Response to the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO), National Science Foundation Request for Information on the Development of an Artificial Intelligence (AI) Action Plan

Submitted by:
Simon Szykman, Ph.D.
President
Cambio Digital Transformations
1278 Woodbrook Ct.
Reston, VA 20194
[redacted]

This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the AI Action Plan and associated documents without attribution.

Moore's Law and the Future Economics and Bifurcation of the AI Market

Bottom Line Up Front: The trends associated with computing power and the accompanying history of the computing industry tell a compelling story that is highly relevant to the emerging AI market. It is essential for policy makers, funders of R&D, corporate technology leaders, and investors to make decisions based on an understanding of these trends, and a recognition that the evolution of AI technology will, if anything, accelerate more rapidly than it did with traditional computers. Failing to recognize the inflection point that lies ahead could result in policies or investments that incentivize suboptimal behaviors, decisions that are misaligned with where the market is going, or placing bets that fail to capitalize on the true potential of AI technologies. And here's why…

Intel co-founder Gordon Moore's most famous observation-turned-prediction was that the number of transistors on a microchip was doubling and would continue to do so approximately every two years. Now known as Moore's Law, this trend has evolved and is often referenced in ways that are more general than just referring to quantity of transistors, but has nevertheless turned out to be a consistently effective characterization of the exponential growth of computing power. (For the curious, at the high end of performance, Top500.org shows trends of supercomputer speeds for the past 3+ decades at https://top500.org/statistics/perfdevel/.)

In the domain of AI, we are now in an era where AI performance, across varying measures of performance ranging from the practical (cost) to the technical (training floating point operations per second) to the abstract ("intelligence"), appears to be increasing at a rate that exceeds that of Moore's Law. I am by no means the first person to note this trend – it has been publicly discussed by academic researchers, OpenAI CEO Sam Altman, as well as NVIDIA CEO Jensen Huang, who referred to the trend as with a catchy term: "Hyper Moore's Law". This rapid pace of change has led to important outcomes, such as significantly greater AI compute capacity, more complex and sophisticated (i.e., "intelligent") AI models, increasing efficiencies and decreasing costs for a unit/fixed amount of AI capability, and increasingly rapid development/training/release cycles for new and more capable models. As a result of these improvements, AI technology has become far more accessible, leading to new use cases, and applications of AI in ways that had not been previously contemplated.

On the research side of the computing industry, investments continue to be made in the development of new technologies to sustain further advancements. Just as GPUs evolved into the foundation of many AI systems when traditional CPUs could not support increasingly high AI workloads, new types of specialized hardware are being developed to support modern and future needs. These include things like tensor processing units,

neuromorphic chips, other new technologies still under development, and almost certainly future ones that have not yet been conceived.

There is little doubt that AI will have transformational impacts on society by driving improvements in virtually every sector of the economy. These changes will contribute to significant economic growth as well as evolution of entirely new types of value creation. The untapped potential has yielded massive jumps in valuations of companies in the AI market, and has led to planned investments on the scale of tens of billions of dollars for hardware and data centers to support this rapidly accelerating market. Some reporting suggests that the trending of these infrastructure investments is heading toward trillions of dollars (though conversely, there are also early indications that the market may have overestimated the future demand for AI datacenter capacity).

But interestingly, while there is plenty of talk about growth in the AI market, there is not much in the way of analytical projections of the economics of this market viewed through the lens of Moore's (or perhaps more accurately Hyper Moore's) Law. The history of the computing market has important implications for the future of the AI market. More specifically, past computing trends suggesting a future time where there might be a downward inflection in some of the AI-driven economic trends, or when the market might split into two separate paths – one for commodity and another for high-performance – as it has with the traditional computing market.

It has been said that the modern smartphone has more processing power than the computers that landed people on the moon or that were in the space shuttle. This was already true for earlier generations of smartphones that existed years ago, so let's look at comparisons with some more modern commodity hardware. One variant of the new Qualcomm Snapdragon 8 Elite processor used in some of today's flagship smartphones has a GPU benchmark speed of over 3.6 trillion floating point operations per second (teraflops, or TFLOPS). Twenty years ago, fewer than 100 of the fastest supercomputers in the world had the processing power that people can now hold in their hands. The high-end NVIDIA GeForce RTX 5090 graphics card has a benchmark speed of 314 TFLOPS. Ten years ago, fewer than 220 of the fastest supercomputers in the world had the processing power that consumers can now buy for their gaming computers. I acknowledge that not all benchmarks are the same, so please grant me some literary (or should I say numerical) license, since it would take only a shift in the number of years cited for the general point to remain valid.

