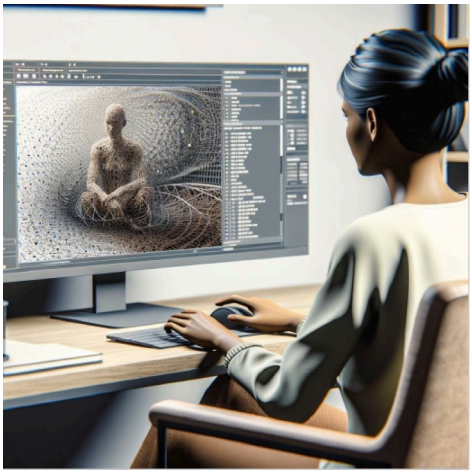
5 July 2024, 15:18 BST
Chief Investment Office GWM
Investment Research

Artificial intelligence: Sizing and seizing the investment opportunity

Transformational Innovation Opportunities (TRIO): Artificial Intelligence

Ulrike Hoffmann-Burchardi, Head CIO Global Equities, UBS Financial Services Inc. (UBS FS)
Kevin Dennean, CFA, CIO Equity Strategist, US Technology & Telecom, UBS Financial Services Inc. (UBS FS)
Achille Monnet, CIO Equity Sector Strategist, UBS Switzerland AG
Alexander Stiehler, CFA, CIO Head of Longer Term Investment Themes, UBS Switzerland AG
Sundeep Gantori, CFA, CAIA, CIO Equity Strategist, UBS AG Singapore Branch
Delwin Kurnia Limas, CFA, Equity Strategist, UBS AG Singapore Branch
Bennett Chu, CAIA, Equity Strategist, UBS AG Singapore Branch
Linda Mazziotta, CFA, Strategist, UBS Switzerland AG
Nikolaos Fostieris, Strategist, UBS Switzerland AG
Antonia Sariyska, CIO Sustainable and Impact Investing Strategist, UBS Switzerland AG

- ChatGPT is a watershed moment for artificial intelligence (AI) and its adoption. We are witnessing the start of a major investment boom and technological advance that may fundamentally affect all economic sectors.
- However, use cases need to develop further to justify these investments. In the early innings of the AI era, we recommend investors to focus on vertically integrated players across the AI value chain, as those businesses combine good visibility on monetization and strong competitive positioning.
- In this report, we outline our investment framework to identify AI opportunities; shed light on how AI works; highlight its implications for the global economy, sustainable development, and individual sectors; and share where we think investors should focus.



Source: Dall-e, UBS as of June 2024
Note: This image has been generated.

Introduction

November 30, 2022. We will look back at this date as the watershed moment for a new innovation cycle, one that may prove to be the biggest and most profound in human history. The launch of ChatGPT on that day was an inflection point for broad adoption of artificial intelligence (AI), much like the Netscape browser was for the internet and Windows for the PC. The term artificial intelligence dates back to 1955 when the word was first introduced by John McCarthy at Dartmouth College. With the launch of ChatGPT, AI has now gone mainstream. ChatGPT saw the fastest user adoption ever, attracting 100 million users in just two months after its launch.

Three features distinguish this new wave of generative AI from past versions: its ability to generate content; the scale of data inputs, model parameters, and compute; and its adaptive learning. Generative AI can generate human-like long-form content such as text, images, and video. It uses vast amounts

This report has been prepared by UBS Financial Services Inc. (UBS FS), UBS Switzerland AG, UBS AG Singapore Branch.

of data to identify patterns to generate content. For example, ChatGPT was trained on around 300 billion words, representing 570 gigabytes of data[1]. The models adaptively learn over time as they process more data and improve their ability to generate relevant output.

With more than 18 months into this cycle, some are beginning to wonder whether AI has become overhyped—and others whether it's underappreciated. The rate of innovation is so rapid that the potential of generative AI applications continuously exceeds already lofty expectations. Over the last two years alone, large language models have gone from mediocre to expert test takers[2]. Text-to-image models have gone from designing avocado chairs to creating videos of red pandas running around a bamboo forest inside a Petri dish. The context window has gone from 16,000 tokens to 1 million in less than a year.

The range of problems that AI can address keeps growing, enhancing the productivity of knowledge workers every day. With 1 billion knowledge workers worldwide, the productivity gains could easily surpass that of the internet—making it one of the biggest investment opportunities over the next decades.

The potential of this new productivity boom has kicked off a capex cycle to build AI data centers that will likely surpass the size of traditional data centers in the next few years. To justify these investments, we have to see use cases materialize at scale—whether in advertising, customer service, personal and coding assistants, R&D, cybersecurity, and more.

The purpose of chapter 1 is to outline a framework for the investment opportunity set of this new technology. For equities, nothing impacts returns more profoundly than innovation. A mere 2.4% of companies have been responsible for close to 100% of net wealth creation in the stock market since 1990[3]. These companies have driven most of the growth, fueled by innovation around the internet, mobile, and cloud. We believe over the next decades, the companies that provide and adopt AI will be at the top of this league table.

Content overview

Ten predictions about AI

1. AI will be the most profound innovation and one of the largest investment opportunities in human history
2. AI will kick off a data center capex cycle that will dwarf general purpose data center capex in the next years
3. The ratio of monetization of the AI application layer to the costs of the enabling and intelligence layers will become a key metric for investment returns
4. The race to artificial general intelligence (AGI) could trigger a capex cycle that inflates an investment bubble where the capex of the enabling layer is dissociated from near-term monetization potential of the application layer
5. The AI enablers will be the first adopters of AI, driving both revenue and margin upside
6. Monolithic players will emerge along the AI value chain and over time, the AI market will be dominated by an oligopoly of vertically integrated "AI foundries"

7. The AI silicon moment: AI chips will capture a large part of the AI value creation
8. The application and intelligence layers will merge with AGI
9. Software will become ubiquitous
10. Data assets will emerge as the competitive differentiators for AI adopters

Chapter 1: Just how large? Sizing the AI opportunity

"The potential size of this market is hard to grasp—somewhere between all software and all human endeavors." [4]

1. **How large could the AI opportunity become?**

Top-down

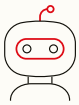
Estimates for the size of the AI market vary widely—from Bloomberg's USD 1.3tr by 2032[5] to McKinsey's USD 4.4tr[6]. Still, most agree that annual AI-related revenues could reach the trillion-dollar threshold over the next decade. A top-down approach to estimating AI's value creation is to assess the productivity improvement that results from adding AI tools to knowledge workers. Generative AI can improve workers' ability to do cognitively challenging knowledge work. Examples of such tasks include using AI to develop new products or services, problem solve, or create business strategies. Based on estimates from Gartner, there are 1 billion knowledge workers worldwide. Only two of the International Labor Organization (ILO) categories of workers can be considered as consisting of knowledge workers: managers and clerical staff. For both, the average annual salary is USD 29,000[7].

While it is too early to accurately quantify the aggregate productivity enhancements from AI, anecdotal evidence suggests substantial efficiency gains. For example, developers code up to 55% faster with the use of GitHub Copilot[8]. Boston Consulting Group estimates that customer service operations will become 30–50% more efficient when generative AI is implemented at scale.[9] The table below lists more examples of efficiency gains in the order of 25–30%.

Examples of productivity gains from generative AI

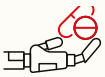
As shown in a Harvard study of GPT-4 versus a large population of high-performing consultants at Boston Consulting Group, GPT-4 was 25.1% faster and produced outputs at more than 40% higher quality than human experts.

Research (NBER working paper "Generative AI at work") found that at one company with more than 5,000 customer service agents, the application of generative AI increased issue resolution by 14% an hour and reduced the time spent handling an issue by 9%.

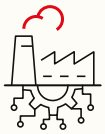


Payment service company Klarna says its AI assistant is doing the equivalent work of 700 full-time agents and it is more accurate in errand resolution, leading to a 25% drop in repeat inquiries.

Study (Kalliamvakou 2022) showed that software engineers can code up to twice as fast using a tool called Codex, based on the previous version of the large language model GPT-3.



As reported by the Capgemini Research Institute ("Harnessing the value of generative AI", 2023) German biotech firm Evotec announced a phase-one clinical trial for a novel anti-cancer compound. By using an AI design platform, the drug candidate was identified in only eight months (vs. average 4–5 years).



Germany-based Claudius Peters produces processing equipment for cement, coal, alumina, and gypsum plants. Its aim was to reduce costs and product weight. In this example mentioned in the same report by the Capgemini Research Institute, the generative design produced components with a remarkable 20–60% weight reduction while meeting performance requirements. The design was used as a re-engineering template for conventional manufacturing, resulting in a 30% lighter final design that lowered component costs.

Another approach is to study the productivity enhancement from past innovation cycles. Studies have shown that the PC increased labor productivity by 18% from 1986 to 2000 and the internet has similarly increased labor productivity by 20% from 2000 to present[10]. If we assume a 15% productivity boost from AI and 1 billion knowledge workers with an annual salary of USD 29,000, AI value creation could amount to USD 4.4tr.

Figure 1 - Potential value creation from AI productivity enhancements, in USD tr

Based on USD 29,000 salary for average knowledge worker and 1bn knowledge workers globally

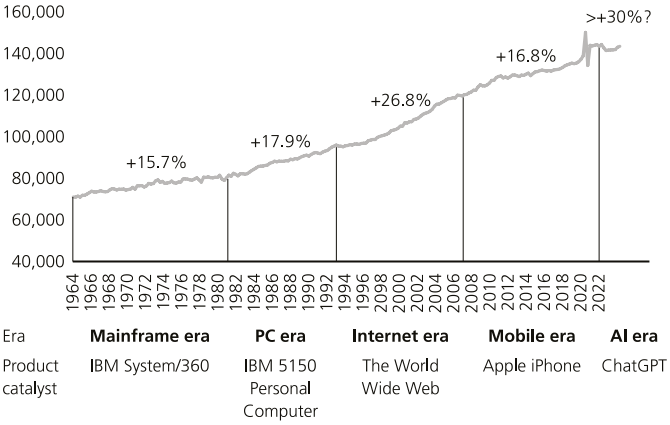
Productivity increase	10%	15%	25%
Value creation	USD 2.9tr	USD 4.4tr	USD 7.3tr

Source: International Labor Organization (ILO), UBS estimates, as of May 2024

And lastly, a third way to assess the value creation from AI is by looking at real GDP per worker. In the US, one worker now produces about twice the amount of real GDP than in 1964. The biggest increase in output per worker occurred during the internet era. Some of the efficiency gains during this period coincided with the offshoring of labor, which likely made this period stand out compared to other innovation cycles.

If the productivity examples above are applicable more broadly, we could see this metric accelerating further with the adoption of genAI.

Figure 2 - Real US GDP per worker during innovation cycles



Source: Bloomberg, UBS, as of May 2024

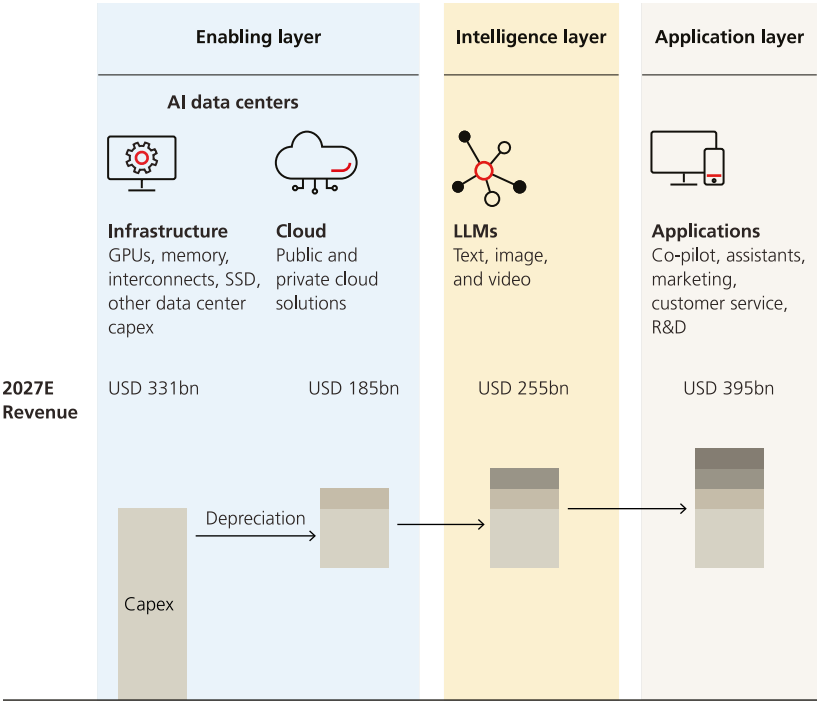
Prediction: AI will be the most profound innovation and one of the largest investment opportunities in human history

Bottom-up AI market size estimates: AI value chain

To complement the top-down approach, it is useful to follow the value creation along the AI value chain from a bottom-up perspective. There are three layers in the AI value chain: It starts with the enabling layer, which powers the intelligence layer, which in turn feeds the application layer. The enabling layer is the physical layer of data center infrastructure and compute that is necessary to train and run generative AI models. The intelligence and application layers are both software layers and refer to the data and algorithms that are used to build models for applications such as generating a piece of text on a particular subject area.

An overview of the AI value chain illustrates the economic value creation. From a bottom-up lens, we see an annual value creation of USD 1.16tr by 2027. We see the largest near-term opportunities in the enabling layer and still expect the ratio of applications/enabling and intelligence layer to imply limited bottom-line profitability for the application layer during first stages of the cyclical and structural ramp of genAI. Figure 3 illustrates the composition of the value creation by each layer of the AI value chain.

Figure 3 - The AI market opportunity: a bottom-up perspective



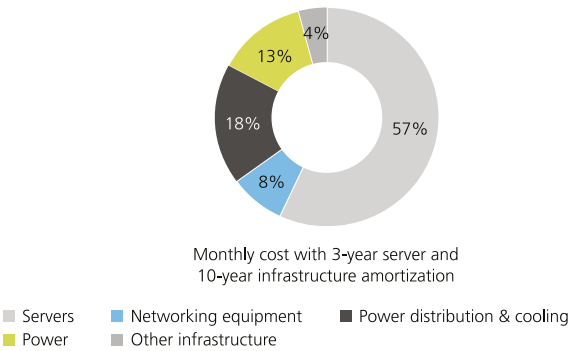
Source: UBS estimates, May 2024

1. **The enabling layer**

The enabling layer consists of AI data centers that can offer computing resources at scale. They can be either on-premise in a company's private cloud datacenter or made available through public cloud service providers such as Amazon AWS, Microsoft Azure, or Google GCP. The AI server sits at the heart of the data center with components such as graphics processing units (GPUs), memory, interconnects, and storage. Besides an AI server, data centers need cabling, power supplies, cooling, and other equipment. Figure 4 below illustrates the typical monthly cost structure of a data center.

Figure 4 - Data center cost structure

In %



Source: https://perspectives.mvdirona.com/2010/09/overall-data-center-costs/, UBS

We expect USD 331bn in annual capex for the enabling layer by 2027, comprised of AI server and AI data center infrastructure capex. Most of the value in the enabling layer is likely captured by AI servers, much like in a general-purpose data center illustrated above. Because of the scale of AI compute, most companies will likely consume compute resources in the form of cloud services. As a result, we expect USD 185bn in value creation to be generated by 2027. Part of this value creation will be realized through internal productivity gains inside companies that are AI enablers. The AI chip companies and cloud hyperscalers both operate in oligopolistic markets with attractive gross margins. This suggests that companies in these markets will be able to capture sizeable profits.

The data center numbers may prove conservative if there's a rush to build out supercomputing facilities among the large cloud providers. Reports suggest that Microsoft and OpenAI are embarking on a data center project that could cost as much as USD 100bn, with an AI supercomputer called Stargate[11] to be deployed by 2028. With more powerful models requiring compute scale, we may see a competitive race of data center build-outs between the largest cloud vendors.

Prediction: AI will kick off a data center capex cycle that will dwarf general purpose data center capex in the next years

2. Intelligence layer

The next layer is the intelligence layer. This part includes the generative AI algorithms and large language models (LLMs), which use the scale of computing resources from the enabling layer. This layer assumes the strategic function as the "brain" of generative AI. We are in the early stages of monetization, and pricing of this brain varies from usage-based models to monthly subscriptions. LLMs are developed by private companies like OpenAI, Anthropic, and Mistral, or as part of larger public companies such as Google and Meta. Therefore, there is no publicly reported data on the financials of these businesses. According to the FT, OpenAI hit the USD 2bn revenue mark in December 2023[12] with 25% gross margins. We expect this layer to show the strongest growth into 2027 given its small base.

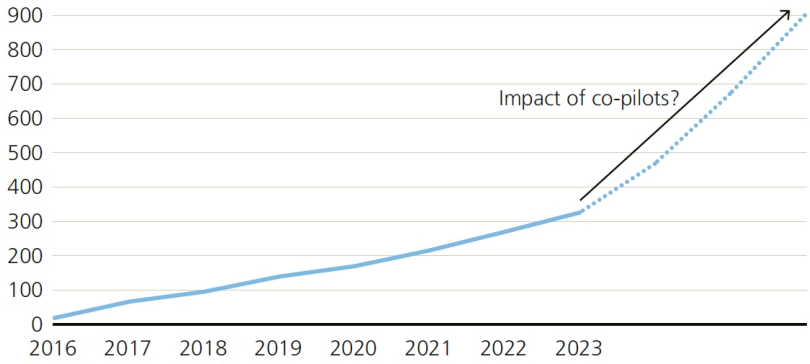
3. Application layer

The application layer likely offers the largest monetization potential over time, yet this opportunity is difficult to quantify at this early stage. Like the intelligence layer, the application layer is a software layer ranging from AI-powered software applications and services that are readily apparent to the user to embedded AI functionality in other software applications.

