

Figure 6. Comparison of von Neuman and Neuromorphic Computing Architecture.

—

18. Reservoir Computing
19. Brain Organoid Reservoir Computing
20. Bridging Biological and Artificial

Finally, sustainability is another compelling advantage of a biologics-based computing paradigm. Neurons, unlike silicon, do not need to be mined from the ground and processed at complex manufacturing processes that contribute to environmental degradation. Neurons can be grown from biological sources, such as cells derived from rats, or replicated through advanced biological processes researchers have developed in the past century. This method of production is infinitely sustainable, providing a renewable resource for computing that lessens the environmental impact associated with the extraction and processing of silicon and other materials used in traditional computing hardware[21].

Already, current biologics-based chips have been used in rudimentary computer vision[21] and speech recognition tasks[22-23]. In one study, researchers converted audio clips into spatiotemporal sequences of bipolar pulse stimulations to the brain organoid. The evoked neural activity was recorded and fed into a logistic regression function for classification, then trained and optimized.

A future goal is to approach one-shot learning in neurons similar to human learning experiences. For example, traditional deep learning models are successful when there is a large, labeled, and unnoisy dataset. But humans are adept at learning from just one or few instances of data[24]. This is due not only to the computational power of human brains, but also to the ability to synthesize and learn new object classes from limited information about different, previously learned classes. One-shot and few-shot learning aim to achieve a similar goal. spatiotemporal sequences of bipolar pulse stimulations to the brain organoid. The evoked neural activity was recorded and fed into a logistic regression function for classification, then trained and optimized.

—

21. Face Classification
22. Analogue Signal
23. Temporal Data
24. Bridging Biological and Artificial

Conclusion

That's not to say AI and advancing it needs to stop because it's incredibly useful for important applications like accelerating the discovery of therapeutics. The world just needs to remain cognizant of the effects and keep pushing for more sustainable approaches to design, manufacturing, and consumption.

NEED TO WORK ON MORE

Bibliography

"AI and Compute." 2018. OpenAI. May 16, 2018. https://openai.com/research/ai-and-compute.

Appenzeller, Guido, Matt Bornstein, Martin Casado, Guido Appenzeller, Matt Bornstein, and Martin Casado. 2023. "Navigating the High Cost of AI Compute." Andreessen Horowitz. November 15, 2023. https://a16z.com/navigating-the-high-cost-of-ai-compute/.

Cai, Hongwei, Zheng Ao, Chunhui Tian, Zhan Hao Wu, Hongcheng Liu, Jason Tchieu, Mingxia Gu, Ken Mackie, and Feng Guo. 2023. "Brain Organoid Reservoir Computing for Artificial Intelligence." *Nature Electronics* 6 (12): 1032–39. https://doi.org/10.1038/s41928-023-01069-w.

"Computational Power and AI." 2023. AI Now Institute. October 11, 2023. https://ainowinstitute.org/publication/policy/compute-and-ai#:~:text=This%20trend%20has%20borne%20out,in%20only%205.7%20months%2012.

De Vries, Alex. 2023. "The Growing Energy Footprint of Artificial Intelligence." *Joule* 7 (10): 2191–94. https://doi.org/10.1016/j.joule.2023.09.004.

"Decadal Plan for Semiconductors - SRC." 2021. January 2021. https://www.src.org/about/decadal-plan/.

Dhar, Payal. 2020. "The Carbon Impact of Artificial Intelligence." *Nature Machine Intelligence* 2 (8): 423–25. https://doi.org/10.1038/s42256-020-0219-9.

"Electricity 2024 – Analysis - IEA." 2024. International Energy Agency. January 2024. https://www.iea.org/reports/electricity-2024.

Leffer, Lauren. 2024. "The AI Boom Could Use a Shocking Amount of Electricity." Scientific American. February 20, 2024. https://www.scientificamerican.com/article/the-ai-boom-could-use-a-shocking-amount-of-electricity/.

Li, Can, Miao Hu, Yuning Li, Hao Jiang, Ning Ge, Eric Montgomery, Jiaming Zhang, et al. 2017. "Analogue Signal and Image Processing With Large Memristor Crossbars." *Nature Electronics* 1 (1): 52–59. https://doi.org/10.1038/s41928-017-0002-z.

Mann, Tobias. 2023. "How thermal management is changing in the age of the kilowatt chip." *The Register*, December 18, 2023. https://www.theregister.com/2023/12/26/thermal_management_is_changing/.