Let me offer some thoughts on what this means for the AI market of the future. Supercomputers existed twenty years ago, but the needs driving the market for the fastest supercomputers in the world were not driving the consumer market for computers. The same was true ten years ago, and still today the market for supercomputers exists, but is not driving the consumer computer market. There is an extremely wide span between the economics of (both investments required for and profits emerging from) building computers at the highest of the high end vs. the commoditized market where the average consumer makes purchases. I believe we will see very similar trends in the AI market. Numerous companies today have built AI capabilities that the average person can access via their computer or smartphone. Massive investment has gone into making this possible, and one needs only an internet connection to reach the massively-powered (and costly) infrastructures that deliver these capabilities.

Following a Moore's Law trend for AI technologies that parallels the trend of computers, one could reasonably assume that in 10 years people may have the same capability running locally on their desktop computers with consumer-level investment, and in 20 years those capabilities may run locally on a smartphone at a cost point that is absorbed into the cost of a phone (just as you don't pay extra for a GPU when you buy a smartphone today). Now take note once again that the AI technology trends have followed the so-called Hyper Moore's Law trajectory – faster than Moore's Law – and consider that the progression from massive computational infrastructures to PC cards to chips on a smartphone might happen much more quickly for AI technologies.

A handful of years from now, there will almost certainly continue to be a need for computing infrastructures that run uniquely high-end AI workloads (analogous to the specialized applications that drive the need for today's supercomputers), and there may be significant investment going into those systems. But that market will be an offshoot that splinters off from a much larger commoditized consumer-driven market for AI. Just as today's consumers have no need for the power of today's supercomputers (at least not at the price point they come with), there will be a time not too far off when the vast majority of people don't need more than what they have available from a machine on their desk, or in the palm of their hand. And as was the case with the computing market, the bulk of the investment and profits will be driven by the consumer market rather than the most capable of the high-end systems.

Today, there is a race to achieve faster, more accurate, more “intelligent” AI systems. But once AI becomes fast enough and intelligent enough, the market will reach an inflection point, and the race will shift to one that aims to commoditize “good enough”, make it less expensive, put it on a PC card, and then put it on a chip. In the long term, there will continue to be companies that make massive infrastructure investments to stay at the bleeding edge – the top 10 or top 100 fastest, most accurate, most intelligent AI systems. But those are not necessarily the companies that will reap the biggest rewards from the market. The companies that do will be the ones that identify when the point of “good enough” has been reached, and instead work to drive the cost of AI to where it becomes a significant feature that is accompanied by a modest if not negligible cost.

Reiterating the bottom line (now at the bottom): The trends associated with computing power and the accompanying history of the computing industry tell a compelling story. It is essential for policy makers, funders of R&D, corporate technology leaders, and investors to make decisions based on an understanding of these trends, and a recognition that the evolution of AI technology will, if anything, accelerate more rapidly than it did with traditional computers. Failing to recognize the inflection point that lies ahead could result in policies or investments that incentivize suboptimal behaviors, decisions that are misaligned with where the market is going, or placing bets that fail to capitalize on the true potential of AI technologies.

Government thought leaders involved in drafting the **AI Action Plan** should not only anticipate that this change will occur at some point, but should draft a plan built around the expectation that it will, and should be designed in a way that can rapidly react to both evolving technologies and a changing market for AI.

About the Author

Simon Szykman, Ph.D., is the co-Founder and President of Cambio Digital Transformations, where he is leveraging nearly 30 years of technology and business leadership experience to carry out leading-edge technology-enabled transformations of how clients deliver their mission and services to stakeholders.

Prior to founding Cambio, Simon served as the Senior Vice President for Client Growth at Maximus, where he was responsible for leading a 45+ person growth team and enabling the go-to-market strategy for the company’s Federal Services Segment. He joined Maximus when the company acquired Attain, where Simon was a Partner and Chief Technology Officer.

Prior to joining Attain, Simon spent nearly 20 years in diverse government technology roles. His previous positions include serving as the Chief Information Officer (CIO) of the U.S. Department of Commerce, CIO of Commerce’s National Institute of Standards and Technology (NIST), two stints at the White House Office of Science and Technology Policy (including serving as the Director of the National Coordination Office for Networking and IT R&D), and the first Director of Cyber Security R&D at the Department of Homeland Security.

About the Company

Cambio Digital Transformations is a team of visionary technologists, strategic consultants, and government innovation experts who empower our clients to imagine the art of the possible, and then bring their imagination to reality. Building on years of digital transformation experience, Cambio synthesizes a fusion of leading-edge technical capabilities, and an intense focus on customer experience, and human-centered design, to help fundamentally transform how our clients do what they do.