One of most tangible use cases is to use LLMs as co-pilots as coding and personal assistants. These assistants are able to operate with different modalities, text, audio, and video. In 2023, Microsoft's coding advisor GitHub Copilot exceeded USD 100mn of revenue[13], was profitable, and grew 40% year-over-year with 1.3 million users[14]. With developer productivity gains of 50–60% (based on Peng et al. (2023): The Impact of AI on Developer

Productivity: Evidence from GitHub Copilot), we expect an acceleration in the creation of software code.

Figure 5 - Total number of GitHub repositories
In millions



Note: Assuming a growth rate 55% higher than the previous 5-year average

Source: GitHub

Note: This is for educational purposes only. For investment ideas and our stock picks, see our companion report "Investing through the AI platform shift."

Prediction: Software will become ubiquitous

Additionally, there have been successful proof points for using LLMs in other large markets such as digital advertising (USD 667bn market in 2024[15]) and call centers (USD 315bn[16] market in 2022). Further, we see successful generative AI applications in healthcare R&D, cybersecurity, and fintech.

At this stage, we expect a directly addressable market of USD 395bn in revenue opportunities for the application layer by 2027, predominantly driven by AI assistants (USD 75bn), applications of generative AI for advertising (USD 100bn), and customer service (USD 75bn).

Economic value by layer

It's useful to illustrate the AI value chain by layer of economic value creation, because each layer has to create enough economic value to justify the costs of the preceding layer. In other words, the revenue of the enabling layer becomes the cost for the intelligence layer, and the application layer has to create enough value to absorb the costs of both the enabling and intelligence layers. For example, a company like Jasper that uses OpenAI as its intelligence layer and Microsoft's Azure cloud services for the enabling layer will have to generate sufficient revenue to cover the costs for both OpenAI and Azure services.

The example below shows that USD 100 of operational expenditures on the enabling layer (such as cloud services) will require USD 400 of revenues on the application layer if both the intelligence and application layers have 50% gross margins.

Figure 6 - The AI value chain – margin layers (assuming opex)

	Gross margin	Revenue	Cost of goods sold
Enabling layer		USD 100	
Intelligence layer	50%	USD 200	USD 100
Application layer	50%	USD 400	USD 200

Source: UBS estimates, May 2024

Gartner estimates that data center systems generated USD 236bn revenue in 2023, while software generated USD 915bn and IT services USD 1.39tr. This implies that every dollar of hardware spent generated roughly USD 10 of revenue. While there are certainly many variables, we think this simple example highlights the point that the enabling layer in the AI stack should become a fraction of the services it supports. In looking at AI-driven businesses, we observe that capital intensity (capital spending divided by revenue) has been in the mid-teens to the low-thirty percent range for both Alphabet and Meta Platforms.

One of the key ratios to watch is therefore the ratio of monetization potential of the application layer to the costs of the enabling and intelligence layers. This is especially true for companies that only operate in single layers of the value chain and cannot trade off margin in one layer for another. We need to see continued advances in the performance of LLMs and successful applications to justify the current capex build-out. A lack of material model advances, slowing consumer and enterprise uptake, and new algorithms requiring less compute could lead to a glut in infrastructure components such as GPUs, ASICS, and memory.

Prediction: The ratio of monetization of the AI application layer to the costs of the enabling and intelligence layers will become a key metric for investment returns

Multi-layer versus single layer: Layer cake vs. monolith

With an attractive margin structure along the value chain, it is likely that an increasing number of companies will try to expand into different segments of the value chain. Amazon already designs its own AI chips, Trainium and Inferentia, while Microsoft has been working on its custom AI chip Athena. We expect that companies operating in the intelligence layer will explore value-add applications specific to certain use cases, such as medical diagnostics or customer service, while also working on an AI super-agent that can be universally useful. Given the scale effects of cloud compute and generative AI, it is likely that over time, the application layer will be dominated by the same players as in the intelligence and enabling layers and that there will be an oligopoly of fully vertically integrated AI foundries.

Prediction: Monolithic players emerge along the AI value chain and over time the AI market will be dominated by an oligopoly of vertically integrated "AI foundries"

Figure 7 - The AI market opportunity: a bottom-up perspective

Company	Enabling layer		Intelligence layer	Application layer
	Chips	Cloud		
Google	TPUs	GCP	Gemini (formerly Bard)	Duet AI, Advertising
Microsoft	Athena	Azure	Open AI (investment)	GitHub, Office Co-pilot
Amazon	Trainium, Inferentia	AWS	?	Chatbot recommendations
Meta	MTIA	?	Llama	Advertising
Nvidia	GPUs	DGX	?	?
Tencent	?	Tencent Cloud	Hunyuan	Advertising
Baidu	?	Wangpan	Ernie	Baidu Comate
Alibaba	?	Aliyun	Qwen	Qwent-Agent
Huawei	GPUs	Huawei Cloud	PanGu	PanGu Drug Molecule Model

Source: UBS estimates, May 2024

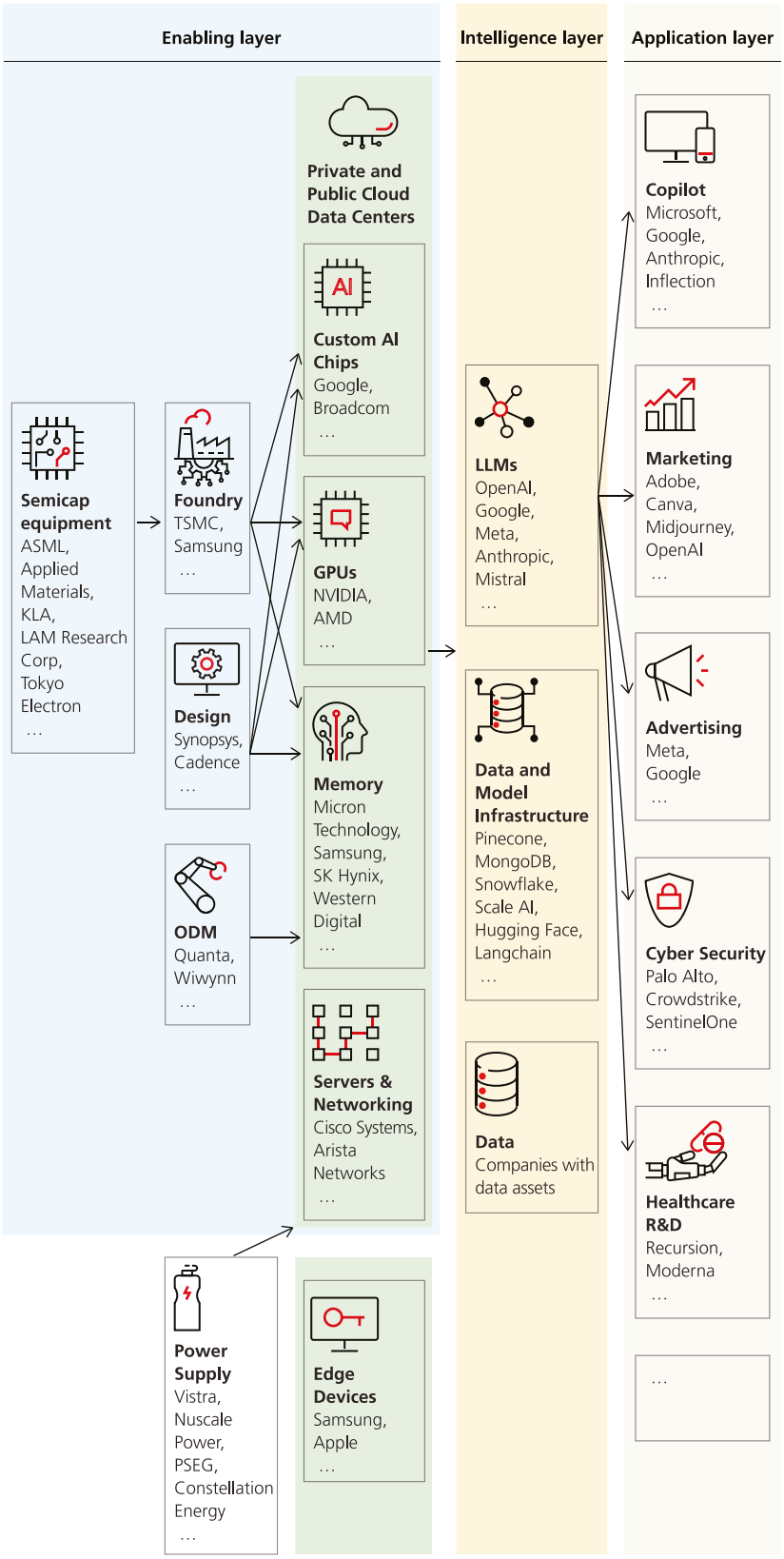
Further, the enablers of AI are those that are most adept at using it and, hence, will be their own customers, driving productivity gains.

Prediction: The AI enablers will be the first adopters of AI driving both revenue and margin upside

2. Opportunities along the AI value chain: A deep dive

This section provides a deeper look into the different components of the AI value chain and discusses the AI investment opportunity set for each of the components (see Figure 8). Companies can operate in multiple segments of the value chain. Google, for example, designs its own chips, large language models, and productivity and advertising products and thus covers all three layers of the value chain. Other companies (such as NVIDIA) currently only operate in a single segment (GPUs). Lastly, companies can have exposure to multiple segments of the value chain through investments in other segments, as in the case of Microsoft, which holds a 49% stake in OpenAI.

Figure 8 - The AI value chain – illustrative with select companies



Source: UBS, May 2024

Note: This is for educational purposes only. For investment ideas and our stock picks, see our companion report "Investing through the AI platform shift."

Monetizing and capturing value in the enabling layer

1. **AI chips**

Chips, GPUs, and custom AI chips such as tensor processing units (TPUs) are the key components for the enabling layer. Figure 9 below shows that over 70% of the bill of materials of an AI server is made up of GPU costs. This makes it the largest beneficiary, followed by memory and interconnects, which make up 15% and 9% of the total costs, respectively.

Prediction: The AI silicon moment: AI chips will capture a large part of the value creation

As reported by the FT in December 2023, Lisa Su, the CEO of AMD, estimates that AI chips will make up USD 400bn of the semi industry's global revenues by 2027—a significant increase from the USD 150bn prediction she made less than four months earlier, and a figure equal to the entire global semiconductor market in 2019.

Figure 9 - Representative AI server bill of materials

Item	Category	Cost (USD)	Quantity	Total (USD)	% of total
NVIDIA HGX H100 80Gb	GPU	34,000	8	272,000	74%
Intel Xeon	CPU	1,555	2	3,110	1%
32 x 128GB DDR5	Memory/DRAM	1,700	32	54,400	15%
M.2 SSD – Max 2	Storage	770	2	1,540	0%
U.2/U.3 NVMe 2.5	Storage	142	16	2,272	1%
SATA SSD 2.5 – Max 3	Storage	468	3	1,404	0%
Mellanox MCX653106A-HDAT-SP Connect	Interconnects	1,802	18	32,436	9%
				367,162	100%

Source: Dihuni.com with UBS inputs, May 2024

Note: This is for educational purposes only. For investment ideas and our stock picks, see our companion report "Investing through the AI platform shift."

There are two types of AI chips: GPUs and custom AI chips such as Google's TPU, which are an example of an application-specific integrated circuit chip (ASIC).

a) GPUs

The AI GPU computing segment provides chips for both AI training and inference. The segment's market cap was already close to USD 2.5tr in March 2024, and we believe as the key infrastructure spending beneficiaries, GPU companies should ride the AI wave over the years ahead. The combined revenues for the AI GPU computing segment were close to USD 40bn in 2023, with the largest GPU maker maintaining a dominant 98.5% share. Thanks to robust training and inference demand (as evident in recent comments that inference already accounts for almost 40% of one player's AI GPU demand),

we expect the AI GPU industry to post strong revenue growth in both 2024 and 2025. We forecast the industry's revenues will more than double to USD 88bn in 2024 and grow by another 25% in 2025 to reach close to USD 110bn.

b) Custom AI chips

Custom chips mostly refer to the accelerator chips designed by big tech and other startups as a complement to expensive GPUs, which currently cost somewhere between USD 15,000 and USD 30,000.

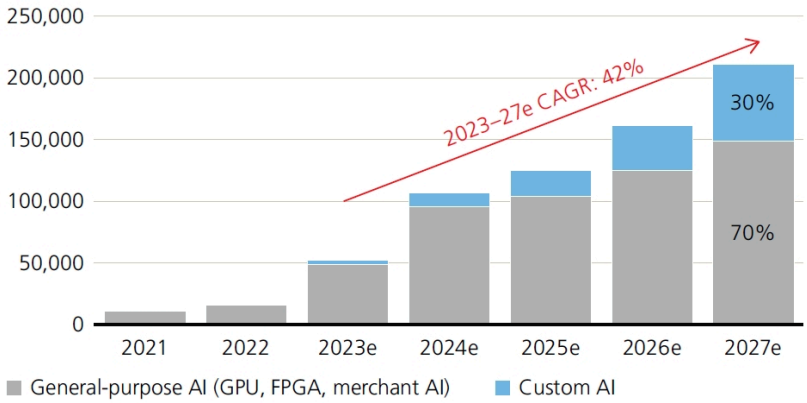
Custom AI chips, through the ASIC (application-specific integrated circuit) route, can be cost-effective for the leading platform companies in special cases like inference and training smaller LLMs, even if they may lag the performance of AI GPUs.

Unlike AI GPUs, where disclosures are clear given the segment only has two major players, data for AI custom chips in areas like pricing is not freely available, as the custom chips are meant for internal consumption. For instance, the total cost of goods sold for NVIDIA's H100 is only USD 3,367, based on our calculations, compared to a retail price of USD 25,000 (effectively translating to an 86.5% gross margin for NVIDIA on that chip). The cost dynamics for AI custom chips may not be too different compared to H100 chips, but considering they are mostly used for internal purposes and not for sale, we don't think the sales data of NVIDIA's AI GPUs can be compared with the cost data of AI custom chips.

According to Morgan Stanley estimates, custom AI chips (ASICs) currently represent less than 10% of the total AI computing market. But Morgan Stanley expects the segment to outgrow GPUs and potentially take up to 30% of the cloud AI semi market in 4–5 years.

Figure 10 - Custom AI chips (ASICs) are expected to outgrow GPUs and potentially take up to 30% of the cloud AI semi market in 4–5 years

Cloud semi breakdown: general AI vs. custom AI, in USD million



Source: Company data, Gartner, Morgan Stanley Research estimates (e)

2. Foundries

The foundry industry is dominated by a single company that controls more than half of the market, with a disproportionately high market share in leading-edge and AI-related GPUs and custom chips. During its January conference call, that company raised its medium-term AI outlook (consistent with our recent upward revision) to now expect data-center AI revenues to contribute more than 20% of its 2027 revenues, up from its previous guidance of 10–15%.

With a strong contribution from AI custom chips, we think semiconductor foundry companies have a bright future. While some may see a slightly lower contribution from AI, it is clear to us that AI will be the key growth driver for foundries in this decade. For reference, smart devices today contribute almost 40% of revenues to foundries; in the long term, AI may grow to a similar share.

3. Semicap equipment

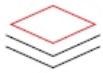
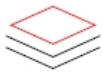
Driven by strong growth in global semiconductors, we expect global semiconductor capital (or semicap for short) equipment to report steady growth, with industry revenues likely to cross USD 100bn for the first time in 2025. While the industry is still absorbing excess investments made during the pandemic, particularly for legacy or less-advanced nodes, we still see decent 7% y/y growth in 2024 and 11.5% y/y in 2025, with AI-driven spending trends remaining the key catalysts.

Memory should be a key growth driver in the near term thanks to increased spending on, among others, the high bandwidth memory (HBM) segment, which is primarily used for AI processing in data centers.

While foundry capex spending should be muted in 2024 with a likely year-on-year decline due to increased discipline, we expect a decent recovery in 2025 with a mid-to-high single-digit year-on-year increase. In particular, custom chips and edge-computing should drive strong spending, with a potential hardware refresh cycle acting as an additional catalyst. At this stage, we do not project any major capex contribution from OpenAI's CEO Sam Altman's proposed USD 7 trillion investment into semiconductor manufacturing or other such initiatives, but any progress would be an additional positive for the industry.

In summary, the semicap equipment industry could be one of the few standout segments within semiconductors in 2025. With revenue growth set to accelerate, and together with strong margin support and the sector's oligopolistic characteristics, we continue to see share price catalysts for the industry.

Figure 11 - Overview of select semicap players and process exposure

	2023 total revenues (USD bn)	Key process	
Applied Materials	26.5	Deposition & Etch	
Tokyo Electron	13.1	Deposition & Etch	
ASML	23.7	Lithography	
Lam Research	14.3	Etch	
KLAC	9.7	Process control	

Source: Company reports, UBS estimates, as of March 2024

Note: This is for educational purposes only. For investment ideas and our stock picks, see our companion report "Investing through the AI platform shift."

4. Memory

While we believe the GPU supply chain still has among the most visibility in the AI computing segment in terms of near-term growth thanks to strong training and inferencing demand, we see memory as a key beneficiary as AI infrastructure demand broadens. This is based on our view that AI will potentially drive mid-teen demand growth for dynamic random access memory (DRAM) in 2024 (vs. a low-mid single-digit contribution in 2023). As a result, we see strong near-term momentum for DRAM pricing fueled by a rising contribution from the high-bandwidth memory (HBM) segment used in AI computing.

High-bandwidth memory is dedicated DRAM used in high-performance computing applications, like how AI uses vertically stacked chips. We expect HBM penetration to keep rising in generative AI applications. For instance, NVIDIA's popular H100 GPU used in generative AI can support 80 gigabytes (GB) of HBM, whereas AMD's upcoming MI300X can support 192GB of HBM. The rising HBM penetration is positive for the memory industry, considering the significant premium HBM enjoys versus traditional DRAM chips (HBM price is 5–6x standard DDR4 memory pricing).

