GW&K INVESTMENT MANAGEMENT

WILLIAM P. STERLING, PH.D.

Global Strategist at GW&K

GLOBAL PERSPECTIVES

OCTOBER 2024

TAMAY BESIROGLU

Associate Director of ***Epoch AI***

WHAT'S NEXT FOR AI: INTERVIEW WITH AI RESEARCHER TAMAY BESIROGLU

HIGHLIGHTS

- AI progress is largely driven by increased computing power, with training compute, or resources required to train AI models, growing at about 4x per year amid plans for $100 billion AI clusters that could revolutionize the job market.
- Power availability is emerging as an important constraint for AI training, with recent large models requiring power equivalent to the consumption of thousands of households.
- AI's potential to automate cognitive work could boost economic growth substantially, but regulatory hurdles and unforeseen bottlenecks may slow progress.

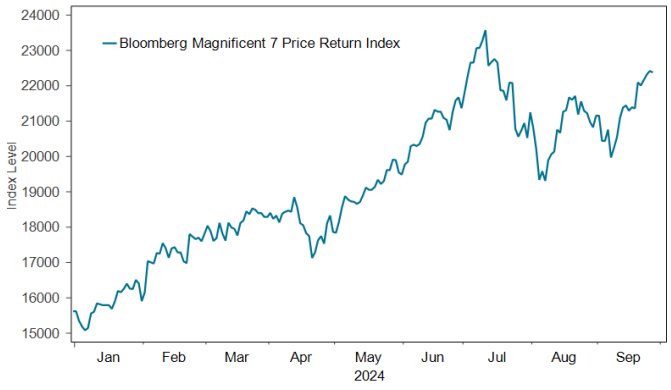
EXPLORING THE LONG-TERM PROSPECTS OF AI-DRIVEN GROWTH

AI fever gripped the US equity market in the first half of this year as mega-cap tech stocks soared while smaller companies posted less impressive returns. But some investors have started to question the durability of the AI growth theme. For example, in late June Goldman Sachs published a piece with a self-explanatory title called "Gen AI: Too Much Spend, Too Little Benefit?"[1] This piece also highlighted the views of MIT economist Daron Acemoglu who predicts only limited economic upside from AI over the next decade.

Since mid-July, the so-called Magnificent 7 Index of tech-driven leaders (Alphabet, Amazon, Apple, Meta, Microsoft, Nvidia, and Tesla) experienced a sharp correction before rebounding. As of late September, the index stands at a 43% year-to-date gain, having recovered most of its losses (**Figure 1**). This volatility raises questions about the sustainability of the AI-enabled growth surge and its long-term trajectory. For some perspective on what's next for AI, I emailed some relevant questions to AI researcher Tamay Besiroglu, the Associate Director of Epoch AI, a highly respected AI research firm.

FIGURE 1

The "Mag 7" Index Has Been a Strong Performer in 2024 But Was Notably Volatile in the Third Quarter



Source: GW&K Investment Management, BLS, Cleveland Fed, and Macrobond

The Magnificent 7 Index is an equally weighted equity index based on seven mega-cap technology-driven stocks: Alphabet, Amazon, Apple, Meta, Microsoft, Nvidia, and Tesla.

BILL STERLING:

Let's start by asking you to sketch a basic model of the key factors that are fueling AI progress.

TAMAY BESIROGLU:

I think a "compute-based" view of AI progress has been particularly insightful. This is the view that supposes that the increase in the compute that is used to train and run these systems has been central to the improvements in the capabilities of AI systems. Whereas early large language models (LLMs) five years ago used hundreds of graphics processing units (GPUs) in training, models today, such as GPT-4, are scaled up and trained on tens of thousands of GPUs. Time and time again, scaled-up models that are trained on more data have seemed to outperform smaller models in the depth and breadth of capabilities.

This result, that scaled up models outperform smaller ones, is in part formalized in so-called "scaling laws," that describe the extremely regular and predictable relationship between some measures of performance and the amount of compute that is used to train these models.

The amount of compute used to train models has accelerated since the 2010s and is now growing at an extremely fast rate of about 4x per year (**Figure 2**). To put this pace of the growth in AI training compute into perspective, it outpaces even some of the fastest technological expansions in recent history. It surpasses the peak growth rates of mobile phone adoption (2x/year, 1980 – 1987), solar energy capacity installation (1.5x/year, 2001 – 2010), and human genome sequencing (3.3x/year, 2008 – 2015).

