The Henry Fund
Henry B. Tippie College of Business
Landon Skinner [landon-skinner@uiowa.edu]

Artificial Intelligence

April 21, 2023

Information Technology – Problem Solving with Computers

Stock Rating **OVERWEIGHT**

Investment Thesis

We are OVERWEIGHT on the artificial intelligence industry, driven by its transformational role in global productivity and infrastructure transformation. As AI adoption accelerates across sectors and countries, we see sustained demand for compute, software, and AI-native services. With a multi-year investment cycle underway, falling unit costs, and real commercial use cases emerging, we are bullish on artificial intelligence.

Drivers of Thesis

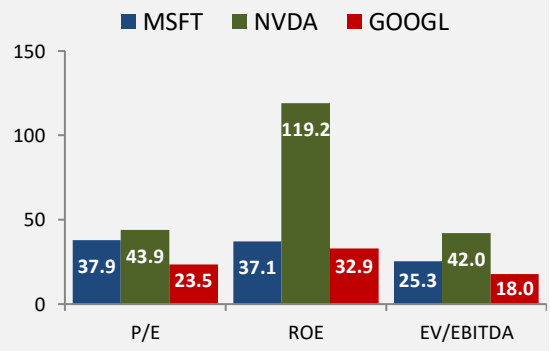
- **Massive CapEx Commitments:** Over $300B in announced AI infrastructure investment from hyperscalers and sovereign initiatives supports multi-year demand.
- **Industry-Wide Adoption:** AI integration is expanding in healthcare, finance, defense, and manufacturing, driving new monetization pathways.
- **Inference and Agentic AI Growth:** As model deployment outpaces training, inference workloads are expected to scale across enterprise applications as models improve, allowing for agentic AI.

Risks to Thesis

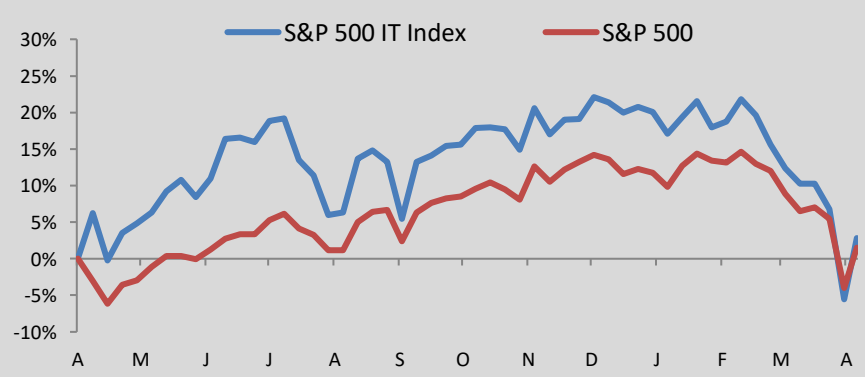
- **Legal and IP Exposure:** Training data ownership challenges could raise costs or lead to restrictions.
- **Overbuild Risk:** Infrastructure investment may outpace near-term monetization, compressing returns and stagnating investment.
- **Regulatory Uncertainty:** Global AI laws are evolving and may introduce compliance burdens or restrict data access.

Key Metrics

Market Cap	($ in millions)
MSFT	$2,734,070
GOOGL	$1,854,734
AMZN	$1,831,800
META	$1,270,579
NVDA	$2,476,356
AMD	$142,155
EV/EBITDA	
MSFT	25.34
GOOGL	17.97
AMZN	19.15
META	17.66
NVDA	41.95
AMD	39.26
P/E	
MSFT	37.88
GOOGL	23.68
AMZN	39.70
META	24.54
NVDA	43.90
AMD	122.15



12 Month Performance



Industry Description

The artificial intelligence industry develops systems that replicate human cognitive functions, including language understanding, image recognition, decision-making, and automation. These technologies are increasingly embedded across sectors such as healthcare, finance, manufacturing, and defense, with adoption accelerating globally as organizations invest in data, compute infrastructure, and specialized talent.

INDUSTRY DESCRIPTION

Intelligence

Intelligence is the ability to use logic and reason to interpret external stimuli and demonstrate one's agency to act and change the world and environment around them. Using our five senses along with our brains, which are the largest relative to body size, humans have applied intelligence to solve problems and make decisions to modify the environment. Intelligence, long thought to be tied to consciousness, was questioned by Alan Turing when he drew a distinction between consciousness and intelligence.[1]

Artificial Intelligence

Artificial intelligence is the ability of machines to use reasoning to make logical conclusions. Once thought to be science fiction, AI has gained prominence in recent years due to major breakthroughs. Most recently, artificial intelligence has been able to pass the Turing Test, an assessment of whether people can distinguish between humans and machines through text-based conversations.[1] Passing this test supports the idea that machines are capable of learning.

Machine perception is fundamental to the growth of AI, enabling computers to interpret the world by mimicking human senses. By processing sensory data and applying reasoning, computers have made significant progress in their ability to interpret information and generate conclusions.

The benchmark tests for true artificial intelligence have become more rigorous. For example, the Turing Test evaluates a computer's intelligence by assessing whether a machine can be perceived as human. Simply put, can a person tell the difference between a computer and a human?

Artificial intelligence has been commonplace in many industries for decades, often working behind the scenes. AI has powered Google's search engine since 2001, helping to correct spelling and predict searches.[2] It drives personalized recommendations on Amazon by analyzing purchase behavior and browsing patterns.[3] Meta has long used AI in its algorithms to curate personalized feeds on its social media platforms.[4]

Foundations of Artificial Intelligence

Artificial intelligence researchers set out to replicate human thought processes using machines. Originally coined at Dartmouth in 1956,[5] artificial intelligence initially pursued a symbolic or traditional approach, based on the belief that intelligence could be explicitly coded into machines. By encoding human knowledge, reasoning, and logical rules to create algorithms, or sets of computer instructions, researchers limited a machine's expertise to the domains they directly programmed.