Hence, despite the muted recovery in other tech hardware segments, AI should provide a strong catalyst for the global memory industry, which should benefit from a cyclical recovery due to a favorable demand-supply balance. Our checks suggest utilization rates for DRAM supplies are set to touch 100% very soon—a V-shaped recovery from last year's lows—and a current run rate of about 87% thanks to strong procurement from leading AI customers. In terms of pricing, we see upside risks to DRAM prices with a potential rebound of close to 80% in this cycle from last year's lows. Against this backdrop and

with increasing HBM qualifications, we think leading memory vendors with strong HBM capabilities will likely benefit.

5. Chip design companies

Semiconductor stocks are generally volatile given the cyclical nature of the industry, versus the steadier growth witnessed in the software and internet industries. However, there is one segment within semiconductors that exhibits the software kind of steady growth and high recurring revenues: the electronic design automation (EDA) and intellectual property (IP) segment. Representing around 3% of the broader semiconductor sector's revenues, we see the chip design software industry growing by low-mid teens on average over the medium term and with high operating margins of 30–35%, on par with software. The EDA and IP industry is dominated by two companies, which provide computational software, custom hardware, and IP building blocks to chip designers and manufacturers on a subscription model. With rising chip design complexity and an increasing trend of custom chips by leading internet and AI platform companies, the industry is in an enviable position, in our view. While the industry is not immune from geopolitics, we think the subscription business model of the EDA and IP companies and structurally above-average growth and margins put them in a sweet spot, particularly for investors looking for defensive opportunities within semiconductors.

On top of strong AI infrastructure demand likely translating into solid spending on chip design software, we believe the industry is also at the forefront in terms of AI adoption. For example, mask synthesis software products have been used for computational lithography, which we think is one of the most compute-intensive workloads in the semiconductor manufacturing process, consuming tens of billions of hours per year for CPUs. In 1Q24, the leading GPU provider and software design companies announced a partnership to accelerate computation lithography. The new process uses only 350 H100 systems, which can replace 40,000 CPU systems and can accelerate production time and reduce costs, space, and power.

6. Original design manufacturers (ODMs)

ODMs play a critical role given the higher assembly and customization needs for new generation AI servers, providing strong tailwinds in the form of premium pricing. The revenue of an AI server rack can be 2–3x more than a regular server rack. It is estimated that the top four server ODMs will account for over three quarters of the total assembly market over the next few years, which doesn't come as a surprise given that they are the same players within the traditional compute space. This largely rewards leading Taiwanese ODMs, in our view, particularly those with deep-rooted partnerships across major cloud service providers (CSPs) globally. Recent guidance on the strength of CSPs' cloud capex should translate to more sustainable demand growth as we progress into the AI era.

The major ODMs are also responsible for most of GPU baseboard supplies globally. The GPU baseboard, a type of printed circuit board (PCB), is an essential component designed to house and provide interconnectivity among the multiple GPU modules. These baseboards have become increasingly complex given the rise in computation needs, as showcased by the rapid

growth of GPUs and LLMs in recent quarters. The new generation AI GPU baseboard form factor contributes to higher AI training and inference throughput, and we expect existing players to remain dominant due to the decade-long collaboration with both CSPs and GPU designers.

7. Power supply

Rising computation needs result in increasing demand for more efficient, stable, and higher density power sources. Given the recent growth in new power-hungry technologies, electricity supply and related solutions are gradually becoming the bottleneck for AI infrastructure growth. There are only a handful of leading AI server power suppliers in Asia, and we believe high barriers to entry provide a solid growth backdrop for these players.

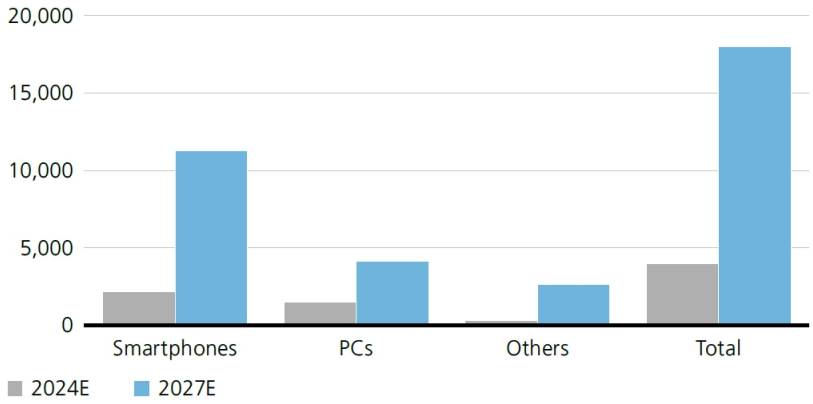
In parallel, optimal heat dissipation is necessary to sustain the high power consumption and to maintain AI server performance while preventing hardware failure. Average selling prices for thermal solutions for AI have doubled, particularly for air cooling, relative to traditional servers, given the required upgrades in components needed to handle the heat load. Air cooling should remain the dominant solution, as a shift to liquid cooling would require a substantial shift in the current data center infrastructure design. However, as rack densities transition beyond 70 kilowatts (kW), liquid cooling (which is a direct-to-chip cooling) becomes the only viable solution to sustain the heat generated from AI workload. In 1Q24, a leading GPU maker designed the AI server infrastructure specifically catered for liquid cooling solutions. We likewise expect the industry to gradually transition toward liquid cooling.

8. The AI edge opportunity

Early this year, Samsung unveiled its Galaxy S24 products integrating AI into its smartphones, which should herald a new wave of consumer devices integrating AI, including smartphones, PCs, and other household appliances and Internet of Things (IoT) devices. The ability to process data locally without being exposed to external data security risks could push many consumer electronics companies into exploring opportunities like edge computing, which can provide low latency and personalized generative AI services. For instance, some basic image generation and translation services may not need a model with trillions of parameters and may require training only a few billion parameters.

With AI democratizing, we should see the proliferation of such AI edge-computing devices. The potential addressable market from such a trend is significant, as it includes opportunities in peripherals and memory, as edge devices need more memory and storage. Hence, we believe AI edge computing offers promising growth opportunities. Accordingly, we expect industry revenues to grow from USD 4bn in 2024 to USD 18bn in 2027. Our estimates may prove to be conservative if we see rapid adoption of AI consumer devices and higher AI edge-computing chip pricing. But even at our current expectations, the segment should offer a nice tailwind for the broader semiconductor foundry industry.

Figure 12 - AI-edge computing should emerge as one of the fastest-growing segments within semiconductors
Revenues in USD million



Source: Company reports, UBS estimates, May 2024

Monetizing and capturing value in the intelligence layer

1. **Large language models**

The ability of a company to capture the value it creates is always dependent upon the quality and sustainability of its competitive advantage. Academic research is the canonical example. For instance, the University of Chicago Booth created substantial value for savers with the introduction of the efficient market hypothesis but captured very limited upside for itself. We can clearly envision the value created by LLMs; we can point to specific examples already such as the Klarna customer support case. However, the amount captured will depend on the long-term competitive advantages of respective foundational models. On this point, it's much harder to speak with clarity.

There are two main categories of pricing models: usage based and subscription plans where users pay a fixed monthly or annual fee to access the model with certain usage limits. Subscriptions can offer cost predictability and may come with tiered pricing based on usage levels. For example, Anthropic differentiates between three models: Opus offers top level performance, intelligence, fluency, and understanding; Sonnet offers maximum utility at a lower price and balanced for scaled deployments; and Haiku offers quick and targeted performance.

Monetization will depend on a lot of parameters, such as whether the queries are submitted via APIs (application programming interfaces; low margin) or interactively through a prompt with a subscription model (likely high margin), the size of the model (see pricing differences for Anthropic Opus vs. Haiku – 100x), and the nature of the LLM market (competitive vs. oligopolistic). In a steady state with a few dominant commercial models in each region of the world, LLMs could earn 40–50% operating margins, much like large software companies.

Figure 13 - LLM pricing examples

	Description	Cost (Input / Output per MTok^)	Subscription tier
Anthropic			
Claude 3 Haiku	Fastest and most compact model for near-instant responsiveness. Quick, accurate, targeted performance.	$0.25 / $1.25	Enterprise
Claude 3 Sonnet	Fastest and most compact model for near-instant responsiveness. Quick, accurate, targeted performance.	$3 / $15	Enterprise
Claude 3 Opus	Most powerful model for highly complex tasks. Top-level performance, intelligence, fluency, and understanding.	$15 / $75	Enterprise
Google			
Gemini Pro 1.0	Best performing model with features for a wide variety of text and image reasoning tasks.	$0.50 / $ 1.50	Pay-as-you-go
Gemini Pro 1.5 *	Best model for scaling across a wide range of tasks. Breakthrough experimental feature in long-context understanding. Highly sophisticated reasoning tasks for different modalities.	$7 / $21	Pay-as-you-go

*Gemini 1.5 Pro currently in preview mode, 'Pay-as-you-go' for both 1.0 and 1.5 listed as 'coming soon.'

Source: Company websites, UBS, May 2024. MTok- Millions of tokens

Note: This is for educational purposes only. For investment ideas and our stock picks, see our companion report "Investing through the AI platform shift."

Why they could capture value

First, it makes little sense for enterprises to build their own LLMs, as the technical know-how and cost to train are both strong barriers. Moreover, the useful life of any pre-trained model is quite short given the constant release of improved closed- and open-sourced models. Therefore, we expect enterprises to leverage and build on top of existing LLMs. Furthermore, given the number of leading foundational models (e.g., OpenAI's ChatGPT, Google's Gemini, Anthropic's Claude, Meta's Llama, XAI's Grok, Mistral's models and Open Source), it makes little sense for incremental startups to build models. Instead, we expect most of the incremental competition to be in the enabling or applications layer.

Training a GPT today depends on three core components: the amount of compute resources available, the scale and quality of the dataset, and the underlying model architecture algorithms. Since 2017, underlying model architectures have remained eerily consistent. As a result, GPT model improvements have come from scaling compute or data. This is why we call transformer models capitalist AI and why AI is not cheap.

The scale required to be successful is both large and growing fast. We note OpenAI's estimates that the amount of compute used in the largest AI trainings doubled approximately every 3.4 months, leading to a 300,000x increase between 2012 and 2018. This is an extremely high sunk cost and does provide some defensibility to LLMs, but we note that Meta's commitment to open-source and more efficient model architectures—such as the ones pioneered by Mistral—can lessen the scale barrier for end users.

LLMs also benefit from scaled use where live usage improves performance. Moreover, OpenAI's leading model, now GPT-4, has consistently ranked

ahead of peers in terms of performance for the last 18 months despite the mass acceleration of spend and focus by competitors. Thus, being the first mover does entail certain advantages. But the ability to create switching costs depends on how big the differences in model quality are from the proliferated use and head start.

Differences in model features, such as the maximum context window length, could create some defensibility. But we think it's too early to take a view on what the important key features are and if any of these will be unique to specific players. We do believe differences in model reasoning or embedded memory can drive more differentiation. However, for now, it seems that model choice is a function of application needed, cost optimization, and which hyperscaler most of the user's data sits in.

Data plays an important role, as it serves as competitive differentiation. For instance, it is unlikely that "off-the-shelf" foundational models with no fine-tuning will provide defensible structural competitive advantages, as these low-hanging fruits will be available to most participants. Therefore, to capture the value of AI, businesses will have to invest in their data usability and security. These investments involve curation to produce high-quality, labeled, and large training examples of specific applications and securitizing data through tools such as vector databases. The more differentiated and proprietary the data the more value will accrue to the owners of these data assets.

Prediction: Data assets will emerge as the competitive differentiators for AI adopters

In the longer term, the advent of artificial general intelligence—which we define as AI that is smarter than the collective of human intelligence—could prove defensive for the model layer. If we have "one model to rule them all," as Sam Altman believes will be the case, we should expect significant value to accrue to this layer of the stack.

Prediction: The application and intelligence layers will merge with AGI

Why they may not capture value

The argument as to why LLMs could commoditize is underpinned by very simple reasoning: We expect strong unit economics for LLMs given the level of competition from both closed and open-source proprietary models. For instance, Meta CEO Mark Zuckerberg recently stated that the company's "long-term vision is to build general intelligence, open source it responsibly, and make it widely available so everyone can benefit." At the start of the year, Mark Zuckerberg also disclosed that Meta plans to acquire over 350,000 H100s in 2024 and expects to have 600,000 H100 equivalent compute available generally. That's enough internal compute to compete with the leading proprietary models. Meta's most recent model, Llama 3, is comparable to the leading proprietary foundational models in terms

of performance and is available freely. While Llama 3 probably lags the leading models in terms of context window size and multi-modality, we don't consider the gap insurmountable.

Furthermore, Stanford University estimates that in 2023 alone, 149 new foundational models were released—more than the previous five years combined. Of these newly released models, 65.7% were open source—versus only 44.4% in 2022 and 33.3% in 2021. We therefore expect substantial competition for LLMs at varying model sizes. Enterprises looking to implement genAI may leverage many models throughout their IT stack, rather than one centralized model that rules them all. We suspect that customer lock in and platform entrenchment may arise from the application layer instead of the intelligence layer.

The vast majority of LLMs are still privately held or a subsidiary in a larger public enterprise (e.g., Google's DeepMind). So we should take the above information with a pinch of salt, as revenue estimates and developments are fast changing. Nonetheless, valuations remain lofty and given the lack of clarity on LLM monetization over time, we find it difficult to conclude that investing in an LLM today is an attractive proposition for anyone other than the cloud computing providers.

Why is it different for the hyperscalers, and what are they doing?

Hyperscalers are synergistic buyers of LLM companies because partial or full acquisitions of leading models typically come with exclusivity agreements with the respective public cloud instances. For example, since investing an aggregate USD 13bn in OpenAI for a 49% stake, Microsoft Azure has become the exclusive provider of cloud computing for OpenAI's research, products, and developer interfaces. Microsoft has made further investments into LLMs with its partial investment in Mistral and full acquisition of Inflection AI.

Google, meanwhile, continues to be mostly focused on internal model development, given its early investments in AI. Google's acquisition of DeepMind in 2014 for USD 400mn ranks alongside its acquisitions of YouTube and Android as the firm's most prescient, in our view.

Finally, Amazon has taken a mostly neutral distributor stance with Amazon Bedrock. Amazon Bedrock is a full management service that allows companies to build generative AI applications by providing access to a variety of foundational models. Until recently, Amazon had not invested directly in a model provider. However, that changed with its September 2023 initial and March 2024 follow-up investment into Anthropic, bringing its total to USD 4bn.

A recent trend in genAI has been the advancement of sovereign AI. Governments, alongside corporations, have woken up to the importance and impact of genAI. Most governments will likely want to develop their own foundational LLM to ensure that it matches the region's values, ethics, data privacy laws, and security requirements; an example is the UAE's Falcon Model.

Monetizing and capturing value in the application layer

1. Copilots

Copilots are at the center of AI applications due to the strong role they can play in boosting office productivity. For instance, at its 4Q23 earnings call, Microsoft reported developers' productivity can rise by around 50% when using its copilot. The term "copilots" refers to AI companion tools that are integrated within office workflow/productivity software to boost employee productivity. For other software companies, AI is also providing a catalyst to raise product pricing—for example, Adobe raised prices on its Firefly service around 10% and ServiceNow increased prices by over 50% for its Vancouver Platform. As a group, we expect copilots and software assistants to reach USD 75bn in revenues in 2027, which at only 7–8% of global software industry revenues is very conservative, in our view. This is because we believe faster-than-expected productivity gains should drive significant upside.

To understand the AI upside case, it is imperative to first appreciate the size of the addressable market. According to World Bank data, the global working population is close to 3.5 billion currently. And as mentioned, there are around 1 billion knowledge workers. So we assume that 2.5 billion workers are frontline employees (who may not need access to premium office features). As for the remaining 1 billion knowledge workers, we assume 50% work in large organizations and the rest in small and medium-sized enterprises (SMEs).

While not all of the working population use office commercial products, we can calculate the potential addressable market for this segment using the above assumptions. For the frontline employees, we assume a low-tier level pricing of around USD 48/year (assumptions are based on Microsoft's pricing). This results in a USD 120bn potential market. For the 500 million SME employees, we assume a mid-tier level pricing of around USD 120/year, which results in a USD 60bn potential market. For the 500 million employees in large enterprises, we assume a high-tier level pricing of around USD 456/year, which results in a USD 228bn potential market. In summary, assuming every employee subscribes to office commercial products, and using the above assumptions, the potential addressable market should be around USD 408bn a year.

This implies a very low penetration rate of 11%; even if a USD 65bn estimate is used over the next three years, the implied penetration rates would stand at only 16%. In a scenario where strong AI features lead to swift uptake of office productivity software in emerging markets, we could see significant upside to these penetration levels. Instead, if the segment manages to grow at 20% each year (our upside case for copilot growth), its revenues could reach USD 80bn, implying around 5% upside to street revenue forecasts—which expect growth rates to decelerate to 10% y/y in FY2027—and mid-high single-digit upside to operating profits—considering office products are higher-margin products. With a typical copilot product priced at USD 20–30 per month, these estimates could prove to be even more conservative if there is faster-than-expected adoption. For instance, we estimate a mere 20% adoption of AI copilots by current office users would present an upside of around 10% to overall revenue forecasts.

In summary, we believe the street's estimates on the office commercial segment are very conservative. Over the next three years, we think that AI will drive an acceleration in office commercial revenues and not a deceleration to 10% growth.

Workflow management software is another segment we favor, and we believe the leader in this space is well positioned. Generative AI is able to speed up issue resolution by analyzing alerts and providing critical context for operators or increase case deflection from live agents with a conversational and self-service experience. A leading company claims that there are more than 20 use cases internally, which are generating millions of dollars in cost savings per year and are helping service agents close incidents in half the time.