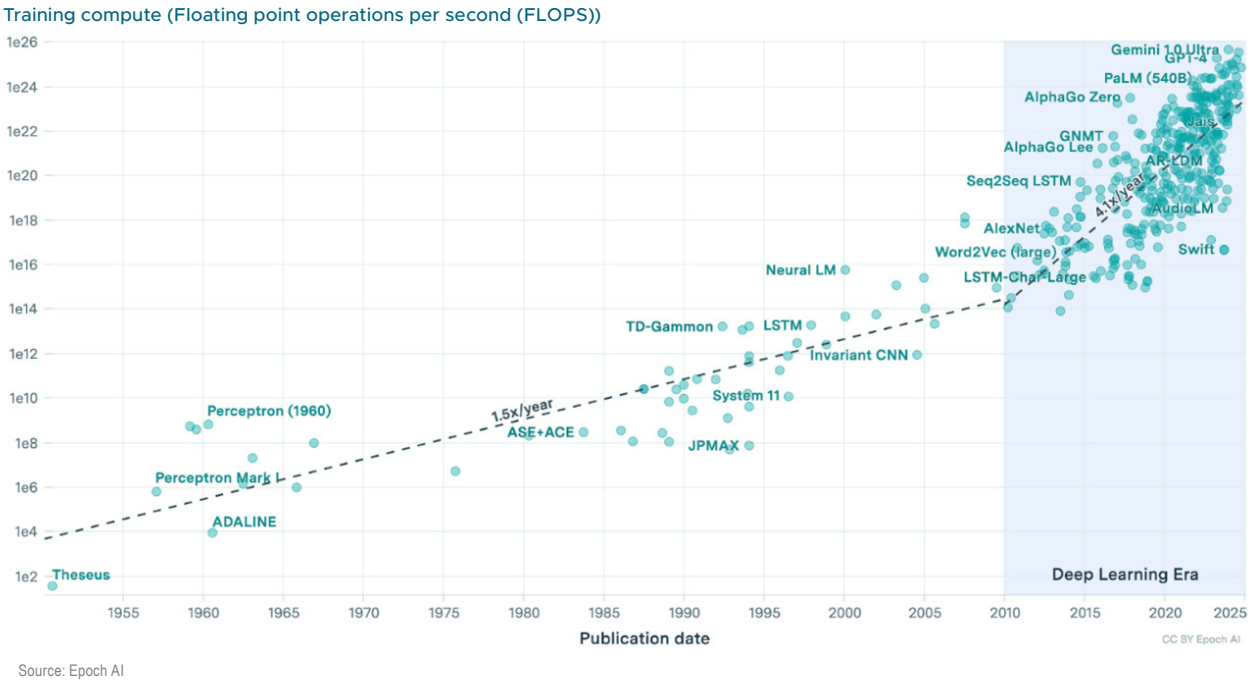
This growth is predominantly the result of greater investment by so-called "hyperscalers" — companies such as Google, Microsoft, and Meta — buying more data center GPUs and dedicating an increasing number of these to training and serving the latest AI models. This has mostly been a story of buying more chips and getting Taiwan Semiconductor Manufacturing Company Limited (TSMC) to produce more chips. In part, it is due to innovations in chip design by Nvidia, improvements in specific components such as memory and interconnect technologies, and improvements in lithography by companies like ASML Holding N.V. (ASML).

Increases in inference compute matter too. Recently, OpenAI has previewed their o1 models, which involves giving the model more

1 Allison Nathan, "GenAI: Too Much Spend, Too Little Benefit?", *Top of Mind*, Goldman Sachs Global Macro Research, Issue 129, June 25, 2024.

FIGURE 2

Notable AI Models



Source: Epoch AI

At a recent growth rate of 4x per year, growth in AI training compute outpaces some of the fastest technological expansions in recent history.

compute to increase its "thinking time." This seems to drastically improve performance on tasks that involve long chains of complex reasoning.

BILL STERLING:

What are some of the key bottlenecks that are most likely to hamper AI progress in coming years? Limitations of training data? GPU shortages? Energy limitations?

TAMAY BESIROGLU:

The provision of power and AI chips could slow down the scaling of AI. Power could mostly bottleneck the scaling of training, whereas AI chips could limit both the scale of training and the serving of AI models ("inference").

AI chips are needed to train AI models and run these models to serve models to customers. We've recently seen AI training clusters that are as large as 100,000 H100s, Nvidia's state-of-the-art data center GPU. However, Nvidia is producing only around two million per year, and these need to be divided between around five or more hyperscalers. Current production is not sufficient to permit AI labs to scale their training runs for more than a year or two.