John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude Shannon, the pioneers of artificial intelligence, employed this approach in the field's early development. This traditional method, often referred to as "Good Old-Fashioned AI" (GOFAI),[6] reflected the rationalist tradition of Western philosophy, where reason is considered the primary source of knowledge.[7]

The symbolic era of AI research aligned with the dominant academic paradigms of the time, influenced by figures ranging from mathematicians like Claude Shannon to linguists like Noam Chomsky. Chomsky's cognitive theory, which argued that language is biologically innate,[8] had a significant impact on the early cognitive framework guiding artificial intelligence.

From AI Cold Winter to an AI Boom

GOFAI was limited and struggled with the complexities and intricacies of the real world. By the beginning of the 1980s, the limitations of GOFAI were stalling the field's advancement. Hard-coding algorithms to simulate human reasoning proved infeasible, and funding for artificial intelligence research began to decline. As support diminished, a resurgence in connectionism emerged, using mathematical model networks to simulate the human brain.

Academics like Geoffrey Hinton, now often referred to as the "Godfather of AI," pioneered connectionist learning models.[9] These networks of neuron-like processing units allowed machines to learn from data and solve problems that hand-crafted algorithms could not. Hinton believed that intelligence could emerge from pattern recognition in data.[10] His view followed an empiricist philosophy that knowledge originates from sensory experience, which conflicted with the dominant belief that truth exists independently of experience.[7]

By challenging the prevailing academic thinking, Hinton's idea that computers could think like humans using intuition from data rather than rules remained outside the mainstream. As a result, neural network research received limited funding, fewer research opportunities, and less academic visibility. It survived in small, dedicated academic communities maintained by enthusiasts working at the edges of the field.[11]

In 1990, Yann LeCun, one of Hinton's students, demonstrated at Bell Labs that neural networks could read handwritten numbers.[12] Since 2013, LeCun has served as Meta's Vice President and Chief AI Scientist.[13] Geoffrey Hinton also joined Google in 2013. Interest in neural networks among major technology firms increased after Hinton and his students, Alex Krizhevsky and Ilya Sutskever, competed in the ImageNet visual recognition contest, which involved classifying over a million images. Using a neural network called AlexNet, their model learned to recognize images without being explicitly programmed.[15]

Krizhevsky went on to work on Google's autonomous driving projects, and Sutskever later co-founded OpenAI. He now leads a startup called Safe Superintelligence.[16] AlexNet, which was trained using two NVIDIA graphics processing units, is widely seen as the starting point of the current AI boom.[15]

Artificial intelligence began as a rule-based system with ambitious goals, but symbolic AI reached its limits when applied to real-world complexity. Learning-based techniques, once considered fringe, finally gained recognition after the success of ImageNet in 2012. Neural networks and machine learning, supported by advances in GPU hardware, launched the present era of artificial intelligence as former skeptics adopted data-driven methods.

Machine Learning (ML)

Most of today's breakthroughs stem from advancements in machine learning, a branch of AI that focuses on algorithms capable of learning patterns from data. Machine learning algorithms allow machines to make predictions or decisions without being explicitly programmed. These models improve as they are trained on larger datasets and process more information. The greater the volume of data, the more advanced and accurate the models become. Machine learning models adjust their internal parameters to improve performance on tasks such as prediction or classification by using statistical techniques to minimize errors on training sets, which allows them to generalize to new data. Machine learning has proven especially effective for pattern recognition and prediction.[17]

Deep Learning (DL)

Deep learning is a subset of machine learning that uses neural networks to model complex patterns in data. By using neural networks with multiple layers, deep learning can gradually extract higher-level features from raw, unlabeled data. It excels at recognizing patterns in large, unstructured datasets such as images, audio, and text. Deep learning models improve as they are exposed to more data and supported by greater computational power. With sufficient data, computing resources, and time, machine learning can recognize intricate relationships and subtle nuances across various data formats.[18]

Neural Networks (NN)

Inspired by the human brain, neural networks are composed of interconnected nodes and artificial neurons that process information in hierarchical stages, increasing in complexity as data moves through each layer. Neural networks are a central architecture in modern machine learning. Through a training process called backpropagation, they learn complex relationships within data.[19]

Convolutional Neural Networks (CNNs) specialize in processing visual or spatial data. Using convolutional layers, these networks scan inputs and apply filters to detect low-level features, which are then combined with higher-level features in deeper layers. This structure mimics how the human visual cortex processes signals.[20]

Recurrent Neural Networks (RNNs) are designed for sequential data such as language, speech, or time series. They include loops that allow information to persist across nodes, enabling the model to retain context from previous inputs. RNNs are primarily used for language modeling.[20]

Large Language Models (LLM)

Large Language Models (LLMs) are a subcategory of AI that specialize in understanding and generating human-like

text. LLMs use advanced techniques to translate binary computing into outputs that resemble natural language. They are built on transformer architecture, which processes input in parallel and uses attention mechanisms to weigh the relevance of different words in predicting coherent sentences.[20]

LLMs are trained on massive amounts of text from sources such as websites, blogs, books, and articles, allowing them to learn probabilistic representations of language. When prompted, an LLM predicts the most likely next word in a sequence to generate a complete response. Because LLMs rely on probability to determine their outputs, performance improves as the size of the training dataset increases.

LLMs highlight both the capabilities and limitations of artificial intelligence. For instance, they are prone to hallucinations, where false information is presented as fact. At the same time, researchers have observed emergent behaviors where LLMs perform tasks they were not explicitly trained or instructed to do.