2. Advertising

Unlike the semiconductor and software segments, which have clear near-term visibility from generative AI, the internet advertising segment doesn't have the same first-mover advantage. However, in the medium to long term, AI should help grow revenues: With a conservative estimated penetrate rate of 13–15% of global advertising industry revenues, we expect the internet AI segment to report USD 100bn in revenues in 2027. While the AI monetization trends for the internet industry will likely evolve, we see four broad ways that internet companies can monetize AI.

The first way is via content creation, as we believe generative AI can help create new content across texts, images, videos, and other multimedia formats that can maximize revenues for internet companies. Early trends already indicate how AI-generated articles are being used by media companies and how AI-created images and videos are being used to display advertisements. Second, chatbots, which can provide nice subscription revenue streams, can be used to improve customer service and as a personal buddy/assistant (Character AI is one such popular paid chatbot today in the market). The third way is via personalized content, which can be used by streaming services as well as advertising companies to increase user engagement. And finally, predictive and other analytics can be used by digital media and e-commerce companies to roll out new products and services.

In summary, the integration of generative AI has just begun, and with promising new generative AI applications expected in 2H24, visibility on the internet industry's ability to monetize should gradually improve. This in turn should also lead to a gradual rerating of the industry.

The impact of AI on our longer-term thematic offering

Authors: Alexander Stiehler, Michelle Laliberte (US strategist)

Artificial intelligence will impact CIO's Longer-term Investment (LTI) themes in a variety of ways, creating risks and opportunities alike. To simplify, we've grouped our 29 LTIs into six major categories with similar investment drivers and based on where the opportunities overlap. We provide more context on AI's impact for each of the six categories herein.

Disruptive technology

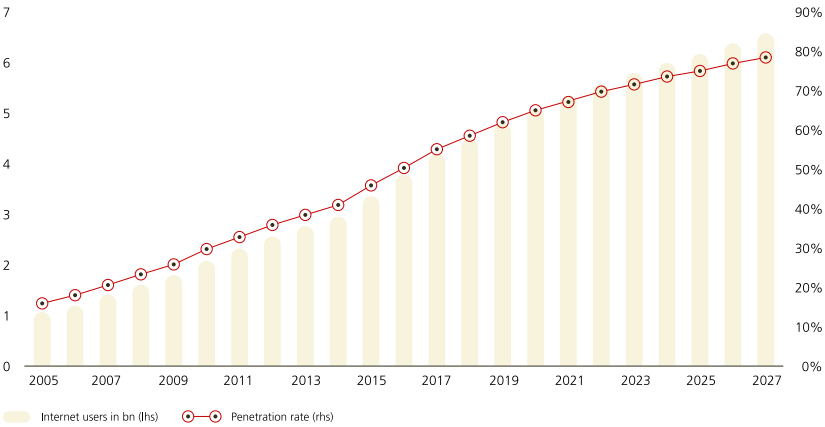
Our tech-related LTI themes are at the heart of the AI revolution. They cover the full value chain, from AI chips in our *Enabling technologies* theme and AI software in our *Digital data* theme to edge AI devices in our *Metaverse* theme and AI applications that could accelerate the smart manufacturing transition in our *Automation and robotics* theme.

The investment opportunity is substantial, in our view. For the *Enabling technologies* theme alone, our analysts expect the market in the five enabling technologies—artificial intelligence, augmented reality/virtual reality, big data, 5G, and other moonshot technologies—to grow in aggregate by 26% on average annually between 2022 and 2027, reaching USD 1tr in 2027. An important input factor for AI is data, which we consider in our *Digital data* theme. Thanks to rapid urbanization in emerging markets, we expect the global internet user base to increase by 2 billion between 2020 and 2030, and internet penetration to approach 80% worldwide (see Fig. 14).

This trend, the proliferation of connected devices, and solid enterprise data trends should lead to an exponential increase in consumer data. Generative AI has further contributed to exponential data growth with the massive creation of synthetic data, which should become the bulk of the future data universe. OpenAI's text-to-video AI model—Sora—utilizes large amounts of synthetic data for the creation of videos. This is just one example of how synthetic data may reshape many aspects of our life in the future.

Figure 14 - Global internet penetration expected to approach 80% by 2027 with almost 7bn users

Strong growth in internet users to fuel further growth in smartphone and technology penetration globally



Source: World Bank, UBS estimates, as of 2024

Energy transition

The energy transition is one of the most important and urgent topics facing humanity. The negative consequences of climate change are visible across the globe. Artificial intelligence is likely to play a role both in terms of solutions and by adding to the challenge of meeting energy demand. We expect AI

to play an important role in mitigating climate change across our thematic offering, but it's also likely to push electricity demand estimates higher on the back of the data center build out.

Longer-term themes such as *Clean air and carbon reduction* and *Energy efficiency* are at the forefront of this transformation. The application of AI tools in smart electricity grids to better forecast energy demand is one example of a potential positive use case. Another is the ability to improve efficiency via building automation. Today, heating and cooling are a relatively static process. AI could make it more intelligent by adjusting for weather forecasts or room occupancy; the latter has become even more important since the COVID-19 pandemic, as more people are working from home. Smarter building automation could also lead to reduced operating costs and better utilization of buildings, providing firms incentives to invest in these solutions. Moreover, the likely swift expansion in AI datacenters, which in some cases consume three to eight times more electricity than traditional datacenters, make energy efficiency considerations even more important.

The electric vehicle (EV) revolution is another area that is supporting the transition from fossil fuels to more environmentally friendly solutions. We discuss this topic in our *Smart mobility* theme. In our view, AI will accelerate developments in autonomous driving and improve the power energy management in EVs. AI models can accurately calculate and predict the vehicle range and provide updates about the health of the battery to expand the lifespan, to name just a few examples.

In sum, we believe that the energy transition offers investment opportunities for many decades, providing investors attractive structural growth opportunities. However, the rapid deployment of AI is likely to push electricity demand higher, exacerbating the challenge of transitioning away from fossil fuels while still ensuring access to reliable and affordable energy supply.

Natural resources

Water scarcity, polluted oceans, and food manufacturers' need to ensure food security while mitigating climate change are just a few examples of natural resource challenges that should ultimately benefit solutions providers. For all these topics, AI should play an important role in dealing with these challenges. In the food sector, AI could be used to develop autonomous tractors with state-of-the-art sensors and vision technology to better distribute fertilizer or to spray herbicides in a targeted way. This is an area we discuss in our *Agricultural yield* theme.

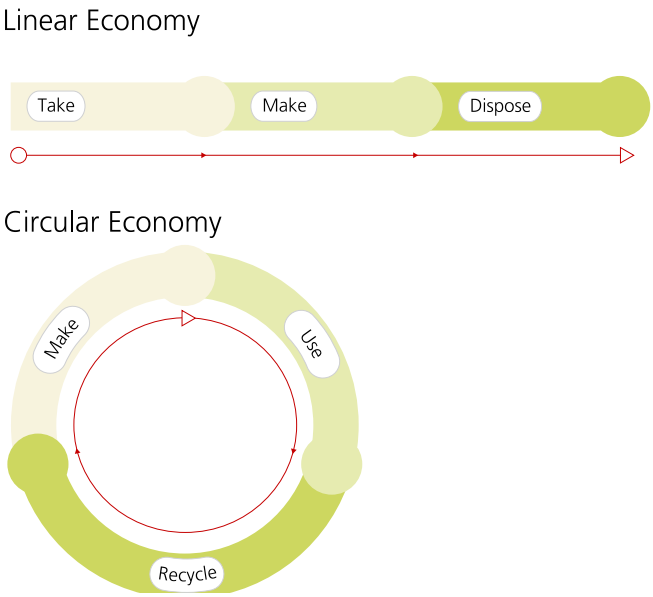
The water sector is suffering from water leakage and inefficient water infrastructures. Here, AI and the application of sensors could help make water infrastructure smarter, detecting leakages in seconds. It could also help in the analysis of water quality or tracking of water use. Irrigation strategies can be improved by AI applications too, as another example.

At present, much of the world is still living in a linear economy, where goods are bought, consumed, and discarded (see Fig. 15 for the difference between a linear and circular economy). To achieve a *Circular economy*, AI can help accelerate the design and prototyping of new products and materials that can be reused in a circular economy. AI is already being used to better forecast

demand and maintenance, using historical and real-time data to avoid waste and improve product utilization.

As another thematic opportunity within natural resources, AI can offer new applications in *The blue economy*. AI-driven robots, for instance, could help analyze marine life and help manage fish farms. Today, robots already feed up to 200,000 salmon per pen in Norway. With AI projections, the output can be optimized. These are a few examples of how themes in our natural resources basket may benefit from the implementation of AI.

Figure 15 - Illustrative comparison of a linear vs. circular economy



Source: UBS

Health and longevity

Artificial intelligence is likely to be used extensively within the healthcare industry, which is currently one of the least digitized industries yet produces a vast amount of data—a combination that makes the sector ripe for AI-driven efficiency improvements. If AI can improve and speed up drug discovery or improve the likelihood of FDA approval success, this could help support long-term themes such as *Genetic therapies,* where much of the industry is still dependent on early-stage trials that involve a high risk of failure. *Oncology* is similar but to a lesser extent, given some of the largest pharmaceutical companies in the world already have products in the market.

In addition, *Healthtech* solutions typically rely on leveraging data and technology, so AI will likely be used to improve the management of those systems and processes. We're also seeing AI tools used in diagnosis and image processing. So far, this has been done alongside a human, which we expect to continue. Evidence has been encouraging for early detection of diseases such as breast cancer and melanoma. Outside of imaging diagnostics, AI's power lies in its ability to save time and analyze vast amounts of data quickly. This could be used to analyze medical history or provide real-time data to

detect or predict future health risks, optimizing preventative care treatment and potentially reducing larger costs down the road.

Medical devices is another healthcare theme where AI and machine learning are already being used, with the FDA seeing an increasing number of applications for AI/ML-enabled devices over the last couple of years. As we note in the industry overview, the revenue benefit for medtech companies integrating AI into their product offering remains unclear so far.

Human capital and consumption

When considering artificial intelligence within the context of human capital and consumption, we see the *Education services* and *E-commerce* themes as most immediately relevant. AI will likely shift demand for certain skillsets and further emphasize the need for lifelong learning and worker upskilling and retraining. For ed-tech providers, AI presents a paradox. It can be a tool for those that properly utilize it and integrate it into product offerings, or it can be a significant threat for those that fall behind, eventually becoming "table stakes" for education companies. AI has several use cases in the education space. For example, it can enable personalized learning for students by tailoring educational experiences according to an individual's learning objectives and needs. AI-driven data analysis can also provide real-time feedback. We discuss the risks and opportunities further in our latest report on *Education services*.

In *E-commerce*, data-driven targeted advertisements have proliferated in recent years and will likely be enhanced even further by artificial intelligence and machine learning. In addition, more websites could start using AI-powered chatbots for customer service applications. There are also some takeaways for the *Consumer experience* theme, as AI will likely be leveraged to enhance the consumer experience in a variety of ways such as premium bespoke product offerings or providing more supply chain transparency—all discussed more within the consumer sector takeaways.

Security and critical infrastructure

AI should fundamentally change security systems and further emphasize the need for cybersecurity, a key pillar of the opportunity in our *Safety & security* longer-term theme. Cyber vulnerabilities in critical infrastructure systems are a key focus for the US and other governments, whose agencies are integrating artificial intelligence into cyber defense systems. AI can be used to identify vulnerabilities, detect anomalies, and aid in early detection of potential threats. Cybersecurity investment is critical for companies as well, with the cost of cyber breaches significant and rising. In the latest IBM Security and Ponemon Institute cyber breach report, firms that systematically use AI as a tool for cybersecurity reported that they had on average a 108-day shorter time to identify and contain the breach and USD 1.76mn lower data breach costs compared to those that don't use it (39% lower cost on average).

The *Space economy* also looks ripe for integrating artificial intelligence, thanks to the vast amount of data collected via satellite. AI can quickly analyze data for a variety of applications, including earth observation and natural disaster responses from flood prediction to landslide detection. AI is becoming increasingly vital in forest fire prevention and management.

Drones equipped with AI technology monitor forests for ignition threats, detect wildfires early, and gather valuable insights on fire intensity. Thermal drones can even locate trapped individuals and provide real-time data to help search and rescue strategize and prioritize response.

Most major space agencies, including NASA, have set up groups focused on artificial intelligence and the exploration of its use in space-based applications. For one example, in a coordinated effort between NASA and Google, AI algorithms were used to identify two new exoplanets previously missed by humans. We note that artificial intelligence can be used by malicious actors too and is highly likely to be used in more controversial end-use applications related to defense and space-based defense applications. We think this creates an incentive for governments to invest not only in their AI and technological capabilities, but the related space-based infrastructure too. The defense industry's use of AI is discussed in more detail within the sector takeaways chapter.

Positioning for AI in sustainable investments

Authors: Catherina Campedelli, Amanda Gu, Antonia Sariyska, Merilin Tedeschini

Enabler of environmental and social solutions

Thanks to its ability to treat large amounts of structured and unstructured data (text, audio, images, videos) and to make predictions or recommendations, AI models present significant opportunities for the development of products and services that tackle key sustainability challenges. Given increasing consumer demand for more sustainable solutions, coupled with the need for climate change mitigation and adaptation, food security, and access to quality healthcare and education services, we think allocation to sustainable thematic investments within equities and private markets could future-proof and benefit portfolios.

Many of these opportunities build on our previous discussion around implications for longer-term investment themes.

For example, the adoption of AI can play a critical role in reducing greenhouse gas emissions by increasing the efficiency of energy management systems and transforming the way power is generated, distributed, and consumed. AI-driven solutions are already used in renewable energy forecasting for balancing supply and demand and in smart homes, where energy is dynamically managed. They also find extensive applications in emissions measurement and transparency, often decreasing the cost of measurement while improving its accuracy. Furthermore, AI technologies are increasingly being integrated into forecasting for (extreme) weather events, such as flooding and wildfires, which can help improve climate adaptation and reduce physical climate risk to businesses, households, and insurers.

AI is also revolutionizing waste and water management and pollution control. Paired with robotics, specialized companies integrating AI into their service delivery can benefit, among others, from higher precision in sorting and

recycling, thanks to the ability of AI to analyze a vast amount of data and to detect a broad range of waste types. Application of AI in agriculture, jointly with the use of drones and satellite imagery, can help farmers explore soil health for an intelligent and selective use of fertilizers and pesticides and help prevent water waste through autonomous irrigation.

Within healthcare, AI can be used for more effective cancer detection, in mental health assessment, or to support people with disabilities via human augmentation. In education, AI can create adaptive, personalized, and inclusive learning models. Similar applications can be found in financial inclusion, where financial risk assessments can be more precise and tailored to underserved borrowers or entrepreneurs.

Within professional services, AI enables recruiting practices that help reduce bias by screening candidates fairly and objectively. In turn, this could increase innovation and productivity and promote diversity and inclusion in the workplace.

While we see a broad range of opportunities in the adoption of AI to develop sustainable products and services, unlocking its potential calls for addressing many challenges. For example, the use of AI to achieve energy efficiency requires a careful design of the AI algorithms with a particular focus on minimizing usage of computational resources and power. Also, for an effective use in recruiting practices, appropriate techniques and processes need to be put in place to tackle biases, which are often inherent to the input data.

Many of these AI applications find themselves in the early stages of development and often require patient financing through venture capital or private equity investments. Impact investing—a subset of sustainable investing, where the capital is used to drive additional and measurable environmental and social outcomes—has emerged as a preferred way for private market investors to embrace new sustainable enabling technologies and business models.

Enabler of sustainable supply chains

Corporate sustainable development and transparency stand as pivotal forces reshaping organizational strategies. As companies navigate the complexities of global markets, the pursuit of sustainable practices and transparent operations emerges as a strategic necessity, which can be further facilitated through the adoption of AI.

We think investors should consider environmental and social risks stemming from companies' operations and supply chains, even if their portfolios are not explicitly focused on sustainability. An assessment of AI adoption and implications is likely to factor into such considerations.

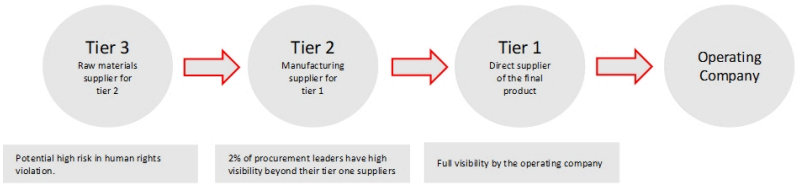
Supply chains exert a substantial environmental influence, encompassing the extraction of raw materials, transportation, and concluding with the disposal of end products. Moreover, issues such as child and forced labor as well as all kinds of human rights violations are gaining prominence in a business landscape increasingly focused on ethical practices. Consequently, regulatory bodies are acting, exemplified by the recent adoption of the EU Corporate

Sustainability Due Diligence Directive (CSDDD), which will require companies to conduct due diligence on sustainability matters, value chains included, as of 2027. More broadly, regulation has continued to drive the need for supply chain transparency and systematic assessments.

Traditional planning methods have proven inadequate for the complexities of the current world. AI contributes by improving planning, optimizing processes, and bridging gaps in supply chain management. It can predict and ensure efficient production flows, logistics optimization, and route planning —ultimately reducing the environmental impact. For instance, AI-based predictive analytics identifies optimal navigation routes considering weather, traffic, and fuel consumption, while also improving demand forecasting and inventory management, minimizing waste. Additionally, AI enables automation and predictive maintenance, reducing downtime and energy consumption.