However, one key issue that results in the supply of these GPUs undershooting where labs would want this to be is that there are time lags between launching a great AI model and seeing chip production expand substantially. First, it usually takes around a year from training a new AI model to launching that model. AI labs can therefore not immediately convince their backers to let them place much-expanded orders with Nvidia for new chips. And once they do, TSMC needs to expand key production capacity, such as advanced packaging, and building new fabrication facilities for this takes around two years from construction to production. This creates a challenging cycle for scaling AI chip production, slowing down the overall scaling of both AI chips and the scaling of AI.

This can be short-circuited. If a company like Google had enough conviction, they could pay TSMC via something like advanced-market commitments, promising them to buy large amounts of future chips at a high price, which could be sufficiently compelling to TSMC to expand its production well ahead of where it otherwise would be. There have been rumors of OpenAI offering something like this, but they currently don't have chip designs as mature as Google.

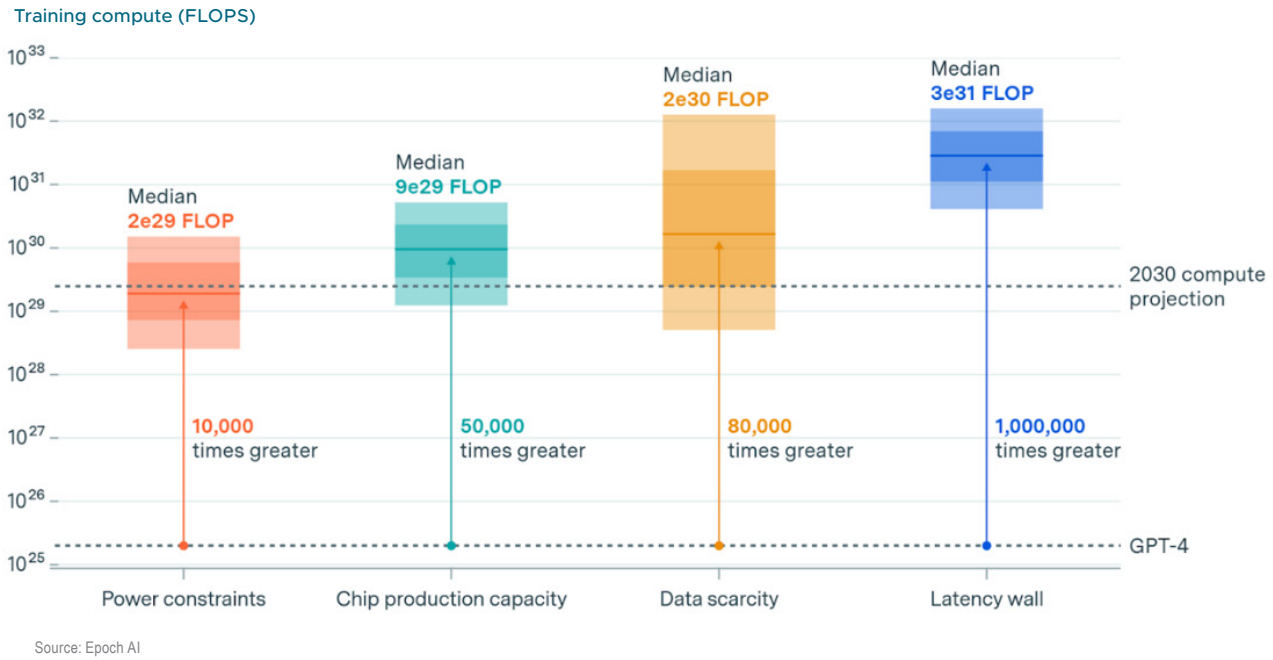
Power availability is emerging as an important bottleneck for AI training (**Figure 3**). Unlike model serving, which can be distributed across multiple data centers, it is much better for training to be concentrated in a single location. This is due to the need for frequent, synchronized updates across all computing units, which requires high-bandwidth, low-latency communication. Consequently, AI training creates intense, localized power demands that might be hard to meet.

When might energy constraints begin to limit AI model training? The recent Llama 3.1 405B model required 27 megawatts (MW) of power during training, equivalent to the average yearly consumption of 23,000 US households. If we assume model energy demands continue to grow rapidly, potentially doubling yearly, power requirements could reach 1 gigawatt (GW) in about five years. This approaches the upper limit of what current on-site power generation can likely supply data centers, potentially creating a bottleneck for AI development.