The transition from binary outputs to natural language generation has been enabled by the sharp decline in data storage and processing costs. This shift allows LLMs and other AI systems to be trained on large-scale datasets using powerful data centers capable of processing and understanding complex text.[21]

Big Data

Machine learning relies on massive datasets; the more examples an AI system is exposed to, the better it can learn and the more accurate its predictions become. Over the past two decades, the world has seen an explosion of digital information, driven largely by the internet. Writing shared online, sensor data from smartphones, digitized records, uploaded images, and other digital sources have all become valuable inputs for training AI systems. In the last decade, researchers have increasingly leveraged this abundance of data to improve model performance.[22]

The value of data to AI depends on the models' ability to access it. Open-data advocates like Aaron Swartz, co-founder of Reddit, challenged institutional barriers to information and helped democratize access to data. Swartz's legacy, particularly following his death after being prosecuted for downloading millions of JSTOR articles to share freely, became symbolic of the movement toward open, empirical AI that learns from the world around it.[23]

In the era of Big Data, data lakes have replaced traditional databases as the preferred method for storing large-scale information. Unlike structured databases that organize data into fixed tables for efficient processing, data lakes allow organizations to store raw data in its native format. By collecting and preserving all potentially useful information in a centralized repository, organizations gain the flexibility to process and analyze that data later, as needed.[24]

Cloud Computing

To train and run AI models, there is a need not only for data but also for the infrastructure to store and process it. In the past, only large federally funded universities or corporations had the resources to manage massive datasets and maintain computing clusters. Today, with the development of internet infrastructure, a global interconnected network of computers, data can be transmitted across cables to remote servers.

Cloud computing enables access to storage, processing power, and software applications over the internet. These services can be used remotely, eliminating the need for physical cables and on-site servers. The rise of cloud computing has significantly improved efficiency, scalability, and connectivity by allowing data to flow seamlessly across the globe to connected devices.[25]

Data Centers

Information Technology (IT) infrastructure refers to the computer hardware, software, and resources that support organizational operations. IT infrastructure enables business processes and increases productivity across industries.

A major component of IT infrastructure is the data center. Traditionally, data centers were rooms located on-site that housed a company's IT infrastructure. Today, they have evolved into million-square-foot facilities containing thousands of servers that can be accessed remotely. This shift from in-house infrastructure to outsourced services coincided with the rise of cloud computing. Cloud technology allows organizations to access computing

resources over the internet, removing the need for physical cables and local servers

The widespread adoption of cloud computing has transformed the way IT is managed. Rather than maintaining in-house IT teams, companies increasingly outsource these responsibilities to experts and specialists. Many organizations now partner with colocation facilities, which own and operate the infrastructure and rent out space, servers, and bandwidth to clients. This arrangement gives companies access to modern data center capabilities without the burden of owning the physical infrastructure.

To access data center resources, companies subscribe to service models based on their specific needs. These models include Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). Each offers varying levels of virtual access to storage, networking, and compute resources. Software-defined infrastructure (SDI) further enhances flexibility by allowing dynamic allocation of resources to meet changing workloads.

Modern data centers have significantly advanced beyond traditional IT setups. With cloud architecture, physical components such as CPUs, GPUs, storage systems, and networking equipment can be accessed virtually, avoiding large upfront capital expenditures. By leveraging shared access to routers, switches, and firewalls, users can optimize utilization, scale operations rapidly, increase processing efficiency, and reduce provisioning time.

Data centers that exceed 10,000 square feet and contain more than 5,000 servers are classified as hyperscale data centers, or hyperscalers. Google launched the first hyperscale data center in 2006, spanning 1.3 million square feet. Most hyperscalers consume between 100 and 300 megawatts of power, reflecting the immense scale and capability of these facilities.[26]

Revolutionized Processing

The Central Processing Unit (CPU) performs linear computation using a cache hierarchy with three levels: L1 being the fastest and L3 being the slowest. Data movement between caches is managed by the memory management unit (MMU), which directs information between the processor core, cache, and Random Access Memory (RAM).

Graphics Processing Units (GPUs) support CPUs by offloading simpler, parallelizable tasks such as processing large datasets and rendering graphics. This allows the CPU to focus on tasks that require sequential execution and are less suited to parallel processing.

GPUs have transformed data processing by enabling parallel computation. Unlike CPUs, which typically handle a few complex tasks at a time, GPUs use thousands of small, efficient cores to perform many simple calculations simultaneously. This makes them ideal for executing the same operation across large volumes of data. Within GPUs, Streaming Multiprocessors (SMs) serve as the primary processing units. While simpler and less powerful than CPU cores, SMs excel at running thousands of threads in parallel rather than relying on context switching.[27]

Combining CPUs and GPUs has significantly reduced the time required to solve problems involving linear computations. This hybrid architecture has accelerated processing at scale and enabled rapid advancements in artificial intelligence, making it possible to train complex models on massive datasets housed in modern data centers.

The Evolution of Moore's Law

Gordon Moore, who co-founded Intel with Robert Noyce and was known as the "Mayor of Silicon Valley," made a prediction in the early days of integrated chips that would shape the semiconductor industry. Moore stated that the number of transistors on integrated circuits would double every two years. Although he never intended this to become a permanent rule, Moore's observation has held true for more than 50 years. This consistent doubling of transistors has allowed computing devices to become smaller and up to 100 times faster every decade.[28]

NVIDIA CEO and founder Jensen Huang has declared that Moore's Law is now at its limits, as transistors are being produced on atomic scales as small as three nanometers wide. Huang introduced the idea of "hyper Moore's Law," noting that GPUs are lowering the cost of inference and are now 30 times faster than they were five years ago. In 2024, Huang stated that NVIDIA's AI chips are 1,000 times better than those from a decade ago, which exceeds the pace set by Moore's Law.[29]

These advancements have been made possible not only by increasing the number of transistors, but also through

simultaneous improvements in chip architecture, libraries, and algorithms. Moore's Law has set the standard for progress in the semiconductor industry over the past half-century, and ongoing innovation continues to drive down computing costs, reduce device size, and increase processing power. By the early 2010s, just twelve NVIDIA GPUs could deliver the same deep learning performance as 2,000 CPUs.[15]

Graphic Processing Unit

The GPU is designed for parallel computing, enabling complex computations to be broken down into smaller subtasks. With thousands of cores, each contributing computational power, GPUs can manage thousands of instructions at the same time. This significantly reduces execution time for both data processing and visual rendering. By using GPUs, which are highly effective at parallel operations and capable of performing many multiplications simultaneously, the time needed for machine learning tasks has been greatly reduced.