In the evolving supply chain landscape, visibility and transparency beyond tier-one suppliers are increasingly vital. Buyers establish direct contractual relationships with their tier-one suppliers, excluding the numerous subcontracted service providers and lower-tier suppliers. This results in traceability and transparency challenges, constraining buyers' ability to enforce desired corporate codes within a loosely connected network. According to the 2023 Deloitte survey on Commercializing sustainable supply chains, only 2% of the procurement leaders state that they have "high visibility" beyond tier-one suppliers, 26% of procurement leaders face limitations in visibility beyond their tier-one supplier, and merely 6% claim full transparency across their entire supply chain. Without a clear vision on the value chain, organizations reduce their ability to effectively mitigate potential risks like forced labor, child labor, neglect of health and safety standards, and disparities in fair and equal treatment. The November 2023 allegations against Mars about employing child labor in its value chain shows the importance of having clear visibility beyond tier-one suppliers and raised greater concern about the importance of knowing its value chain.

Figure 16 - Beyond tier-1 suppliers



Source: UBS CIO, 2024

Historically, gaining visibility into sub-tier suppliers presented significant difficulties, requiring considerable time and labor resources. However, contemporary advancements in AI tools can help organizations gain greater visibility into their supply chains, offering real-time insights into suppliers' performance. AI can efficiently manage a comprehensive database of first-tier suppliers, enabling the creation of a tangible map of the supply chain. AI tools can integrate data from various sources, including suppliers at different tiers, and centralize this information into a unified database. This consolidation allows for a comprehensive view of the entire supplier network. Companies

equipped with a comprehensive understanding of their value chain are positioned to minimize the risk of environmental, social, and governance non-compliance and monitor social engagements with their suppliers.

AI can also help measure CO_2 emissions. Carbon accounting has always been a challenge, often showing many inaccuracies. The lack of data or unstructured datasets can be attributed to imprecise reporting and to the cost of measurement. However, carbon accounting has witnessed advancements through machine learning applications. Existing models provide estimated carbon data, but with the implementation of refined methodologies and smarter technologies, the potential for more accurate assessments is substantial. AI has the potential to analyze enormous data sets with precise results as output. By leveraging AI applications, organizations can enhance the precision of their carbon accounting processes, allowing for more nuanced understanding of environmental impact.

Enabler of impact measurement, tracing and attribution

The integration of AI offers more effective and sophisticated solutions for the measurement of environmental and social outcomes. This measurement, coupled with tracing and attribution methodologies, can help improve investment decision-making and corporate capital allocation decisions, potentially leading to better and more effective distribution of sustainable investing and financing.

In impact measurement, AI facilitates the analysis of multidimensional data sources, ranging from structured financial reports to unstructured social media feeds. Through natural language processing algorithms, AI systems can extract and contextualize qualitative information, enabling a more nuanced understanding of the social and environmental outcomes associated with investments. Additionally, machine learning models can trace patterns and correlations within the data, enabling investors and corporate decision-makers to see hidden relationships and gain deeper insights into the impact of their investments in real-time.

AI-powered tools can enhance impact measurement by providing more granular and context-specific metrics. While traditional metrics may fail to capture the full extent of impact—especially in complex social systems—AI algorithms, together with technologies like blockchain and Internet of Things, can analyze qualitative data such as text and images to extract valuable insights. Smart contracts deployed on blockchain networks ensure transparent record-keeping, while IoT applications can provide granular data on supply chain activities and environmental performance metrics. AI algorithms can then analyze this data to identify discrepancies or potential areas for optimization, thereby enhancing transparency and accountability.

Impact attribution is a key component of impact investing and often a hard one to measure. AI techniques, such as causal inference algorithms and counterfactual analysis, enable investors to identify the causal relationships between their investments and the observed impacts. By simulating alternative scenarios, AI-driven attribution models can isolate the true impact attributable to specific investment actions, informing decision-making and capital allocation.

Operationalizing AI: ethics and energy consumption

While we expect the deployment of AI to bring numerous benefits to productivity, resource efficiency, innovation, and sustainable development, adopters and investors should also consider potential adverse impacts that come with any technological development. In the case of AI, two critical issues continue to be flagged when it comes to operationalizing: ethics and energy consumption. We think investors should consider companies' governance mechanisms, ethical behavior and approach to managing greenhouse gas emissions as part of decisions to allocate capital to AI-related technologies and business models.

The question of AI ethics is in fact very real and present, with "deep fakes" distorting information, and the effect of human thinking on the development of the technology. In November 2021, the 193 members of the United Nations adopted UNESCO's "Recommendation on the Ethics of Artificial Intelligence" to summarize, standardize, and potentially manage the ethical framework around this technology. While mostly directed at policymakers, we think the framework represents a useful guide on assessing the potential risks of deploying AI technologies more broadly—both for companies and for investors. Ongoing monitoring and lifecycle management form an essential part of this process as well.

UNESCO's framework on assessing AI ethics

Values:

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

Principles:

- Proportionality and Do No Harm
- Safety and security
- Fairness and non-discrimination
- Sustainability
- Right to Privacy, and Data Protection
- Human oversight and determination
- Transparency and explainability
- Responsibility and accountability
- Awareness and literacy
- Multi-stakeholder and adaptive governance and collaboration

Source: UNESCO, 2021

For example, when thinking of the broader economy and labor, it might be essential to evaluate the change in skillset that is required to deploy AI, and to what extent this reflects the current labor mix. Failure to build essential skills might slow down the speed of development and adoption, while rapid skills development, on the other hand, might create an economic divide and labor shortages in other parts of the economy. To illustrate the potential scale, one study shows that large language models can already impact 80% of the US workforce and 10% of the work tasks.[17] The impact on small and medium-sized enterprises, which are typically the backbone of developed and developing economies, should also be considered, depending on the volume of resource and capital expenditure that AI deployment requires.

Another area of potential risk is formed by issues around fairness and non-discrimination, and whether AI can perpetuate biases, which—in addition to exacerbating societal challenges—might also lead to poor business outcomes, for example when serving clients or developing new products and services more broadly. As with every digital technology, safety and security are also key. The ethical dimension of this risk has to do with potential misuse of AI when hacked by ill-meaning groups or organizations, particularly in the field of terrorism and warfare.

These risks are difficult to assess at the individual company level, lending themselves to broader impact considerations across technological development as a whole. According to a 2021 Gartner study[18], 86% of corporate decision-makers in their sample agree that businesses are not taking ethical risks seriously when deploying AI, with 31% stating that the chief information officer should be ultimately responsible for the ethical management of the technology, begging the question of whether companies are truly equipped with the right governance frameworks, leadership, and skillset to effectively tackle this issue. Therefore, we think investors should pay close attention to governance mechanisms, the history of ethical corporate behavior, and companies' approach to selecting and targeting their consumer base.

The issues around energy consumption—namely, the carbon footprint of AI—seem better defined, with clear mechanisms for management, in our view. The International Energy Agency estimates that data centers and data transmissions account for 1–1.5% of total energy-related emissions[19], and while absolute emissions linked to information and communication technologies continue to grow, driven by computing power and cooling, this growth is quite modest compared to the overall deployment rates of such technologies. The IEA also points out that new generations of existing technologies, such as internet broadband, mobile networks, and data centers are becoming more resource efficient. Still, optimizing energy consumption and ultimately reducing absolute greenhouse gas emissions remain critical as the world aims to curb global warming. The IEA analysis highlights that the electricity demand for an OpenAI search request is almost 10 times higher than for a Google search request. Many regions and economic areas, especially across emerging and frontier markets, still rely on legacy technologies and networks, making the transition to "greener" AI infrastructure more challenging. In this regard, investors should pay attention to companies' and technologies' overall development in carbon intensity—i.e., the additional carbon emissions generated as revenues and deployment grow—to be able to assess the energy and resource efficiency of deployment. We think climate transition risk is likely to be material to all industries, including IT, and have an effect on larger AI adopters within the energy, industrials, and materials sectors.

State of the private AI market

Author: Karim Cherif

Private investors (PE) and venture capitalists (VC) in particular often get a sneak peek of the future. The same holds true for the AI opportunity. First investments by private equity and VC funds in AI/ML firms date back almost 15 years ago. Since then, the investment pace has gradually accelerated with almost a quarter trillion dollars invested in the past five years.

Last year saw global investments in AI/ML companies reach USD 43bn across 2,500 deals, of which 32% can be attributed to VC funds, 21% to corporates and corporate venture capital funds, and 6% to private equity firms, according to CB Insights. Overall activity slowed down year-over-year and was below 2021 records, a consequence of broader macroeconomic uncertainty and elevated interest rates. Compared to the rest of the VC market, however, AI startups showed resilience in terms of funding, valuations, and exits.

The AI industry is maturing, and deals are getting bigger. Average deal size reached USD 23.5mn in 2023 from USD 15.2mn five years ago. At the end of 2023, there were 186 private AI companies valued at over USD 1bn (unicorn), up 10% y/y. Regionally, the US is racing ahead, capturing 42% of total AI/ML deals, followed by Asia (30%) and Europe (23%).[20]

Investment approaches, however, differ depending on the investor base. Earlier-stage investors such as venture capitalists and growth equity managers are focusing on technology investments in areas such as natural language interfaces (e.g., chatbots, personal assistants), AI core platforms (for data training, model development, deployment, and management), two-dimensional digital media content (images, video, avatars), and biotechnology. Larger private equity firms, meanwhile, are focusing on building the physical infrastructure that will be key to the AI revolution, datacenters in particular. According to Synergy Research Group, private equity share in data center M&A activity steadily grew to reach 91% of overall deals in 2022. Other large PE firms are partnering with AI startups to adopt new solutions to drive productivity and efficiency and generate additional sales—notably in the fields of software, finance, media, healthcare, and cybersecurity, with AI increasingly seen as a new lever of long-term value creation.

More broadly, general partners are laser focused on understanding the implications of generative AI advancements on existing portfolio investments and on identifying threats and opportunities to existing economic business models. AI technologies also hold the promise of revolutionizing the way private equity firms operate. From identifying future targets to conducting due diligence or automating back-office workflows, AI has the potential to dramatically improve data collection and sharing in an industry renowned for its lack of information transparency and efficiency.

We think PE and VC managers are uniquely positioned to offer investors a holistic exposure to AI opportunities across various verticals at different stages of a company life cycle. Most promising opportunities currently lie in the fields of datacenters, software, media, and biotechnology, in our view.

We note, however, that many of these opportunities have yet to prove commercial viability and are mainly accessible through VC investments, which carry several risks. Besides the inherently higher failure rate of startup companies, VC investments differ significantly from investing in listed shares

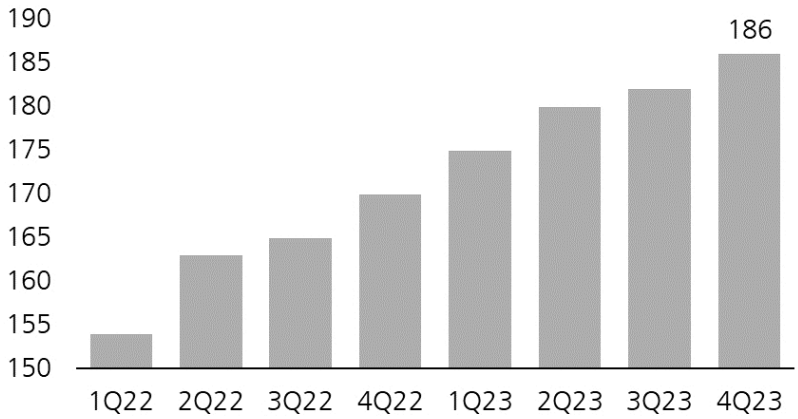
on the stock exchange. Venture capital is essentially an opaque and illiquid market, where information is not equally available to everyone. Financing usually consists of multiple rounds of private financing by investors who purchase newly issued, unlisted securities. Each round triggers a capital increase and raises the risk of ownership dilution. Also, shareholder rights such as voting, veto, exit, or liquidation rights can differ greatly between founder-entrepreneurs, angel investors, independent or corporate venture funds, and other industrial or financial groups.

Securing access to the best fund managers to mitigate these risks is paramount to maximizing the chances of success. Competition among VC managers to fund the best companies is also very high. Investors should seek partnerships with managers who are actively sourcing deals and taking a leading role in the company in which they invest, as opposed to employing a follower strategy and focusing solely on unicorns and bigger deals. Manager selection matters, but so does portfolio construction. Investing in one VC fund has historically proved to be a rather inefficient way to access the asset class. Investors should commit to a long-term plan and build exposure across vintage years, geographies, managers, etc. Importantly, VC exposure should be considered within a global private market portfolio diversified across various private equity and debt strategies and sized according to an investor's risk appetite and goals.

AI unicorns

While we continue to see multiple trillion-dollar AI opportunities within public listed companies, we also highlight another trillion-dollar AI opportunity in the unlisted space—in AI unicorns and other startups. While there are more than 1,000 unicorns globally, according to data from CB Insights, there were 186 AI unicorns by end-2023, with 16 new unicorns added in 2023. Currently, AI represents around 15% of overall global unicorns; considering our view that AI is the tech theme of the decade, we expect AI's share of global unicorns to continue to rise, as we believe unlisted companies today offer a better way to participate in breakthrough AI innovation and strong long-term growth potential than through public equities. Within the tech space, we believe some of the most exciting AI opportunities in the unlisted sphere are mostly in AI platforms, software and infrastructure, and data center companies.

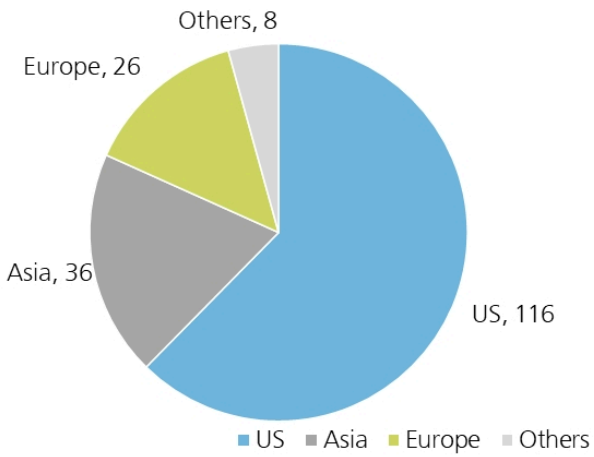
Figure 17 - AI unicorns globally on a steady rise
Number of AI unicorns (private companies with USD >1bn valuation)



Source: CB Insights (cbinsights/research), UBS, as of February 2024

Figure 18 shows the list of AI unicorns by region in 2023. With a significant first-mover advantage, including strong leadership in LLMs and AI computing, the US market is in the driver's seat, accounting for 116 AI unicorns in 2023 and 62% of global AI unicorns. Coincidentally, the US market today accounts for 63% of global equities (as represented by MSCI AC World). Asia has the second-largest number of unicorns with 36, followed by Europe at 26 and the rest of the world at eight. While broadening AI trends mean we should expect more unicorns outside the US, we still expect the US market to remain at the center of AI innovation in both the public and private spaces for the foreseeable future.

Figure 18 - The US dominates with more than 100 AI unicorns



Source: CB Insights (cbinsights/research), UBS, as of February 2024

Economic value of AI, an investor's perspective

Author: Dean A. Turner

The impact of AI on the global economy will in many ways be similar to that seen with the introduction of new technologies during the first three industrial revolutions. In our view, given the numerous and as yet unknown and undetermined paths the development of AI is likely to take, it is less important for investors to attempt to quantify an overall impact and more beneficial to consider how AI will benefit or impair various sectors and regions of the global economy.

New technologies are only adopted if they can deliver economic efficiency. Historically, greater efficiency has resulted in productivity gains that have fueled economic growth. AI has the potential to boost productivity in many sectors of the economy, by enabling workers to produce more with the same or fewer inputs. However, unlike previous technological developments, AI is happening against a backdrop of shrinking workforces in countries that currently make up over half of the world's GDP. Moreover, increasing resource constraints and environmental concerns may mean that productivity gains are merely used to stabilize output with fewer inputs. Thus, it may be that the main influence of AI will be to slow the pace of declining growth that has affected many economies over the past decade, not reverse it.

Economic growth isn't the only factor for investors to consider. Corporate pricing power is also likely to change if, as we expect, firms use AI to introduce greater price discrimination. This would likely come from new technologies giving firms a greater understanding of the price every customer is willing and able to pay for a good or service. Of course, this could mean some companies can charge more, potentially boosting margins, but there is also the potential that AI only serves to lower barriers to entry for new competitors, which could erode margins for established industry leaders.

With regard to employment, the IMF estimates that around 40% of the global workforce will see their jobs impacted by AI. For some, this will result in changes to established working practices alongside the potential productivity gains. But it is also estimated that roughly half of those jobs affected will be replaced by AI entirely. Fears of mass unemployment are not new when it comes to new technologies such as AI, but history shows that the labor force adjusts to accommodate such changes, with new roles and industries replacing old ones. However, this takes time, and one concern with AI is that the speed of adoption of the new technologies will mean that workers cannot retrain quick enough, resulting in a large and prolonged bout of frictional unemployment. This can lead to negative political outcomes, as disaffected workers seek to apportion blame, causing greater political polarization. It could also mean that politicians are more willing to pursue policies that prevent the adoption of AI. A more volatile political backdrop combined with bad policymaking can often lead to less positive outcomes for investors.

A final consideration is the regional impact of AI. All economies should benefit from new technologies that promote economic efficiency compared to the alternative. However, not all will benefit equally or at the same speed. Regional differences in regulation, availability of capital, levels of education and training, and protectionist policies will likely mean that some regions will enjoy the benefits sooner and to a greater extent than others.

Chapter 2: Repeating and rhyming – why AI is similar to and different from prior IT cycles

American writer Mark Twain is often credited with saying that "history doesn't repeat itself, but it often rhymes."

In 1996, Microsoft co-founder and former CEO Bill Gates said "…we always overestimate the change that will occur in the next two years and underestimate the change that will occur in the next ten…"

We think both statements may be helpful in thinking about the path to monetization for generative AI and comparing the current environment to the dotcom era of 1997–2000, the prior great technology boom.

With the benefit of hindsight, we see three distinct phases of dotcom-era investment that may be helpful in thinking about the path to AI monetization.

In the first phase, there is significant investment in the underlying infrastructure and an exploration of the potential economic benefits the new technology may bring.