There are technical solutions and policy solutions. On the technical side, geographically distributed training could tap into multiple regions' energy infrastructure to scale further. This would require large-scale investments in connecting data centers together via a country-wide network of fiber optic cables. On the policy side, the US government could make it much easier to expand energy production and connect new power plants to the grid.

FIGURE 3

Constraints to Scaling Training Runs by 2030



Source: Epoch AI

Four key factors might limit the rapid pace (4x per year) of scaling up AI models: power availability, chip manufacturing capacity, data scarcity, and the "latency wall," a fundamental computational speed limit.

Power availability is the most immediate constraint being addressed by the AI leaders but is not expected to prevent AI models from making a large leap in capabilities by 2030.

BILL STERLING:

How do you see AI investment spending evolving over the next few years?

TAMAY BESIROGLU:

I think a lot of the spending will be on AI hardware and supporting infrastructure, in the form of data center GPUs, power purchase agreements, and investments in new power generation facilities, such as solar farms and possibly natural gas plants.

In terms of data centers, the largest expenditures today are on the order of a few billion dollars for individual AI clusters. Microsoft and OpenAI are rumored to be planning a cluster estimated to be in the hundred-billion-dollar range. I expect within two to four years we will probably see something like $100 billion expenditures from AI clusters from single hyperscalers.

There's a possibility that spending on compute-related capital — including semiconductors, data centers, and dedicated energy infrastructure — will significantly exceed current estimates. Industry leaders like Sam Altman and Satya Nadella have suggested that scaling current AI techniques by approximately 1,000 times could potentially automate a substantial portion of human labor. If this perspective proves accurate, it might justify investments far beyond even the rumored $100 billion data centers.

One intuition of this is just to consider that the global wage bill currently stands at approximately $60 trillion per year. Thus, developing an artificial intelligence system capable of capturing even 10% of this labor value over a few years could potentially yield returns in the tens of trillions of dollars. Admittedly, this would require widespread automation, given that capturing 10% of the value of labor will require the automation of much more than 10% of tasks humans can do. However, this suggests if there is a good chance of developing AI systems that can flexibly substitute for human labor, investing in compute-related capital on the scale of a trillion dollars could potentially offer a favorable return on investment.

BILL STERLING:

A key source of uncertainty about how the economics of the AI revolution will unfold regards how rapidly current advances in AI will culminate in Artificial General Intelligence (AGI), defined as the ability of AI systems to perform all tasks that humans can perform. We read with interest that Geoffrey Hinton, one of the godfathers of AI research, estimated last year that AGI may be reached within 5 to 20 years. How would you assess that view?

TAMAY BESIROGLU:

While it is unclear what definition of "AGI" Hinton has in mind, I agree that this is the right time frame for expecting the capabilities to be developed that would have substantial effects on the rate of economic and technological change.

In the next few years to a decade, I think it's plausible that AI will reach human-level abilities in enough domains to enable something you might call "drop-in remote workers." These AI systems would effectively function as human employees in remote work environments. They could be onboarded like new hires, utilize company software and communication tools, and autonomously complete complex projects over long periods. I expect these to be sufficient to automate enough cognitive work to noticeably boost economic growth.

We've consistently seen AI models improve in both scope and ability when we increase their training compute and data. The substantial performance gap between GPT-2 and GPT-4 can largely be attributed to the 10,000-fold increase in compute used for training and the 1,000-fold increase in training data. Studies on scaling laws and performance tests show that boosting both training compute and inference compute (which gives AI more "thinking time") has been key to expanding what these models can do.

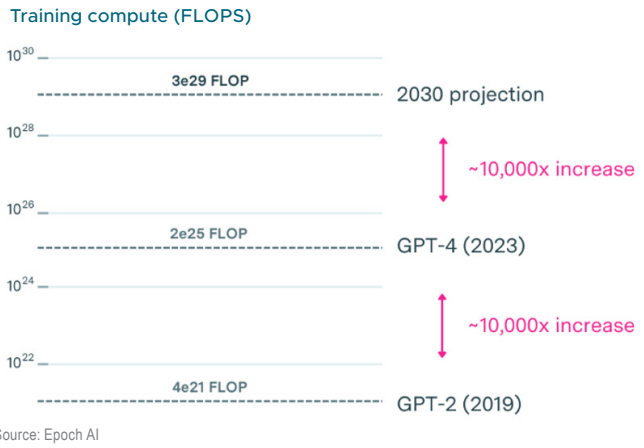
The strong connection between more resources and better AI suggests a clear route to much more powerful AI systems. I expect this will be vigorously pursued, with hundreds of billions of dollars being spent on expanding the resources for training and running these models. My **estimates suggest**[2] that over the next five years, this will likely be sufficient for scaling to models that exceed GPT-4 in scale to the same degree that GPT-4 exceeds GPT-2 in scale (**Figure 4**). Once these advanced models are developed, I expect that complementary software will emerge, enabling AI systems to integrate seamlessly into work processes and autonomously complete projects, much like human remote workers. This would likely boost economic growth and technological change at least by a few percentage points.