Matrix multiplication is a mathematical operation in which the rows of one matrix are multiplied by the columns of another, and the results are summed to produce a new matrix. This operation is central to training neural networks. Because the training process relies heavily on repeated matrix multiplications, GPUs have become essential for accelerating artificial intelligence.[30]

Shovels of AI

GPUs have become the foundation of artificial intelligence. Machine learning was once slow, with computation-heavy tasks taking too long to complete. NVIDIA invented the GPU to accelerate computing, originally focusing on graphics processing. By enhancing image generation on monitors, NVIDIA secured a strong foothold in the gaming industry. The company entered gaming due to the demand for parallel processing in 3D graphics, which requires complex computational power. With a passion for virtual worlds and a belief in gaming's potential as a dominant form of entertainment, NVIDIA introduced its GPUs to personal computers and later to gaming consoles. Notably, NVIDIA's GPU powered the first Microsoft Xbox.

Through continuous innovation in both hardware and software, NVIDIA has positioned itself as a cornerstone of the AI revolution. Rather than simply keeping pace with emerging technology, the company has helped shape one of the most transformative shifts of the modern era. NVIDIA's core strategy focuses on technological leadership and platform development across AI, graphics, computing, and autonomous systems. The company enhances performance by tightly integrating its software and hardware, as demonstrated by the 2006 launch of CUDA, a parallel computing platform and programming model.

Today, NVIDIA provides AI solutions such as DGX Cloud, an AI training-as-a-service platform, and NVIDIA DGX Systems, a supercomputing solution optimized for deep learning workloads. Beyond hardware, NVIDIA supports innovation through initiatives like the Inception Program, which accelerates AI startups, and the Deep Learning Institute, which trains developers in neural network techniques. These efforts not only advance the broader AI ecosystem but also attract new companies and talent into NVIDIA's orbit, reinforcing its role as an industry leader and a critical enabler of the AI gold rush.[31]

ESG Analysis

Environmental, Social, and Governance (ESG) is a framework that helps stakeholders, or individuals with an interest in a company's outcomes, understand the risks and opportunities associated with the business. In the context of artificial intelligence, ESG considerations are especially important due to the high energy demands and environmental impact of the infrastructure that supports it, particularly data centers.

Electricity is the largest operating cost for data centers, as significant power is required for both computing operations and cooling systems. To address sustainability concerns, hyperscalers are increasingly turning to renewable energy sources such as wind, solar, hydropower, and geothermal. However, challenges related to intermittency and grid reliability make it difficult to rely entirely on these sources

To ensure uninterrupted service for AI workloads, data centers must be supported by redundant power solutions, including diesel generators and backup battery storage. As the industry expands, especially in high-density technology hubs like Northern Virginia, Silicon Valley, and the Dallas-Fort Worth area, access to a stable and scalable power grid becomes essential. This need for reliability will only grow as AI is adopted in more regulated sectors such as healthcare and finance.[33]

Despite these challenges, artificial intelligence remains a rapidly growing industry. ESG considerations will continue to shape its future, influencing how companies balance sustainability with operational efficiency. Ongoing monitoring of industry commitments to renewable energy and carbon neutrality will be essential to understanding the long-term environmental impact of AI.

RECENT DEVELOPMENTS

Pushing the Laws of Physics

For over 50 years, the semiconductor industry has followed Moore's Law, which states that the number of transistors on a chip doubles approximately every two years. However, as transistor sizes shrink to the atomic level, currently as small as 2 nanometers and built atom by atom, they are starting to face fundamental physical limitations. Challenges such as quantum tunneling, heat dissipation, and increased leakage make continued miniaturization more difficult and more expensive.

Advanced manufacturing techniques, including those developed by ASML using Extreme Ultraviolet (EUV) Lithography, are now required to produce these tiny components. While these innovations make scaling possible, they also introduce high production costs.[34] The wafer industry is approaching a point where the advantages of reducing transistor size are outweighed by the limits of physics and the rising cost of manufacturing.

Fabrication

In the fabrication of advanced semiconductors, Taiwan Semiconductor Manufacturing Company (TSMC), the world's largest foundry, plays a vital role in producing high-performance chips that are essential for artificial intelligence applications. TSMC operates the most advanced fabrication facilities, currently capable of manufacturing 2-nanometer transistors, and benefits from economies of scale by serving major clients such as NVIDIA, Apple, and AMD.[35] The company is responsible for producing over 90 percent of the world's most advanced semiconductors.[36]

As its name suggests, TSMC is a Taiwanese company, with the majority of its facilities located in Taiwan. It was originally founded as a joint venture between the Taiwanese government and private investors to establish the world's first dedicated semiconductor foundry.[37] Today, TSMC has expanded its global footprint and is actively investing in international growth. The company has committed $165 billion to build three fabrication plants, two advanced packaging facilities, and a research and development center in the United States to strengthen its presence abroad.[38]

DeepSeek and Compute Efficiency

DeepSeek's chain-of-thought approach encourages the model to self-evaluate by thinking out loud, allowing users to identify where logical errors occur. This method reinforces meaning and guides the model in a different way from traditional approaches. Instead of relying entirely on supervised learning, which uses labeled data, or unsupervised learning, which identifies patterns independently, DeepSeek applies reinforcement learning to optimize its policy. This helps reduce training costs, as it eliminates the need to retrain an entirely new model when the policy changes.

Through model distillation, DeepSeek allows a large language model to train a smaller one using reinforcement learning. This reduces the resources required for training and deployment, including GPU consumption. DeepSeek has also open-sourced its code, increasing transparency and accessibility.