The second phase is characterized by experimentation, disruption, and revolution. Incumbents in the technology experiment with integrating the new technology into their offerings, some with success but also some failures as legacy business models are disrupted by new entrants.

Lastly, disruption accelerates as the new technology enables new business models and drives further innovation.

The last phase is the longest and perhaps the most economically important. The new technology is adopted broadly, contributing to economic growth and productivity.

From ARPANET to the dotcom era – lessons in the first phase

ARPANET began as a US Department of Defense Project in 1969 and yielded two key technologies—packet switching and the TCP/IP protocol—that remain the key underpinnings of the modern internet.

The build out of the internet required staggering levels of capital investment. Telecom companies' spending on their networks increased approximately 25% in 2000 and 100% in just three years.

That capital investment drove massive growth for a new set of technology companies. Cisco Systems, the networking behemoth that provided the

switching and routing gear that powered the internet, grew revenue 53% in its fiscal year 2000 on top of 43% growth in 1999.

At the same time, the technology sector was disrupted itself. Optical networking migrated from SONET/SDH to WDM. In networking, frame relay was replaced by ethernet. Mainframe and minicomputer demand declined inexorably as PCs and servers gained share. At the same time, Windows reigned supreme, and Apple carved out its niche.

Just as modern networking had a lengthy gestation before its widespread adoption, artificial intelligence has had a series of "summers" filled with advancement, followed by "winters" of apathy. However, ChatGPT and LLMs appear to be at the same sort of inflection point seen during the commercialization of the internet and the launch of the iPhone.

At the same time, new business models were being explored. The internet removed the need for a local presence in retail, leading to the rise of e-commerce. Some of the early e-commerce companies took too narrow a focus. Others spent more on marketing than they did in building durable franchises. Alongside this, investors struggled to justify the rapid rise in equity prices with valuations that were divorced from financial performance. Valuations based on clicks and eyeballs became common—not the quality of earnings (i.e., the difference between GAAP/IFRS and adjusted earnings, the difference between earnings and cash flows).

Ultimately, reality set in with a bit of help from the Federal Reserve tightening policy and a weaker economy. Pets.com, emblematic of the suspension of disbelief of the dotcom era, is now a cautionary tale. But there are plenty of other examples, ranging from other e-commerce businesses (Webvan, boo.com, and eToys.com) to search engines (Alta Vista, excite.com, Infoseek, Ask Jeeves) and early prototypes of what would evolve into social media (theglobe.com).

What lessons can we draw from the dotcom era's investment and experimentation phase?

First, investment cycles can run hotter and longer than many expect. Cisco grew revenues at nearly a 70% CAGR from 1995 to 1997, the earliest days of the internet build-out. It then grew revenue at more than a 40% CAGR from 1997 to 2000.

Second, infrastructure investment ultimately has to be supported by revenue from the very services it is meant to support.

Third, macro conditions matter even for secular narratives. A tightening Federal Reserve ultimately set a higher bar for valuations, which for many dotcom-era companies proved too high a hurdle as fundamentals deteriorated. Additionally, many of the leading companies of the dotcom era were not profitable and relied on financial engineering (mergers and acquisitions, secondary stock issuance at inflated valuations). Other companies relied on vendor-financing, a method in which a vendor supports a customer's investment via financing. It turned out that some customers of some large technology companies were actually using vendor financing to finance their day-to-day operations.

Applying these lessons to the current AI investment wave yields potential insights for today's investors.

On the positive, a key difference is that AI investment is first and foremost led by some of the largest companies in the world. These firms have high-quality earnings, strong profitability, and healthy cash flows. Additionally, we are not yet seeing a wave of AI startups with unproven and uncertain business models access the public markets. Lastly, if we benchmark the start of the internet era and investment cycle to the 1995 IPO of Netscape, the internet bubble lasted for five years. Benchmarking today's AI cycle to the November 2022 public launch of ChatGPT implies that we are less than 18 months into the current technology cycle.

Aligning these points with valuations is also illustrative. The trailing 12-month P/E for the tech-heavy NASDAQ index was 175x at its peak. Today, the IT sector trades at a rich but, in our view, justifiable trailing 12-month P/E of 33x.

That said, investors should carefully track the progress of AI-service revenues needed to support further investment. Simply put, the hundreds of thousands of GPUs that are expected to ship over the next few years must be put to work generating profitable revenue. Additionally, even the strongest growth narratives can be interrupted by shifting macroeconomic conditions.

The key takeaway is that while the current investment cycle has echoes of the dotcom era, AI has a better fundamental base and is likely still early days.

Phase two – Learning from others' mistakes

While the technology industry was in the throes of a multiyear downturn from 2000 to 2002, the rest of the world went to work figuring out how to use the internet in their own operations. In the technology sector, some of the survivors of the dotcom bubble rose from the ashes to become today's largest companies, while others never quite found their footing.

The internet brought with it significant economic and social change. Entire new industries were built on top of this new mode of communication. Without the internet, there would be no smartphone, no app economy, no ride-sharing, no e-commerce, no tele-health, no streaming media, no online banking, no social media.

Consumers worldwide clearly benefited from the transformation brought about by the internet. Some industries adopted the technology to drive efficiencies. However, these advances came at the expense of some established industries and their incumbents.

Traditional brick and mortar retailers have been decimated by e-commerce. Amazon evolved from an online bookseller to the "everything store" and to one of the most important IT companies in the world with Amazon AWS. Storied retailers including Sears, Dillard's, and Radio Shack suffered long, slow, and painful declines.

But not every retailer was pushed aside. Walmart remains the biggest retailer globally. Target's trademark red bullseye logo sits atop nearly 2,000 physical

stores today, almost 60% more than in 1999 when the logo was first introduced.

Financial services companies aggressively moved to online banking as a complement to their brick-and-mortar branch systems. In recent years, banks have had to fend off a rising cohort of fintech companies, but rather than chase a wholesale change to their business model or retrench from new competition, the incumbents adopted the best features of their new challengers and further leveraged their intrinsic advantages of scale, trust, and incumbency.

To be clear, there were industries and companies that did not survive, much less thrive. The music industry was shaken to its core by file share service Napster, and it took years to adjust to the new digital world of mobile music.

The linear media industry was relatively unscathed for more than a decade, but we can mark the beginning of its secular decline to 2013, when Netflix became the new "must see TV," replacing NBC in American viewers' canon, with its binge-worthy content and innovative direct-to-consumer model via the internet. The industry remains in secular decline and many of the great studios and content providers struggle to find an economically viable model. Roughly 70% of advertising is now digital, as advertisers follow viewership and enjoy higher returns on advertising spending due to better targeting and measurement.

Against this backdrop, innovators saw opportunity. Google emerged as the dominant force in search after beating back a host of competitors. Facebook, as it was known then, changed the way we all shared information with friends and family and now connects nearly 75% of the world's population (excluding countries where it is permanently banned).

What are the key lessons of phase two of the internet era?

First, technology in and of itself isn't a strategy for corporations. Technology alone will not differentiate a company versus its competitors or overcome a poor strategy. Technology is an input, just like flour for a baker, and companies have to understand what technology can and cannot do.

Second, the hangover from over-investment in IT can last for multiple years. Even though the sector's revenue growth finally recovered in 2003 after two years of declines, the landscape remained challenging. Many of the leading dotcom-era companies that survived the crash struggled in the post-crash decade that followed. It took Microsoft a full decade to fundamentally shift its business and adapt to the new reality. Dell Computer, Intel, and Cisco, which along with Microsoft were anointed the "four horsemen of the internet," never returned to their former growth rates and saw their stocks lag the broader indices for years.

Third, after an investment bubble crash, someone will find a way to capitalize on all that infrastructure, as seen with the rise of Google and Facebook.

What do these points mean for investors evaluating opportunities related to artificial intelligence?

Early on, the internet was sometimes called the "information superhighway." Using this image, we see it can be profitable to be the road builder. But eventually the roads are fully built, and then it is better to be a trucking company. From the perspective of AI, the lesson of the internet may be that the opportunity is larger and more durable in the services and applications that run on the huge infrastructure built out in the first wave.

Phase three – broad adoption drives growth and productivity

The global financial crisis (GFC) touched every sector of the global economy, including technology. But the post-GFC era was a fertile time for the technology sector. From the Blackberry to the iPhone, smartphones changed the way 6.7 billion people (more than 80% of the world's population) live and work. Nearly 50 years after its first conception, cloud computing now captures 40% of all workloads and roughly 20% of all IT spending.

The world's largest corporations are all leveraged one way or another to mobility and the cloud. Furthermore, our smartphones are actually "dumb" in the sense that other than processing videos and pictures, most of the apps rely on cloud computing for their actual computing, so we all enjoy the benefits of the cloud. Cloud computing also enabled the rise of software-as-a-service companies, which democratizes technology by divorcing the burden of owning and managing IT infrastructure from the use of an application.

Economist Robert Solow said in 1987, "You can see the computer age everywhere but in the productivity statistics." Despite the vast technological innovation of the past decade, US productivity growth has largely remained below prior decades. Perhaps this lack of productivity is a function of poor economic data, as former Fed Chairman Alan Greenspan once suggested. Or perhaps "productivity is notoriously difficult to predict," as Dr. Greenspan also said.

Either way, it is clear that the technologies born at the dawn of the internet era are enmeshed in corporate IT environments and our personal lives. Furthermore, the benefits compound over time through standardization, which makes the technology more accessible and more affordable, and network effects, whereby the value of a technology increases in a nonlinear fashion relative to the number of users.

What are the key lessons of phase three of the internet era?

First, foundational technologies like the internet tend to have long-reaching and widespread impact.

Second, this very fact may mean that the benefits of technology adoption are shared across the economy and that it is difficult for any one company to gain a sustainable competitive or comparative advantage based on non-proprietary technologies.

Third, it's good to be a tech company. It's clear that while many technology companies didn't thrive or even survive the dotcom crash, the successive waves of tech companies were the main beneficiaries of technology adoption.

How is this relevant for investment in AI today?

We expect AI to "rhyme" with prior major shifts in technology, but not necessarily "repeat" the failures of the dotcom era. AI, like any other technology, is unlikely to be a panacea for economic growth or any individual company. But investors should think about how AI may disrupt some industries and enable others.

Overnight successes are rare in general, and even more rare in the technology sector. It took decades for ARPANET to become the modern internet. It took 50 years for the premise of cloud computing at scale to be realized in Amazon AWS. It was 41 years from the Motorola Dynatac to the iPhone. So investors need to balance longer-term AI optimism against a pragmatic view of the history of technology innovation and investing.

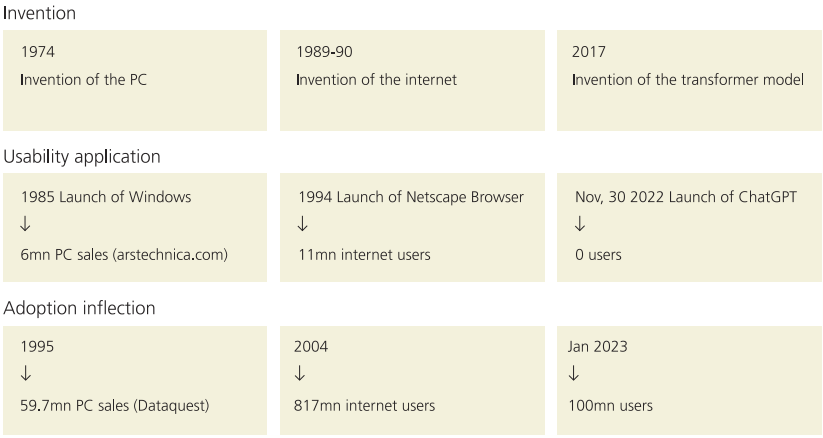
However, the good news is that AI has been a work in progress for over 60 years, and just like prior cycles, the key foundational technologies have seemingly hit an inflection point. History shows us that there will be fits and starts in AI adoption, and that today's leaders may not necessarily be the ultimate long-term winners. But investors should have exposure to artificial intelligence in a balanced, pragmatic approach and be willing to take a fairly long view on the technology given the historical precedent of major waves of technology development enjoying exceptionally long usage.

Chapter 3: Evolution of AI

Introduction

Artificial intelligence is not new. Researchers like Alan Turing were discussing human-like computing systems in the 1940s, and tech giants have been experimenting with AI models for years. It just hasn't had the killer application to propel the technology into the limelight of popular adoption—until now. Launched in November 2022, ChatGPT broke the record at the time for the fastest user adoption ever, attracting 100mn users in two months; it took TikTok, the record holder then, nine months to reach that milestone[21]. Other examples of breakthrough technologies include the Netscape browser, which helped popularize the internet in 1994, and Microsoft Windows, which made the PC accessible to everyday users in 1985. Both the internet and the PC were invented well before these moments, but it was the new interface that made these technologies easy to use. In a similar vein, ChatGPT has opened AI up to widespread use.

Figure 19 - The usability moment for AI



Source: UBS, May 2024

ChatGPT falls into the category of generative AI (genAI)—AI that can create new content, including text, video, images, music and code. The content it produces is very similar to human-generated content, in that it is coherent and contextually aware.

The recent improvements in generative AI can be credited to a new algorithmic architecture called the "transformer model" and the enormous scale that both compute and data have reached. GenAI, alongside other forms of machine learning, will usher in a new age of ubiquitous intelligence, in our view. The next decades, we believe, will be mostly about the AI economy, with wide-ranging implications for jobs, productivity, inflation, and geopolitics.

Large language models (LLMs) may well become a gateway for how we interact with computers, much in the same way the browser became the gateway for our internet access. The model layer may substitute the platform-as-a-service (PaaS) layer, and software development should change quite radically. The value chain in a genAI environment is different. By leveraging LLMs, we can program computers and software through natural language. Software engineers may well work more on data pruning and quality processes than writing source code. What we've described is broadly defined as "software 2.0," where computing infrastructure moves from CPU + software coding to GPU + LLMs/neural nets. Taking it one step further, there may be specific roles for prompt engineers, as the way a business prompts its internal model may become a source of competitive differentiation. While AI seems to be the future of software, there are a couple of important implications.

History of generative AI

Artificial intelligence has been a part of the technology landscape for years, from the success of IBM's Big Blue in chess to Alphabet's DeepMind unseating a world Go champion. Artificial intelligence is a broad term that covers technologies that mimic human intelligence, spanning simple rule-based systems to complex machine learning models.

While there have been several AI cycles in the past, we believe two key moments represent step-changes in AI development. The first is the adoption of GPUs in machine learning (ML), and the second is the emergence of the transformer model for genAI. The latter was developed by a group of Google engineers and published in the paper "Attention is all you need" in 2017.

The transformer model marks a shift away from traditional recurrent neural net architectures that use sequence-based recurrence. Instead, transformer models calculate the relevance of each part of the input data to other parts, allowing the model to focus on different parts of the input data when creating the output data. Transformers handle sequences of data in parallel rather than sequentially, speeding up training and inference.

Large language models like ChatGPT use the transformer model along with large datasets to create a wide range of content, from writing technical texts, engaging in Q&A, and analyzing articles to creating artwork and videos. They can be considered super brains that are getting smarter through learning from data as well as from human and machine feedback.

We believe genAI represents a new technological transition as the way we interact with computers changes and the cost of intelligence drops. To fully understand AI and its potential implications, it is necessary to understand the history and the technology of what has been described as "mankind's last invention."

There isn't a single definition of AI. Even AI systems give different answers to the seemingly simple question of "What is artificial intelligence?" ChatGPT says that AI "…refers to computer systems that are designed to perform tasks that require human intelligence…", while Gemini (Google's genAI model) says AI "…is the field of computer science that aims to create intelligent machines capable of learning, solving problems, and adapting to new situations…". Our working definition is that AI is a field of computer science that creates systems capable of performing many human tasks, but at a scale and level of accuracy that is beyond human potential. This can range from AI systems that are able to find patterns in data that are otherwise undetectable to recommendation engines, to the ability to synthesize vast quantities of information. While AI has powerful potential, it's not without limits in its current form. Most notably, AI suffers from Moravec's paradox: AI models don't perform simple tasks well that humans do every day such as picking up a ball, but they do many complex tasks much better than people could ever hope to. This lack of ubiquitous applicability and the need for accuracy in certain applications limit AI's functionality and potential adoption in the near term.

Categories of AI

Artificial intelligence is not one technology, but rather a range of technologies. Modern artificial intelligence systems primarily fall into the related categories of machine learning and neural networks. While there are various definitions for both, we refer to ML as the classic statistical AI, whereas neural networks are inspired by the human brain's architecture and can detect complex patterns from unstructured data.

The concept of ML was first proposed by Arthur Samuel in 1959. Machine learning uses mathematical and statistical models like linear regression (the y = mx + b that high school students learn in geometry, but with more variables), logistic regression (similar to linear regression, but yielding a discrete value such as one or zero, true or false), decision trees, and more advanced algorithms (e.g., SVM, Naïve Bayes, K-nearest neighbors, K-means, among others). Probability and calculus are also part of the ML math toolbox and are integral to how ML models forecast, optimize, and update model parameters and outcomes.

Neural networks are the fundamental building block of AI deep learning algorithms underpinning genAI, much like how the transistor underpins semiconductors. Their name and structure are inspired by the human brain and mimic the way biological neurons signal to one another. Models use interconnected nodes or neurons in layers to process data. Importantly, because each neuron is adjustable, the systems are adaptive. Each neuron connection, like the synapses in a brain, can transmit a signal to other neurons according to the parameters defining whether one neuron signals to another. Just like transistors process logic by oscillating in an on/off state, neurons process logic by transmitting or not transmitting signals to other neurons. Thus, when we scale neurons and parameters, models become more nuanced, complex, accurate, and useful—however, computational intensity and memory requirements increase.

The evolution of neural nets

As model architecture research has evolved, we have progressed from simple feedforward neurons to recurrent/convolutional neural networks (RNNs/CNNs) and to transformer models today. These different structures of neurons define how the input data is processed by the model.