2 James Sevilla, Tamay Besiroglu, et al, "Can AI Scaling Continue Through 2030?", Epoch AI Blog, August 20, 2024, https://epochai.org/blog/can-ai-scaling-continue-through-2030.

To achieve a much greater degree of automation and much more accelerated economic growth and technological change, we would need to go several steps further. This would likely require AI automation of AI R&D and accelerating AI capabilities this way. Robotics technology would need to advance significantly, both in capability and scale of production. Investment in AI-related capital — such as data centers, semiconductor fabrication plants, and energy infrastructure — might need to reach double-digit percentages of US output. While these developments seem possible within the next few decades, the timeline could be longer.

FIGURE 4

A Leap as Large as from GPT-2 to GPt-4 Is on Trend by 2030



Source: Epoch AI

Research by Epoch AI suggests that by the end of the decade, we might see advances in AI as drastic as the difference between the rudimentary text generation of GPT-2 in 2019 and the sophisticated problem-solving abilities of GPT-4 in 2023.

BILL STERLING:

Your research has examined carefully a number of arguments in favor of the scenario for substantial AI automation to accelerate global economic growth by an order of magnitude, akin to the economic effects of the Industrial Revolution. How would you summarize the key arguments in favor of transformative economic growth occurring in the not-too-distant future?

TAMAY BESIROGLU:

The central argument for why AI results in much accelerated economic growth is that AI enables us to invest resources to increase the number of digital workers. This transforms labor, a crucial production input that traditionally couldn't be expanded through investment, into a resource that can be accumulated by investing more. In other words, AI allows us to "create" new workers by allocating more resources, similar to how we can increase other forms of capital. This capability, combined with the concept of increasing returns to scale in the economy — where doubling all key inputs (labor and capital) leads to more than doubling of output — suggests that AI could drive accelerating economic growth.

The basic idea is this: Eventually, when we have AI systems that can flexibly substitute for human workers, we can automate a wide range of tasks. As the economy grows, we can reinvest the additional resources into building more physical capital — such as factories, machines, and tools — and into creating more AI workers. Some of these AI workers will focus on improving production efficiency by inventing new ideas that enable us to produce more valuable output per unit of input. This means that whenever we double our inputs, we not only double all our key inputs but also enhance our production efficiency. By doing both simultaneously, we achieve more than double the total output, which can then be reinvested again. This results in accelerated growth, at least until we encounter a significant constraint that substantially hinders further growth.

BILL STERLING:

What are the key arguments against transformative economic growth scenarios? Do you think mainstream economists are too pessimistic about prospects for AI to boost worker productivity across a broad swath of business sectors?

TAMAY BESIROGLU:

There are many arguments for why we might not see economic growth or technological change substantially exceeding current rates. Economists generally disagree on which of these arguments are most compelling, but they tend to agree that some of them are indeed persuasive.

The argument I personally find relatively more convincing is the argument that regulation could curtail or slow the development or deployment of AI. The high compute requirements for AI make it easier to regulate for several reasons. Firstly, the supply chain for compute is highly concentrated, with the majority of the world's AI chips produced by a single company, TSMC, which relies on extreme ultraviolet (EUV) lithography machines manufactured only by ASML. Additionally, AI data centers are large and easily detectable, even through satellite imagery, making it difficult for companies to conceal their AI operations from regulators. Finally, unlike intangible elements such as software, compute resources are tangible and can be precisely measured and quantified, making it feasible to impose regulations that are hard to bypass. Overall, these factors make regulating AI relatively straightforward, potentially hindering AI's advancement.

Another argument is just that historically, while the development of farming or the industrial revolution accelerated economic growth, this acceleration did not continue without bound. Instead, each time growth seemed to have accelerated in the past, we quickly encountered some bottleneck that resulted in growth rates stabilizing.