A recent white paper from DeepSeek, a Chinese artificial intelligence company, has drawn international attention. The company's app quickly became the most downloaded on the iOS App Store. Most notably, DeepSeek claims to deliver performance comparable to leading Western models while operating at a fraction of the cost. According to the white paper, the model was developed with just six million dollars in funding, far below the hundreds of millions invested by companies such as OpenAI, Google, and Meta.

Due to restrictions on access to advanced semiconductors, DeepSeek was trained using older NVIDIA H800 chips. Despite this limitation, the model achieved comparable performance while operating at a much lower cost—just fourteen cents per million tokens, compared to OpenAI's rate of two dollars and fifty cents. These results raise questions about the necessity of cutting-edge hardware

and massive capital expenditure in achieving top-tier AI performance.[39]

Following the release of the white paper, the Nasdaq fell five percent in pre-market trading, reflecting growing concerns over the future of U.S. tech leadership. NVIDIA alone experienced a decline in market capitalization of nearly six hundred billion dollars, prompting renewed scrutiny of its central role in powering the global AI ecosystem.

Agentic AI

As AI models have grown more powerful, agentic AI has received increased attention. Unlike basic chatbots that only respond when prompted, an AI agent can take initiative, bringing artificial intelligence closer to a more autonomous form. An autonomous agent is capable of acting within an environment without constant human direction. This development opens the door to a wide range of use cases, from automating business workflows to serving as personal digital assistants.

INDUSTRY TRENDS

Data Center Geography

Traditionally, data centers have been located in dry, arid regions where low humidity made air cooling more efficient and less risky for sensitive hardware. Dry climates reduce the risk of corrosion and minimize the need for complex humidity control systems, which historically made these regions cost-effective for air-cooled infrastructure.

With the introduction of high-performance systems such as NVIDIA's Blackwell architecture, the industry is now transitioning toward liquid cooling. This method allows for improved thermal management, greater GPU and network density, and reduced physical space requirements. As liquid cooling becomes more common, the traditional advantages associated with desert-based data centers are becoming less relevant.

Cooler geographic regions, including the northern United States and Canada, are expected to become increasingly attractive for new data center development. These areas offer access to free cooling, which uses naturally cold air or water to reduce reliance on energy-intensive mechanical systems. This shift supports both cost reduction and sustainability goals.

At the same time, the growing demand for low-latency computing is accelerating the deployment of edge data centers. These smaller, distributed facilities are positioned closer to users and can benefit from local climate conditions while handling real-time workloads. As data centers evolve, geography, cooling efficiency, and infrastructure density will play a defining role in shaping future deployment strategies.

Global Competition

Recently, a white paper from a Chinese artificial intelligence company, DeepSeek, gained significant attention. After its release, the company's app became the most downloaded on the iOS App Store. Most notably, DeepSeek claimed to match the performance of leading Western AI models at a fraction of the cost. According to the white paper, the model was developed for only six million dollars, far less than the hundreds of millions invested by companies such as OpenAI, Google, and Meta.

These claims raised concerns about investment efficiency and innovation within Big Tech. Some argue that large U.S. tech firms, bolstered by substantial free cash flow, have become too reliant on spending power to solve problems. This approach may have stifled creativity and innovation. There is growing concern that these companies have moved away from their original culture of agility and experimentation.

OpenAI responded by claiming that DeepSeek was distilled from its own generative AI model, ChatGPT. It is widely accepted that creating a breakthrough product is more expensive and difficult than replicating one. First movers often set industry standards and build brand loyalty. The ability to replicate an existing product at lower cost remains a persistent source of geopolitical tension, especially amid allegations that China steals up to 600 billion dollars in intellectual property from the United States each year.

DeepSeek also claimed that it trained its models using older NVIDIA H800 chips, which are less powerful than newer models restricted by U.S. export controls. Despite using older hardware, the company achieved similar performance and reported inference costs of just fourteen cents per million tokens, compared to two dollars and fifty

cents for OpenAI's GPT. This has led to renewed scrutiny over the value of large-scale investments and the need for high-end compute.[39]

The market reacted quickly. After the white paper was released, the Nasdaq fell five percent in pre-market trading, reflecting investor concerns over U.S. tech competitiveness. NVIDIA alone saw its market capitalization drop by nearly 600 billion dollars. Investors began to question whether NVIDIA could maintain its dominance and premium pricing if older-generation chips proved sufficient for training advanced AI models.[40]

H800 chips offer only a fraction of the performance of NVIDIA's most advanced GPUs and generate lower margins. If companies can achieve competitive results with lower-tier hardware, demand for the most advanced chips may fall. This could erode the pricing power of U.S. chipmakers and weaken the strategic advantage of the American tech industry. Chinese companies, including Huawei, would likely find it easier to reproduce older-generation chips than to compete with the latest frontier technologies.

This situation challenges the assumption that AI progress is driven primarily by compute power. GPUs have enabled the rise of AI by supporting parallel processing for machine learning tasks. However, if older chips continue to deliver high-quality results, buyers of cutting-edge hardware may find that they overpaid. NVIDIA's ability to justify premium pricing could diminish, especially if performance gains do not scale proportionally with cost.

DeepSeek's research team reportedly spun out of a quantitative hedge fund, bringing with them 10,000 A100 GPUs before export restrictions were enacted. Although these chips are powerful, they are still less advanced than the H100 Hopper models that have been the industry standard, and the newer Blackwell generation recently announced. DeepSeek claims its progress comes not from hardware advantages, but from innovations in algorithm design and model efficiency.

Skepticism remains around DeepSeek's reported infrastructure. We believe is unlikely that the company has operated for two years without expanding its compute capacity beyond the initial 10,000 GPUs. Scalable infrastructure is essential for any serious AI operation. Even if the company used only the original 10,000 units, the cost would still exceed 200 million dollars based on conservative pre-restriction pricing estimates.

There is growing evidence that China has been circumventing export controls by channeling restricted chips through countries such as Singapore. Trade data shows that Singapore, once classified under "other," is now tracked independently, with shipments increasing as restrictions on China have tightened.