Figure 20 - The evolution of neural nets

Year developed			
1965	1986	1980s	2017
Key elements			
Acyclic, sequential	Cyclic (output can feed back as input)	Layered spatial hierarchy	Layers, self attention mechanism
Data input			
Static, fixed size input and output	Variable length sequences, one step at a time	Variable length sequence, processed in batches	Variable length sequences, entire sequence at once
Primary use			
Classification	Speech recognition, NLP	Image processing, video analysis	NLP tasks, genAI

Source: UBS

Note: NLP = natural language processing

For instance, in feedforward networks, the information only flows forward through the network layers and each neuron can only signal to the neuron ahead of it. Information is also processed sequentially.

Increases in processing power allowed the practical application of RNNs, first developed in the 1980s, in areas such as speech recognition and natural language processing. Convolutional neural networks, also developed in the 1980s, saw a similar spike in usage for image analysis and classification. While these systems were powerful and created real economic and business values, there were still significant limitations, mostly due to the underlying technology.

Recurrent neural networks have several specific characteristics that limit their effectiveness for language processing. An RNN analyzes sequential patterns and uses pattern recognition to predict the next likely data point. A simple example follows: "I hear the dog…" is almost immediately followed by "barking." The analysis and pattern recognition can be bidirectional, so that the calculation of nodes on later layers can inform nodes on earlier layers to refine the solution, in a process that mimics human learning to some extent. Recurrent neural networks create a gradient to measure errors in outcomes. This gradient can be thought of as a sort of adjustment factor that helps direct the model toward the proper outcome. However, the use of gradients causes some difficulties when RNNs are used in complex questions. As the RNN cycles through its calculation process, it passes the gradient (think of it as a roadmap for error correction) back and forth through the layers, and the gradient itself is constantly updated. However, this iterative updating of the gradient can cause the gradient itself to "vanish"—i.e., become so small as to be unusable, or to "explode," making the model unwieldy and difficult to train.

Convolutional neural networks faced similar issues but with additional complexity from the need for large amounts of labeled data and high computing power requirements (i.e., significant computing capacity, large amounts of memory).# The transformer model, as introduced in the paper "Attention is all you need," replaces recurrent and convolutional layers with a self-attention mechanism.

Essentially, transformer expands on the notion of word vectors to pay attention to the most important words, contexts, and relationships in language. It does this by tokenizing each word in the prompt and prioritizing certain words based on an assessment of importance.

To decide on importance, transformer models assign three values to each word along the three vectors of query, key, and value.

1. The query vector assigns a value based on the question of what seems most important in the sentence by relating each word to every other word.
2. The key vector assigns a value related to the importance of the word.
3. The value vector is based on the word's actual meaning, if it is important (i.e., has a high key vector value).

These values are then weighted and summed. Additionally, instead of the back-and-forth iterative calculation process in recurrent neural networks, attention mechanisms run in parallel.

Other attractive features of the transformer are that it is parallelizable and generalizable, and thus good at arbitrary problems, optimizable via backward

propagation and reinforcement learning with human feedback (RLHF), and their multi-modal model allows them to ingest text and image-based data.

Conclusion: The transformer model revolutionized natural language processing, enabling parallel data processing and significantly improving the performance of a wide range of tasks.

LLMs and the human brain

Large language models share similarities with the human brain in how both process vast amounts of data to train their mental model. However, LLMs have magnitudes more capacity. Even an early model such as GPT-3.5 used approximately 1,000x more input data than human brains for training. But the model parameters the human brain allows for in the form of synapses are magnitudes higher than that of LLMs—similarly by a factor of 1,000x in the case of GPT-3.5. This suggests that the brain has a much more robust training algorithm than LLMs, requiring less data to estimate parameters or establish synaptic connections. Also, the human brain consumes much less energy than LLMs when estimating parameters. This will likely be another area of intense research, as the energy needs pose high costs for LLMs, especially when the data sets and parameter size increase further.

Figure 21 - The human brain and large language models

	Brain	LLM – GPT-3
Learning	Input data: 350mn words read by college graduate Connections: 150tr synapses	Input data: 500bn tokens Model parameters: 175bn
Inference	Stimulus activates neurons ↓ exceeds threshold ↓ pass on to other neurons	Prompt ↓ exceeds threshold ↓ output
Power consumption	Low (20W)	High (>300W)

Source: UBS, May 2024

Conclusion: Large language models are able to ingest much more text data than humans can but have fewer parameters and need more power. The functioning of the human brain is likely to remain a source of inspiration in genAI research, with parameter size increasing and new algorithms developed with lower computational intensity.

Importantly, the transformer model has not exhibited significant diminishing returns to scaling parameters. This is different to previous AI architectures, such as RNNs or CNNs, which exhibited overfitting bias as model parameters and training data sets were scaled. Furthermore, given that model algorithm changes have been insignificant since the advent of the transformer model, the vast majority of model performance improvements have come from scaling up the compute or data used to train a model. This is why we often refer to GPTs as capitalist AI, as by simply investing more capital into compute and data resources, models have substantially improved. It's unclear

whether achieving artificial general intelligence (AGI) will require much more computing power, algorithmic innovations, or simply more data. But to be sure, the road to AGI will depend upon our ability to tackle the current bottlenecks and bolster compute, data, and model architecture.

Training

Training an LLM model refers to the process of unsupervised learning where the model is presented with word sequences and learns to predict the next word. Once a model is trained, the parameters become fixed—creating a generative pre-trained transformer, or a "GPT," foundational model. Training is not a one-off cost, as the useful life of a model is typically quite short. Once trained, it quickly becomes outdated, which means there is a constant need for re-training.

OpenAI's GPT models and Google's Gemini, for instance, provide off-the-shelf GPTs as "intelligence-as-a-service." These models can be used as standalone applications such as ChatGPT or be integrated via APIs into existing workflows.

Prompt engineering and fine-tuning

After the initial training, the model is evaluated on separate datasets to assess its performance on various tasks such as text completion or Q&A. A customer may then choose to finetune their model. Fine-tuning refers to the process of making small adjustments to a pre-trained model to adapt it to specific tasks or training data sets.

Platforms provide tools for fine-tuning GPT models with a user's own data. We expect fine-tuning to be effective over the long run, but as of right now, early adopters have found a better return on investment in prompt engineering over fine-tuning. Prompt engineering is a lighter way to refine foundational models and it is akin to teaching a child through questions, as well-crafted prompts can steer AI models. Simply by asking the right questions, you can get more value out of today's AI.

Compute requirements

The process of training and inferencing requires significant computational resources and can take days to months, depending on the model size and hardware used. Training takes three times the number of GPUs compared to inference[22].

Conclusion: Inference compute demand will likely increase as more models are put into production, but training demand should stay high as new data is added.

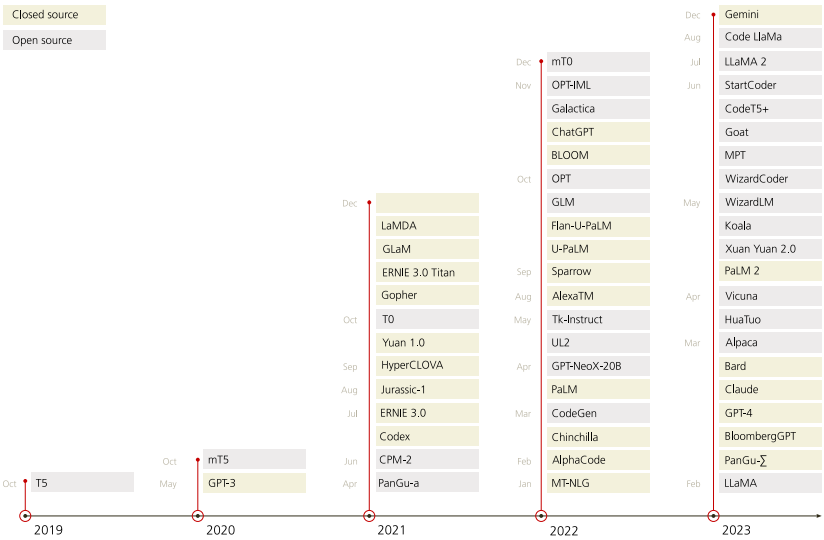
Overview of LLM models

Since the advent of the transformer model, there has been an explosion in the development of LLMs. Large language models both come in the form of proprietary and open-sourced models, with the difference in their licensing, accessibility, and source code availability. According to the 2024 Stanford

Artificial Index Report[23], a total of 149 foundation models were released, more than double the amount released in 2022. Of these newly released models, 65.7% were open source, compared to only 44.4% in 2022 and 33.3% in 2021. The figure below shows an overview of the significant open- and closed-source models released over the last five years.

Figure 22 - Chronological display of LLM releases

The chart illustrates the evolving landscape and trends in natural language processing research



Source: Hugging Face, A comprehensive overview of large language models by Naveed et al, Feb 2024

Conclusion: The largest number of new models are open-sourced, creating significant competition for proprietary models.

Despite many new open-source models, it is noteworthy that proprietary models remain at the top of the Hugging Face leaderboard, which ranks LLMs according to reasoning, general knowledge, and bias across a wide variety of fields.

Figure 23 - LLM Leaderboard

Top 15 positions

1	GPT-4-0513	6	GPT-4-1106-preview	11	Bard (Gemini Pro)
2	Gemini-1.5-Pro-API-0514	7	Claude-3-Opus	12	Llama-3-70b-Instruct
3	Gemini-Advanced-0514	8	GPT-4-0125-preview	13	Claude-3-Sonnet
4	Gemini 1.5 Pro API-0409-preview	9	Yi-Large-preview	14	Reka-Core-0501
5	GPT-4-Turbo-0409	10	Gemini-1.5-Flash-API-0514	15	Command R+

Source: Hugging Face, June 2024

Conclusion: Despite the number of new open-source models, proprietary models remain the top performers. OpenAI's GPT-4 model is still the best-performing model, suggesting that there are scale advantages to a first mover.

The future of the transformer model

The transformer model has become the standard AI model. While model architecture changes are not rare, we believe the transformer architecture may benefit from a first-mover advantage. Since model architecture has been relatively stable since 2017, AI engineers have been able to concentrate their efforts on the transformer.

As a result, an ecosystem around the transformers is developing. As an example, NVIDIA has built libraries to optimize transformer processing. This has made it harder for newer architectures to gain traction, just like how the QWERTY keyboard is still dominant today. By keeping model architecture constant, most model improvements today have come from scaling up compute on the base transformer model. We expect the transformer model to remain a key architecture choice for LLMs for the medium term, but model innovation is possible.

This has led to significant compute demand. As referenced earlier, according to OpenAI, between 2012 and 2018, the amount of compute used in the largest AI trainings doubled approximately every 3.4 months, leading to a 300,000x increase in compute required during that period.

Conclusion: The transformer model is likely to remain a key architectural choice for LLMs.

Limitations

Large language models suffer from significant shortcomings, in particular "hallucinations" (i.e., they can produce inaccurate facts) and inconsistencies (i.e., they can come up with different answers to the same question). At this stage, they cannot be used autonomously but need human supervision, very much like an apprentice who will learn and improve abilities over time.

This issue is particularly pertinent in mission-critical applications such as medical diagnostics and autonomous driving. The implications are that some industries may lag AI adoption and that people working alongside AI "in the loop" is key. Importantly, AI is a copilot—it is not an autopilot—and we should view AI as an intern-to-manager structure where the manager remains accountable for the output.

However, the field is advancing rapidly and several new approaches, such as retrieval augmented generation (RAG), can help provide more accurate, detailed, and up-to-date answers and scenarios where LLMs need to reference data that is not part of the training set.

Conclusion: While genAI suffers from shortcomings such as hallucinations, new techniques such as RAG are rapidly developing to address these issues. Further advances in the abilities of LLMs should provide more accurate and consistent content over time.

New architectures: State-space models

The transformer model is unlikely to be the only game in town. New architectures such as "Mamba" have shown good promise, particularly when handling longer sequences (text prompts). Mamba is a state-space model, and its more efficient computing graph allows for smaller models with

faster throughput while maintaining accuracy. The Mamba model achieves 5x higher generation throughput compared to transformers of similar sizes, and their smaller size allows for less intense and faster inferencing. Despite the different architecture, GPUs still underpin the vast majority of computing for Mamba models but at a lower scale than in a transformer world. However, model architecture shifts should be viewed as an advantage for GPUs, as their general-purpose approach is fitted more for dynamic computing than for custom AI chips.

Conclusion: New LLM architectures are also likely to rely on GPUs; in fact, new architectures are likely to benefit GPUs, as they are more general purpose in nature than custom AI chips.

From data to applications: the LLM

The ability to build high-performing LLMs that power these applications depends on three factors: the size of high-quality datasets, the underlying algorithmic architecture, and the amount of compute resources.

Figure 24 - From the data to applications

Source: UBS

Data

Data is the key input into LLMs. Enormous amounts of data from a range of sources are used in the training process. The latest LLM models ingest content from books, academic papers, and the entirety of Wikipedia, totaling more than 45 terabytes of data.

Figure 25 - Data examples

Data
Examples

	Data inputs						Data infrastructure	
Type	Internet data	Community forums	PowerPoint, Excel, Word documents	Search queries	Books	Scientific articles	Vector database	Pinecone, MongoDB, Snowflake
Company	Common crawl	Reddit	Microsoft	Google	Scholastic	Arxiv	Training	MosaicML, ML Studio

Source: UBS

The collected data is then pre-processed, cleaned, and broken down into manageable pieces (tokens). Cleaning and curating data can be time-consuming; operators can't simply dump their existing database into these models and hope to get great insights. Curating involves removing noisy data or outliers, using techniques like regression to filter out noise from training data, handling missing values, and removing duplicates or clustering similar groups of data together.

Moreover, securitizing data is paramount as data becomes increasingly a vector for competitive differentiation. One approach to data security is to leverage vector databases. Vector databases allow firms to encrypt data without losing its semantic, making it more difficult for hackers to

decode data while ensuring the data remains useful. We believe fears of data sovereignty are proving a major headwind to AI adoption, and the development of vector databases is one technique that can help alleviate such concerns.

Regarding foundational model training and datasets, GPT-4 was trained on 13 trillion tokens—or the Common Crawl dataset. The Common Crawl data consists of petabytes of data from over 50 billion web pages collected since 2008 by a non-profit. The scale of data ingestion alone is impressive; however, training is significantly more effective on curated, labeled, and cleaned data sets. Hence, a major driver of OpenAI's performance advantage comes down to its curation and pre-processing of that dataset. Moreover, because of improvements in pre-processing and the inevitable data drift over time, there is a perpetual need to periodically retrain these largest models.

While we view AI as a rising tide with outsized benefits for the world, the imbalance in data training sets could exacerbate certain existing inequalities. For instance, genAI won't have the same functionality in Welsh as in English, and continents like Africa significantly lag the rest of world in terms of data captured. The implications can already be seen: For instance, many vision models are able to recognize Western weddings by photograph but most fail at recognizing African weddings.

There are limited ways to accelerate the creation of mass reliable training sets. For vision AI, simulating data is a proposed solution. However, they bring questions as to whether an AI trained in simulation can truly operate with real-world factors and raise accuracy concerns. It must follow that there are diminishing returns to data scaling, as each incremental model functionality should be an incrementally more niche application.

Large language models

Because of their generalized applicability and robustness, we will likely see a number of large foundational models led by OpenAI, Google, Anthropic, Meta, and others.

Figure 26 - LLM examples

LLM
Examples

Name	GPT	Bard	LLaMA	Claude	Gemini	Mistral7B	Tongyi Qianmen	Yi-34B
Company	OpenAI	Google	Meta	Anthropic	Google	Mistral	Alibaba	01.AI

Source: company data, UBS

In this regard, we expect several "AI foundries" to be established with customers integrating these models either in existing or new applications. It's simply too expensive, time-consuming, and complicated to develop leading-edge models in-house. End users will have to intermittently fine-tune the models they receive from third parties as their databases drift and as the underlying large GPTs are re-trained.

In deciding which LLM to integrate, end users weigh model performance, inference cost per token, and other features such as permitted prompt length. However, there are also many use cases where the top model would be inefficient and expensive relative to a smaller model. For many use cases, enterprises may prefer a smaller model with cheaper inferencing for specific tasks, as they may not benefit from a model's generalized performance as much as consumers. In other words, the model doesn't also need to be great at poetry if it's trained to detect manufacturing defects in images. Therefore, we expect many smaller models to be developed over time.

Smaller models would also enable AI at the edge, defined as AI processed outside of public and private clouds, the adoption of which we view as inevitable due to their data privacy and latency advantages. For self-contained models at the edge, personal data never has to be shared with a third-party player in the cloud. We believe having clarity on data privacy will be an important building block to building a companion personal assistant AI.

Conclusion: We expect about 10 large dominant foundational LLMs (closed / open, US, EU, Asia) to be long-term winners.

Compute

Large language models require large amounts of compute for both training and inference. The models can be trained in three different locations:

1. In the cloud where shared resources can help scale compute when needed
2. On premise in a data center with full control of data and model
3. On an edge device where training and inference will be local (and private) on a device

Figure 27 - Storage and compute examples

Storage + Compute
Examples

	Cloud				OnPrem			
Product	AWS	GCP	Azure	OCI	GPU	CPU	RAM	SSD
Company	Amazon	Google	Microsoft	Oracle	Nvidia	AMD, Intel	Micron, Samsung SK Hynix	Western Digital, Samsung, Marvell

Source: Company data, UBS

Given the scale of growth in compute, most performance improvements have come from using more interconnected GPUs rather than faster chips. Indeed, underlying chip improvements cannot match the current rate of AI model scaling, and this has led to an explosion in demand for GPU units as well as networking solutions.