For example, the Industrial Revolution significantly increased growth rates compared to the agricultural era. However, this growth eventually plateaued as population growth decoupled from economic output once societies escaped Malthusian conditions.

Similarly, while AI development might overcome current growth limitations, we should consider the possibility that this new acceleration will also encounter unexpected constraints. As a result, growth rates might stabilize at levels that, while higher than today's, are not dramatically so. This historical pattern suggests we should be cautious about predicting unbounded acceleration from AI, and instead consider the likelihood of encountering new, unforeseen bottlenecks.

BILL STERLING:

What financial market indicators would you watch to assess whether investors are taking seriously the possibility of transformative economic growth?

TAMAY BESIROGLU:

I pay attention to the market valuations of key companies in the AI chip production chain, with a particular focus on Nvidia and, to a lesser extent, TSMC and ASML. These firms are significant because of their market power and their crucial roles in AI chip production, and the share of AI-related revenue (which is particularly high for Nvidia but much less so for TSMC and ASML). Other public companies that seem important are Microsoft, Google, and Meta, although none of these are that informative given that each of their lion's share of revenue is non-AI related.

Out-of-the-money call options can provide insights into the chances of a company doing extremely well, and perhaps offer some insight into the likelihood of breakthrough AI advancements that could dramatically increase a company's value. However, these instruments are often thinly traded, and it is challenging to disentangle these signals from general volatility effects.

Other indicators that are useful are the valuations of private AI companies, like OpenAI and Anthropic. These valuations are particularly informative because these companies are almost completely focused on developing much more advanced AI systems than exist today, and valuations therefore directly reflect investor expectations about the potential for much more advanced AI. Other non-monetary investments in these companies, in the form of compute or energy (see for example Microsoft's **rumored** $100 billion investment in clusters for OpenAI)[3] are also informative about leadership priorities at key hyperscalers.

BILL STERLING:

There has been a lot of media attention on the risks of "unaligned" AI and the need for stringent regulation to prevent unintended consequences from an "intelligence explosion." Are these worries overdone? And what key principles should guide regulation of AI systems development and deployment?

TAMAY BESIROGLU:

My impression is that the people that engage with this topic often have extreme views relative to me, either dismissing entirely or assigning some high probability to extinction from AI misalignment. I assign quite a lower chance to serious catastrophe from misalignment.

There are several reasons I think that misalignment does not spell doom. Humans too are often imperfectly aligned with respect to each other, but generally find that the best way to accrue resources and influence is to generate value for others by engaging in positive-sum economic activities. AIs too could be self-interested,

3 "Microsoft and OpenAI Plot $100 Billion Stargate AI," Anissa Gardizy and Amir Efrati, *The Information*, March 29, 2024, https://www.theinformation.com/articles/microsoft-and-openai-plot-100-billion-stargate-ai-supercomputer.

but find that the best way to pursue these goals would be to engage in a positive way with the rest of the world.

AIs would likely recognize the value in creating and supporting institutions that enable coordination and trade among various entities, including humans. This is because trade tends to be more efficient and mutually beneficial than conflict or violence. In general, peaceful resolutions through negotiation are preferable to violence when they can be achieved. With advanced AI systems potentially assisting in negotiations, communication, and the establishment of binding agreements, I believe peaceful settlements in the case of misalignment seems quite probable. I do think that humanity's influence in shaping far places in the distant future will dwindle in the fullness of time. Some might consider this catastrophic, but I personally don't.

BILL STERLING:

Thank you, Tamay!

William P. Sterling, Ph.D.
Global Strategist

RELATED READING

James Pethokoukis, "5 Quick Questions for…MIT Research Scientist Tamay Besiroglu on the Huge Economic Potential of AI," American Enterprise Institute, August 2, 2022.

Jakub Kraus, "#8: Tamay Besiroglu on the Trends Driving Past and Future AI Progress," Center for AI Policy, June 14, 2024.

Matt Clancy and Tamay Besiroglu, "The Great Inflection? A Debate About AI and Explosive Growth," *Asterisk Magazine*, June 2023.

Ege Erdil and Tamay Besiroglu, "Explosive Growth from AI Automation: A Review of the Arguments," arXiv:2309.11690v3 [econ.GN], July 15, 2024.

Leopold Aschenbrenner, "Situational Awareness: The Decade Ahead," situational-awareness.ai, June 2024.

Anton Korinek, "Scenario Planning for an A(G)I Future," *Finance and Development Magazine*, International Monetary Fund, December 2023.