Emerging Technology Supervision

The United States emerged as a global leader in the digital economy by adopting a light-touch regulatory approach to the early internet. A key policy supporting this growth was Section 230 of the Communications Decency Act, which provided legal protection to platforms for content posted by their users. This allowed internet companies to expand rapidly without being burdened by constant litigation, encouraging innovation and risk-taking. As a result, firms like Google, Facebook, and YouTube were able to scale globally, helping cement the U.S. as the dominant force in the digital era. Today, many within the U.S. policy and tech communities argue that a similar hands-off approach should guide the development of artificial intelligence.

Europe, by contrast, took a more cautious stance in the early days of the internet. Emphasizing user privacy, platform responsibility, and content regulation, the European Union positioned itself as a rule-setter rather than a market leader. While this helped establish strong data protection standards, it also contributed to the absence of European tech giants in core areas like search, social media, and cloud infrastructure. Policymakers have since acknowledged that early overregulation may have hindered the continent's competitiveness. That experience now informs how Europe is approaching artificial intelligence, with growing awareness of the need to support domestic innovation alongside regulatory safeguards.

France has emerged as a prominent voice within Europe, advocating for a more balanced approach to AI governance. President Emmanuel Macron has emphasized the importance of building a competitive AI ecosystem that reflects European values without repeating the mistakes made during the internet's rise. At the 2025 AI Action Summit in Paris, France emphasized the need to promote AI research, startups, and infrastructure while still advancing transparency and accountability. This

marked a noticeable shift from earlier rhetoric focused solely on ethics and restraint. The French position reflects an effort to ensure that Europe not only regulates AI effectively, but also plays a leading role in its development.[41]

This recalibration across Europe highlights a broader strategic objective. As the United States continues to prioritize innovation and China expands its state-backed AI capabilities, Europe is seeking a third path that balances responsible oversight with global competitiveness. Policymakers now recognize that building influence in the AI era requires more than drafting laws—it also means enabling the research, talent, and investment needed to shape the next generation of technology. The internet era demonstrated that those who build the platforms also set the standards. In the race for AI leadership, Europe appears determined not to be left behind.

MARKETS AND COMPETITION

Major Players

As of 2025, the artificial intelligence landscape is led by a concentrated group of firms pushing model performance and commercial adoption. OpenAI remains at the forefront with GPT-4o, consistently topping industry benchmarks in reasoning and language understanding, and setting the standard for proprietary foundation models. Google's Gemini 2.0 Pro Experimental has gained traction, particularly in advanced testing environments, and signals Alphabet's intent to compete directly on model sophistication. Meta continues to invest heavily in open-source, with Llama 3.1 emerging as the leading publicly available model, striking a balance between transparency, accessibility, and performance. Microsoft maintains its strategic position through integration, embedding its Copilot model across enterprise productivity software to drive distribution at scale. DeepSeek, a newer entrant, has disrupted expectations by delivering highly competitive model performance at materially lower cost, forcing incumbents to reassess efficiency strategies and cost structures across training and inference.[42]

Microsoft (MSFT)

Microsoft holds approximately 21 percent of the global cloud infrastructure market through its Azure platform.[43] The company is a key partner of OpenAI and maintains priority access to OpenAI's infrastructure needs through 2030. Microsoft has integrated AI deeply into its product ecosystem, including Copilot, an AI assistant embedded across Microsoft 365 applications. The company plans to invest up to 80 billion dollars in AI data centers to support the rapid expansion of cloud and artificial intelligence services.[45]

Alphabet (GOOGL)

Alphabet operates Google Cloud, which accounts for roughly 12 percent of the global cloud infrastructure market.[43] Its AI model, Gemini, has been integrated into Google Search to enhance user query responses.[46] Alphabet has also partnered with Salesforce to embed Gemini into Agentforce, a customer service and automation platform.[47] The company plans to invest 75 billion dollars into AI and cloud infrastructure to stay competitive with Amazon and Microsoft.

Amazon (AMZN)

Amazon is the largest cloud service provider, commanding approximately 30 percent of the global cloud infrastructure market through Amazon Web Services (AWS).[43] AWS contributes significantly to Amazon's operating profit due to its high margins. Amazon has suggested that capital expenditures on data centers could exceed 100 billion dollars in the coming year, with ongoing expansion constrained by AI chip shortages.[48]

Meta (META)

Meta has developed the Llama family of large language models and currently operates 28 hyperscale data centers powered entirely by renewable energy.[49] The company plans to expand its computing capacity to 1.3 million GPUs by the end of the year. Meta's AI research division is led by Yann LeCun, one of the pioneers of deep learning. Its AI models are used to enhance content recommendation, virtual assistants, and advertising, with plans to make Llama widely available to its global user base of 3.3 billion people.[50]

Apple (APPL)

Apple is taking a conservative approach to AI infrastructure, focusing on integrating AI into its ecosystem of high-margin products. The company recently announced "Apple Intelligence," which enhances devices like the iPhone and Siri with AI capabilities. Apple has

committed $500 billion to U.S. investments, including a new AI server manufacturing facility in Houston, Texas, in partnership with Foxconn. Apple remains the largest customer of TSMC, which produces chips at its Arizona plant.[51]

NVIDIA (NVDA)

NVIDIA remains the dominant supplier of GPUs essential for training and running AI models. Its latest architecture, Blackwell, includes chips like the B200 and Rubin platform, designed for high-performance AI and hyperscale data centers. NVIDIA's GPUs are central to nearly every major AI system in production today, making it a foundational pillar of AI infrastructure.

AMD (AMD)

AMD has emerged as a competitor to NVIDIA, producing both CPUs and GPUs optimized for AI workloads. Its ROCm software stack and 2-nanometer chip technology position AMD to gain market share as demand for high-performance, cost-effective AI compute continues to grow. AMD is also gaining traction in data centers that seek alternatives to NVIDIA's ecosystem.[52]

ECONOMIC OUTLOOK

GDP

As artificial intelligence becomes more deeply integrated across sectors, it is expected to drive a significant wave of global productivity gains. The shift from experimentation to large-scale implementation is enabling more efficient workflows, better decision-making, and the automation of complex tasks. Rather than replacing workers entirely, AI is increasingly used to augment human capabilities, allowing for greater output and efficiency across industries.