Moon, J. W., Wen Ma, Jong Hoon Shin, Fuxi Cai, Chao Du, Seung Hwan Lee, and Wei Lü. 2019. “Temporal Data Classification and Forecasting Using a Memristor-based Reservoir Computing System.” *Nature Electronics* 2 (10): 480–87. https://doi.org/10.1038/s41928-019-0313-3.

“New Chip Built for AI Workloads Attracts $18M in Government Support.” 2024. Princeton University. March 6, 2024. https://www.princeton.edu/news/2024/03/06/new-chip-built-ai-workloads-attracts-18m-government-funding-revolutionary-tech.

“Parameters in Notable Artificial Intelligence Systems.” 2024. Our World in Data. March 4, 2024. https://ourworldindata.org/grapher/artificial-intelligence-parameter-count#sources-and-processing.

Smirnova, Lena, Brian Caffo, and Erik C. Johnson. 2023. “Reservoir Computing With Brain Organoids.” *Nature Electronics* 6 (12): 943–44. https://doi.org/10.1038/s41928-023-01096-7.

Statista. 2023. “Electricity Generation in the Netherlands 2001-2022.” July 12, 2023. https://www.statista.com/statistics/450429/netherlands-electricity-generation/.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. “Energy and Policy Considerations for Deep Learning in NLP.” arXiv.Org. June 5, 2019. https://arxiv.org/abs/1906.02243v1.

Tang, Jianshi, Fang Yuan, Xinke Shen, Zhongrui Wang, Mingyi Rao, Yanlin He, Yuhao Sun, et al. 2019. “Bridging Biological and Artificial Neural Networks With Emerging Neuromorphic Devices: Fundamentals, Progress, and Challenges.” *Advanced Materials* 31 (49). https://doi.org/10.1002/adma.201902761.

“The Hidden Costs of AI: Impending Energy and Resource Strain | Penn Today.” 2023. Penn Today. March 8, 2023. https://penntoday.upenn.edu/news/hidden-costs-ai-impending-energy-and-resource-strain.

Yao, Peng, Huaqiang Wu, Bin Gao, Şükrü Burç Eryilmaz, Xueyao Huang, Wenqiang Zhang, Qingtian Zhang, et al. 2017. “Face Classification Using Electronic Synapses.” *Nature Communications* 8 (1). https://doi.org/10.1038/ncomms15199.

Logical Outline

Premise

- Given that AI has undergone recent explosive growth in the past 5 years
- Given that AI development shows no sign of slowing down and is actually accelerating
- Given that AI research requires computers (GPUs) to train, which requires energy
- Climate change is a major problem which is mainly caused by carbon emissions from human activities

Proposition

- Thus, AI training, deployment, and research needs to be made significantly more energy efficient to align with climate change goals.

Reasons & Evidence

1. (How) Current silicon chip design should be abandoned in favor of newer technologies
    a. Silicon chips are not inherently energy efficient due to the memory gap problem
    b. Silicon is unsustainable as the world enters into a silicon shortage
    c. Silicon chips requires massively amounts of energy in order to cool
    d. Silicon technology is reaching the limits of physics and can no longer be advanced at a rate suggested by Moore's Law
2. (How) Focusing on massive clean-energy supplying technologies is a long-term goal that will make AI training more climate friendly
    a. Developments in solar and nuclear technologies are decades away from feasible scalability
    b. Clean energy from solar and nuclear technologies will be a long-term solution for supplying clean energy for data centers
    c. Clean energy from solar and nuclear technologies is not favored or implemented by companies currently due to high upfront costs and low efficiencies and a lack of pressure to reduce their climate impact
3. (How) Biologics based chips should be developed for AI specific training and deployment
    a. Biologics-based chips are inherently 10000x more efficient than silicon chip
    b. Biologics-based chips uses less silicon and is thus sustainable in the long term future
    c. Biologics-based chips do not generate heat and thus use less energy
    d. Biologics-based chips have undergone 4 billion years of evolution, resulting in both energy efficiencies and fast processing power

Rhetorical Outline

- Proposition: AI training, deployment, and research needs to be made significantly more energy efficient to align with climate change goals.
- Audience: Chip companies, deeptech venture capitalists
- Genre: White paper
- Motive of the Author: To promote biologics-based computing for investment and R&D
- Motive of the Reader: To recognize the climate problems caused by AI & to fund biologics-based computing and support climate tech solutions in the long-term
- Plan: Publish as a company white paper research (similar to Bitcoin's whitepaper: https://bitcoin.org/bitcoin.pdf), publish in TechCrunch or related site
- Rhetorical Strategies: No idea what this means!
- Keywords: AI, climate, energy, biologics-based computing, biocomputing, neuromorphic computing