As adoption moves from training to real-time inference, the demand for computing power is projected to rise sharply. This evolution will require substantial growth in digital infrastructure, including high-performance data centers, storage systems, and advanced networking. The ability to scale compute resources effectively will become a key enabler of economic growth in the AI era.

AI's impact is likely to reshape labor markets, influence global supply chains, and redefine competitive advantages across national economies. Continued investment in infrastructure, research, and upskilling will be essential to realizing the full economic benefits of this technological transformation. Looking ahead, artificial intelligence has the potential to become one of the most important contributors to long-term productivity and GDP growth.[53]

Defense

The United States entered the modern AI era with clear advantages: leading academic researchers, eager venture capitalists, early investment from major tech firms, and a well-developed infrastructure of semiconductor companies and cloud providers. These strengths have positioned the U.S. at the forefront of state-of-the-art AI model development and training, with large-scale model runs frequently conducted on domestic cloud platforms.

In 2010, the U.S. Air Force Research Laboratory built the Condor Cluster, one of the most cost-effective and unconventional supercomputers of its time. The system combined 1,716 Sony PlayStation 3 consoles with 78 compute head nodes containing a total of 94 NVIDIA Tesla C2050 GPUs and 62 Tesla C1060 GPUs. In total, the system featured nearly 70,000 processing cores and delivered a theoretical peak performance of 500 teraflops.

The Condor Cluster was designed for tasks such as satellite imagery analysis, radar enhancement, and early artificial intelligence research. It demonstrated how consumer and graphics hardware could be creatively repurposed for national defense applications. As one of the first military systems to integrate GPU-based parallel processing at scale, the project highlighted the potential of graphics processors in supporting advanced compute needs, laying an early foundation for the widespread use of GPUs in defense and intelligence programs. [54]

The Condor Cluster was a milestone in showing how affordable, off-the-shelf technology could be repurposed for defense applications. It also marked one of the first major uses of gaming hardware in military computing, foreshadowing today's reliance on GPUs for running AI models and large-scale data analysis. Namely, the most advanced GPUs from NVIDIA.

Geopolitical Tensions

China, however, has rapidly closed much of the gap by making AI a national strategic priority. With a population

of over 1.4 billion and high adoption of digital platforms, China possesses access to enormous datasets that are invaluable for training AI systems in areas such as recommendation engines, facial recognition, and natural language processing.[55] Domestic firms like Baidu, Alibaba, Tencent, and Huawei have developed their own AI research labs, cloud platforms, and chip architectures, aiming to reduce reliance on Western hardware.

While China has made major strides in AI development and infrastructure, it faces key constraints, particularly around access to advanced semiconductors. U.S. export controls have limited China's ability to acquire top-tier GPUs, such as the A100 and H100, prompting a shift toward domestically designed chips and older fabrication nodes. Additionally, China's energy policy presents another vulnerability. Periodic power shortages and regional electricity curtailments have disrupted data center operations, in contrast to more stable and efficient hyperscale infrastructure in the U.S.[56]

Both nations view AI as a critical frontier for military modernization. The U.S. has invested in AI applications for autonomous vehicles, surveillance, and simulation through agencies like DARPA and the Department of Defense.[57] China, meanwhile, integrates AI across its internal surveillance systems and defense platforms, including swarming drones, missile guidance, and predictive policing. In China, the boundary between commercial and military AI is less defined, as state directives often direct private-sector innovation toward national objectives.[58]

Power

Going forward, efficiency in infrastructure and energy usage will likely shape long-term leadership. AI's growing resource demands require optimized chips, advanced cooling technologies, and reliable access to power. The United States benefits from competitive energy markets and infrastructure innovation, while China's data centers have historically relied more heavily on coal. As both nations invest in greener AI and aim to reduce the carbon footprint of large-scale computing, the ability to scale AI sustainably and cost-effectively may prove to be a decisive advantage in the global AI race.[59]

Regulation Proposals

Recent regulatory proposals, known as the AI Diffusion guidelines, would introduce a global licensing framework for AI-related integrated circuits. The system divides countries into tiers, granting full access to close allies while imposing strict limits or full bans on others. If implemented, these restrictions could significantly reduce demand from key international markets, impacting revenue, cloud services, and future R&D capacity for U.S. chipmakers.[60]

Long-term, the effects of such policies may extend beyond near-term revenue losses for individual companies. Countries restricted under these guidelines, particularly those in lower tiers, are still expected to accelerate efforts to develop or adopt alternative AI hardware providers. As semiconductors become foundational to economic growth and digital sovereignty, national governments are unlikely to remain dependent on a supply chain they cannot reliably access. This dynamic may shift parts of the global AI ecosystem outside the U.S. economic and technological sphere of influence.

Our stance on this policy is that the restriction of advanced semiconductors reflects the broader trend of U.S. protectionism, which has included rising tariffs in recent years. We believe American innovation is a core driver of U.S. power and international influence. However, limiting access to advanced chips can force countries out of the U.S.-centered technology ecosystem and reduce their dependence on American infrastructure and supply chains.

While nations would ideally use the most advanced, energy-efficient chips with high compute capabilities, a less powerful and more energy-intensive chip is still preferable to being excluded entirely. These restrictions risk pushing buyers toward alternative suppliers. Although China is not currently outpacing the United States in innovation, removing its access to U.S. chips could encourage the country to increase investment in domestic development, expand production, and gain market share in both local and global markets.

This is similar to what has happened with AMD. Despite producing chips that underperform compared to NVIDIA's,

AMD has gained market share due to better pricing and broader availability. A similar dynamic could play out if countries restricted by U.S. export controls turn to Chinese alternatives.

Our biggest concern is that limiting access to semiconductors, often referred to as the "new oil," may weaken American soft power. While isolating China makes strategic sense in light of ongoing unfair trade practices such as intellectual property theft and currency manipulation, it also carries risk. China may use initiatives like the Belt and Road to expand its influence, especially through debt-driven diplomacy. Reducing access to critical technology without securing alternative alliances could leave room for China to grow its geopolitical and technological footprint.[61]

Demographic Changes

Global life expectancy has increased due to advances in healthcare, hygiene, agriculture, and information sharing. However, as birth rates decline, particularly in developed nations, aging populations are placing growing pressure on labor markets. Many countries have relied on immigration to support economic growth and fill essential jobs, especially in sectors such as elder care. Yet, rising political pressure and tighter immigration policies have constrained the flow of migrant labor.

In the United States, recent executive actions and growing public support for stricter immigration enforcement have resulted in a shrinking supply of undocumented workers, many of whom occupy low-wage, labor-intensive roles. Combined with demographic trends, this shift raises concerns over labor shortages that could hinder economic output.

To offset this, economies will increasingly turn to automation and artificial intelligence. Agentic AI, autonomous systems capable of reasoning, learning, and acting without constant human input, offers a promising solution to supplement the workforce. These systems can perform tasks in logistics, manufacturing, customer service, and administrative functions, reducing dependency on human labor.

By integrating agentic AI into the workforce, countries can maintain or even increase productivity, despite demographic headwinds, labor constraints, and rising wage pressures.

KEYS TO MONITOR

Despite rapid progress in artificial intelligence, several structural risks remain that could constrain future growth or trigger another slowdown in advancement. Monitoring the following indicators will be critical in assessing the resilience of the current AI cycle and identifying potential inflection points.

Stalling AI Progress

If AI models begin to show diminishing returns from increased scale or improvements in performance plateau, it could signal the early signs of a slowdown. A prolonged period without commercially viable breakthroughs or cost-efficiency gains would raise concern over whether AI is entering another cycle of disillusionment similar to past "AI winters."

Energy Availability and Infrastructure Constraints

AI workloads require immense energy resources. Continued access to reliable and scalable energy, particularly nuclear and natural gas, will be necessary to sustain data center growth. Delays in energy infrastructure development or constraints in grid availability could limit the expansion of high-density compute environments.

Excess Capital Expenditure in Data Centers

A sharp increase in data center investment that outpaces demand for AI services could lead to overcapacity. If return on investment fails to materialize, especially in a high interest rate environment, infrastructure-heavy firms could face balance sheet pressure, and the broader sector may experience a correction.

Legal and IP Risk in Training Data

Much of today's AI has been trained on publicly available internet data. However, legal uncertainty over data ownership and copyright could impose costly licensing requirements. If courts rule against current practices, companies may be forced to re-architect models or limit training inputs, slowing future development.

Regulatory Pressure on Hyperscalers

Governments are beginning to scrutinize the role of hyperscale cloud providers more closely. Regulations targeting antitrust, data usage, sustainability, and AI safety could increase compliance costs and limit compute availability. Policy decisions around permitting and zoning for new data centers will also be important to track.

Limits to Semiconductor Scaling

As Moore's Law slows and transistor scaling approaches physical limits, the cost of compute continues to rise. The pace of innovation in chip architecture, including stacking and AI-specific accelerators, will be essential. A failure to offset the decline of Moore's Law would materially impact model development timelines and overall AI affordability.

Conclusion

Artificial intelligence is positioned to be one of the most transformative forces in the global economy, reshaping how productivity, infrastructure, and influence are distributed across sectors and borders. While momentum behind adoption remains strong, the sustainability of AI's impact will depend on the ability to navigate regulatory uncertainty, energy constraints, and geopolitical fragmentation.

The next phase of growth will rely not only on technological breakthroughs, but also on the capacity to scale systems responsibly while addressing legal, environmental, and economic challenges. Long-term advantage will favor those who secure access to reliable energy, high-quality talent, and resilient digital infrastructure, all while remaining competitive in global markets. The decade ahead will determine whether AI can evolve into a general-purpose technology that drives durable economic value or becomes limited by the same constraints it aims to overcome.

REFERENCES

1. Stanford
2. Google
3. Amazon
4. Wall Street Journal
5. Dartmouth
6. MIT
7. Stanford
8. Arizona
9. Toronto
10. Toronto
11. Toronto
12. Yann LeCun
13. Meta
14. Toronto
15. NVIDIA
16. LinkedIn
17. MIT
18. IBM
19. Amazon
20. Iowa
21. IBM
22. Google
23. Stanford
24. Google
25. IBM
26. IBM
27. Research Gate
28. Intel
29. TechCrunch
30. NVIDIA
31. NVIDIA 10-K
32. Datacenters.com
33. PCIM Insights
34. ASML
35. TSMC
36. New York Times
37. TSMC
38. TSMC
39. DeepSeek
40. NASDAQ
41. New York Times
42. Artificial Analysis
43. Statista
44. Reuters
45. New York Times
46. New York Times
47. New York Times
48. New York Times
49. Meta
50. Reuters
51. New York Times
52. NASDAQ

53. Goldman
54. DTIC
55. RAND Corporation
56. UT Sydney
57. DARPA
58. Brookings
59. Statista
60. Federal Register
61. BBC

DISCLAIMER

Henry Fund reports are created by graduate students in the Applied Securities Management program at the University of Iowa's Tippie College of Business. These reports provide potential employers and other interested parties an example of the analytical skills, investment knowledge, and communication abilities of our students. Henry Fund analysts are not registered investment advisors, brokers or licensed financial professionals. The investment opinion contained in this report does not represent an offer or solicitation to buy or sell any of the aforementioned securities. Unless otherwise noted, facts and figures included in this report are from publicly available sources. This report is not a complete compilation of data, and its accuracy is not guaranteed. From time to time, the University of Iowa, its faculty, staff, students, or the Henry Fund may hold an investment position in the companies mentioned in this report.