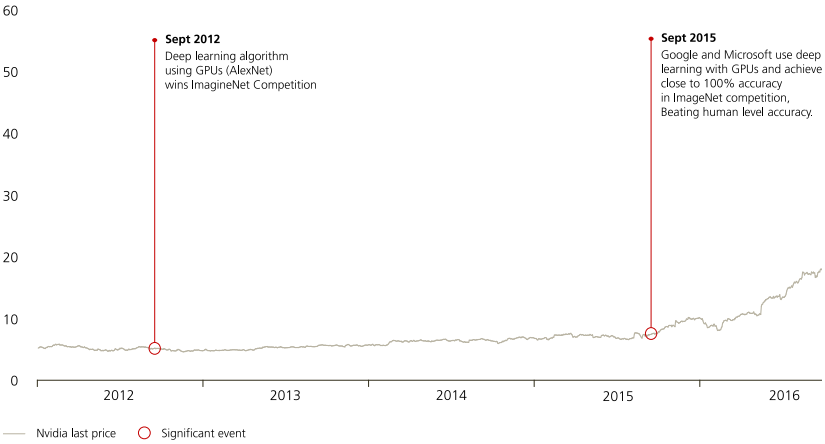
If the transformer model remains the leading system, specialized hardware through custom application-specific integrated circuits could eventually outperform general purpose GPUs for certain use cases. We note that all neural network architectures eventually plateaued, and transformers are likely to follow the same fate.

Case study: The first GPU moment for AI

An interesting analogue to assess the significance of GPUs for LLMs is to study the GPU demand impact from deep learning for computer vision. Graphics processing units were first used in convolutional neural networks (CNNs) for computer vision in 2012, when AlexNext won the ImagineNet competition using the technology. Prior to this date, NVIDIA was known for its innovations in GPUs for computer gaming. With the use of GPUs in computer vision, NVIDIA expanded the breadth of applications for its GPUs. This new market opportunity was one of the factors that helped NVIDIA's share price rise from September 2012 to December 2016.

Figure 28 - Case study NVIDIA: Deep learning and computer vision

NVIDIA price chart from January 2012 to December 2016



Source: Bloomberg

Challenges to adoption

There are some immediate challenges to adoption that could temporarily slow momentum around AI that we discussed earlier, such as data protection concerns from enterprises. Also, enterprises need to make the right investments in their back-end infrastructure to make the most of genAI. This includes modernizing their data stack, integrating genAI into their workforce's workflow, and continuously integrating the latest, most relevant models.

Building and training a model in-house is too costly and not differentiating enough for enterprises to handle, so most of the value should come from fine-tuning the model onto their own dataset. Data has been valuable for decades, but many enterprises are still only just realizing how valuable that data is. After all, we believe that data protection worries from enterprises may slow the immediate pace of adoption. The genie is out of the bottle—and continued ingenuity on model efficiency and data training should drive adoption.

Chapter 4: Everything everywhere all at once: AI's reshaping of key sector and regional dynamics

The sector contributions were written by our regional sector strategists (authors mentioned below) and supported by our APAC team (Chisa Kobayashi, Soekching Kum, Hartmut Issel, and Carl Berrisford).

Introduction

In the fourth chapter, we take a closer look at the different sectors and how (generative) AI may impact them in the coming years. Along with the impact

on business models, we have examined how it could affect sector revenues, operating margins, pricing power, and long-term share price performance. Although AI should have a neutral to positive impact on corporate topline and operating margins for most sectors, pricing power may suffer for several industries if AI has a deflationary impact on product prices. Certain industries are used to technological disruption, but for others, AI will create a bigger challenge and companies must adapt their business models fast to stay competitive in the market. Many of these AI-led changes could also have an implication on sustainable development, as the technology enables us to use resources more efficiently and to deliver much needed products and services to remote and/or underserved communities. We discuss these implications in greater detail in the following chapter.

The impact across sectors is diverse:

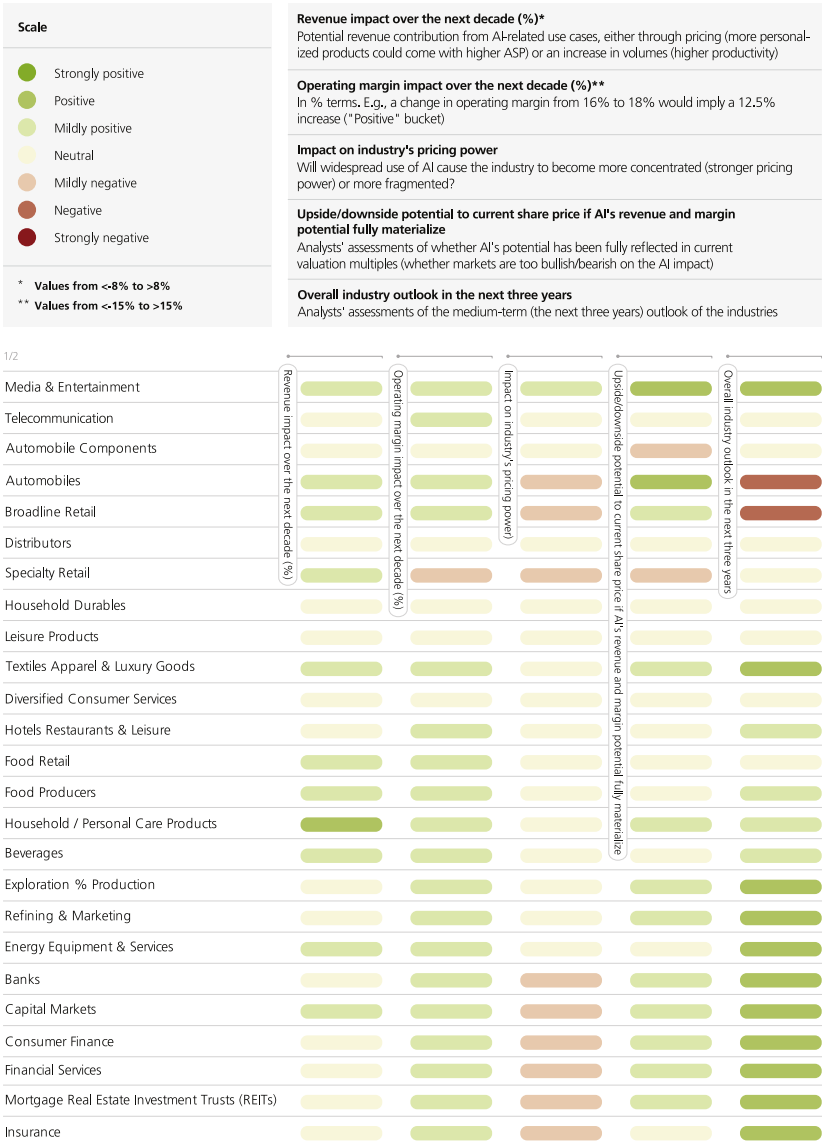
- In the **automotive** industry, AI will likely support the development of autonomous driving but put pressure on pricing and margins due to efficiency gains.
- The **IT** and **communication services** sectors are at the heart of the AI revolution. In our view, AI will be the tech theme of the decade and provide strong tailwinds for them. We believe tech currently offers the best exposure to the AI trend, particularly in the semiconductor and software industries.
- The **healthcare** industry hopes that AI will shorten the drug discovery cycle and accelerate the FDA approval success rate for discoveries.
- The **financials** sector is likely one of the sectors with the largest cost opportunities given that IT and personnel account for a major part of the overall cost base. Chatbots and virtual assistants powered by AI can offer 24/7 support.
- The **consumer** industry can leverage AI to anticipate demand more accurately, thereby ensuring efficient resource allocation, reducing excess inventory, and mitigating food waste. But it will also likely lead to higher competition and lower consumer fidelity in online retail. In the **luxury** segment, AI could play a role in enhancing the consumer journey.
- The increasing demand for AI datacenters has already had an impact on the **real estate** sector. Proptech AI could drive additional business opportunities.
- In the diverse **industrials** sector, a range of AI applications are already in use and monetized by firms. For instance, AI should increase the opportunities of digital twins in industrial applications even more. However, the outsourcing industry (which is also part of the sector) may see disruption due to the proliferation of sophisticated AI chatbots.
- In the **materials** sector, chemical companies are already using AI to streamline prototyping and production processes.
- Finally, the **utility** and **energy** sectors will likely both benefit from higher energy demand. According to the US Energy Information Administration (EIA), an internet search using AI can use up to 10 times the amount of electricity relative to a more traditional internet search.[24]

In our heatmap below, we also included a sector outlook on a three-year horizon. We are positive on most subsectors, but we are cautious on select

areas (e.g., the automotive industry) and two industrial end-markets (air freight and logistics and machinery). In the following pages, we discuss certain companies to highlight generic trends in the industry to better explain how AI may impact business models. Readers should note these case studies are for information only and do not amount to investment views. For investment ideas and our stock picks, see our companion report "Investing through the AI platform shift."

Impact of AI on industries

Heatmap



⊙ Neutral means the potential impact of widescale AI adoption is difficult to predict and includes a range of outcomes spanning from quite positive to quite negative. At this early stage of AI adoption among life science tool companies' customers, we see too wide a range of potential outcomeson overall industry R&D spending to label AI adoption as either clearly positive or negative

Source: Based on MSCI sector classification. UBS estimates as of March 2024

Automobiles

Author: Rolf Ganter

What does generative AI mean for the sector?

Generative AI will likely affect the automotive industry (manufacturers and suppliers) across the whole value chain—from R&D, design, purchasing and logistics, production, sales and marketing, and after-sales to the tech features and services offered to consumers in the vehicles. The impact will likely be felt mainly through lower costs and lower barriers to entry. We see only limited revenue opportunities at this stage (more below). In terms of margins, we

see only temporary improvements due to already high competitive pressure in the industry and new competition—e.g., from Chinese players—and potentially tech firms. On the supplier side, cut-throat price competition and pressure from auto manufacturers have led to low supplier margins. Margins throughout the industry remain low despite all the historical productivity improvements and efficiency programs, and AI is unlikely to change this.

What are the new AI revenue opportunities?

The revenue opportunities are limited, in our view. In-car entertainment features are one area, though consumers may not be willing to pay much as most features are available on their handheld devices or mirrored on their on-board screen. Autonomous driving technology, such as full self-driving (FSD), likely offers the greatest opportunity. Vehicles already offer the necessary hardware for FSD, and the software is being developed and will likely continue to improve over time (beyond the current FSD level of 2+). However, legal and technological hurdles are still high to achieve the required level 4 (fully automated / autonomous) for personal vehicles and level 5 (full automation with no driver) for robotaxi services, in our view.

Which areas may see productivity gains?

Productivity gains driven by AI could be very strong and along the whole value chain. Auto companies are spending heavily on R&D—on average 5–7% of annual sales—and the overall product design and testing phase before market launch could last up to seven years. Competitive pressure to speed up time-to-market is increasing, and genAI should help on the design and R&D efficiency sides (like in pharma). Auto R&D expenses could therefore fall substantially over time. Auto production is already well advanced in terms of just-in-time or just-in-sequence concepts, but AI could help further optimize the process—as well as for purchasing and logistics planning. While the impact on auto production, particularly for blue-collar workers, appears limited, significant efficiency gains could be realized in white-collar sectors, which like other industries constitute approximately 40% of the auto sector.

Product planning, sales, and pricing may be the most impacted. Better predictive power could result in improvements in assessing upcoming repair works, as well as corresponding spare parts planning and sales. Better visibility in projecting consumer behavior—e.g., option uptake during car configurations—could immediately be used to optimize the internal processes and adjust pricing in favor of higher margins. Personalized marketing and sales campaigns may also see gains over time. One should bear in mind that as of today, pretty much all the client-centric knowledge is with the local dealer as a point-of-sale rather than with the auto manufacturer. Thus, acquiring more client data through direct engagement with customers can provide valuable insights for optimization, resulting in lower costs and potentially higher revenues.

What does it mean for consumers?

Software is increasingly defining vehicles more than the hardware. And with traffic increasing throughout the world's motorways, top speed—beyond the German autobahns—has become irrelevant. The latest software and entertainment gadgets are steering consumer decisions. Autonomous driving

technology, which uses troves of collected data to train the systems, and fast computing power in the car itself may be top of mind for consumers by the end of the decade. Manufacturers may apply virtual and augmented reality for better navigation, with windshields also used as a screen, using AI technology to offer a unique consumer experience. More and more cars include voice-controlled systems like MBUX in Mercedes-Benz, Alexa in BMW, or ChatGPT in Volkswagen cars as announced during the CES Vegas. Modern cars are constantly generating data—while driving or stopped—as genAI draws a lot of (right) conclusions about the user and its surrounding environment. This means privacy and data security concerns are likely to be top priorities for regulators and customers alike.

To what extent is AI disruptive for the sector?

We see it mainly as a cost-efficiency story, and if integrated smartly, it could bring costs across the whole value chain substantially down. Barriers to entering the industry have dropped thanks to the shift to less complex electric powertrains. Disruptors, namely from Asia, are already benefiting from applying more technology and are able to beat incumbents on costs, which has started to become a disruptive force. However, AI could lead to the disruptors being disrupted—either by the next generation of companies entering the auto industry or existing incumbents with strong brands using technology at scale across their businesses to improve their positioning. But, in our view, in the auto and auto supplier industry, temporary leads are normally competed away relatively fast, with ongoing pressure on pricing and margins.

Case study: Mercedes-Benz

Mercedes-Benz plans to use AI to bolster its MBUX multimedia system. MBUX collects a lot of personal data within vehicles, and anonymized data records when testing and training AI systems are being used to ensure the highest level of data protection.

Case study: Speeding up F1

Some auto manufacturers like Ferrari and Mercedes-Benz are heavily engaged with their own Formula 1 racing teams. Tens or even hundreds or thousands of a second can decide if a racer makes it to the podium at the end of the race. A Formula 1 car has hundreds of sensors to continuously monitor it. Collecting and analyzing data real-time is key, and with the help of AI, faster analysis can be made and important conclusions can be drawn. This allows drivers to immediately adjust their vehicles and may determine where the team ends up at the end of the race.

Case study: BMW

BMW uses AI to further optimize its production. Real-time communication between the vehicle in the production line and the highly sophisticated production system takes place. By constantly exchanging information, and via the help of cameras and sensors in the production line, any faults can be immediately spotted and addressed, substantially reducing the need for any costly and cumbersome rework. BMW also uses AI in its Acoustic Analytics for quality checks. With the help of microphones and AI, driving noises are identified, analyzed, and classified, recognizing if the respective noises sound normal or abnormal and therefore need to be addressed.

Case study: Preemptive maintenance

The paint shop is a highly complex, energy-intense, and quality-sensitive part in the auto manufacturing process. Any interruptions are costly and troublesome. Dürr, a leading German engineering company for automotive painting technology, uses AI across the board: predictive maintenance to reduce plant downtimes but also to optimize any required maintenance work, as well as recognizing systematic errors in the painting process and using the self-learning capabilities at an early stage to immediately find and address those issues.

Communication services

Author: Kevin Dennean (US strategist)

What does generative AI mean for the sector?

Generative AI brings varying opportunities across the different industry groups within the communication services sector. Interactive media companies are already capitalizing on generative AI to drive user engagement and advertising revenue. We see this group as a continued beneficiary over the next few years. In contrast, we see generative AI as a modest headwind for the media and entertainment industry. The potential cost savings brought about by this new technology will likely be offset by further viewership and engagement headwinds. Lastly, we see generative AI as largely neutral for telecom companies.

What are the new AI revenue opportunities?

Internet companies may see higher levels of engagement that can be monetized through advertising. Entertainment companies may see mixed trends—AI could help create compelling new content, but generative AI may further "democratize" content creation and serve as an additional headwind for media and entertainment companies. Media companies will likely see further cannibalization of traditional publications by online content. Lastly, we don't see AI as a major impact on demand for telecoms, which are dominated by consumer wireless firms and are unlikely to see a material revenue driver.

Which areas may see productivity gains?

We see generative AI as mildly positive for AI sector margins. Internet companies already leverage AI to a large degree and enjoy high margins, so margin gains from here will likely be modest. Media and entertainment companies may see some cost savings from AI-generated content, but those cost savings are likely to be balanced by lower revenue. Telecom companies may see some margin upside from the use of generative AI in customer care.

What are the impacts on industry structure?

We don't expect a significant impact on the structure of industries within the communication services sector. The internet industry is already highly concentrated, but it may benefit from further scale effects. Generative AI may lower the barriers to entry for media and entertainment, in keeping with the adoption of Web 2.0 and the internet. The telecom industry is already an oligopoly, a structure that is unlikely to change from AI.

What is the upside/downside potential for the sector if the benefits/ risks of AI are fully realized?

On balance, the communication services sector should be a modest beneficiary of widespread generative AI adoption. We see this emerging technology as broadly positive for the internet industry, which accounts for most of the sector's revenue, profit, and market capitalization. Media and entertainment companies could see some headwinds (fundamental and valuation), but the impact should be limited given the industry's relatively small contribution to the overall sector.

Case study: Spotify

Spotify, a leader in music streaming and podcasts, uses generative AI to deliver highly personalized content to its more than 600mn monthly active users. With a catalog of more than 100mn songs, 5mn podcasts, and 350,000 audiobooks, the company needs to drive broad engagement to maximize advertising revenue and leverage its licensing costs.

DJ is a new feature that utilizes users' data in an artificial intelligence recommender engine to create compelling playlists tailored to individual users. Additionally, DJ's generative AI capability creates commentary about the recommended artists and songs, delivered in a realistic human voice. Lastly, users can shift moods and therefore the recommended music with a click of an icon. Users benefit from a more dynamic and interesting music experience, along with the discovery of content and information they may never have come across otherwise.

Consumer staples

Authors: Seraina Hold, Carsten Schlufter, Sunny Mehra (US strategist)

What are the key areas of productivity gains from AI?

At this point in time, we see the main areas of productivity gains for consumer staples companies in making production more efficient, sales planning, and optimizing supply chains. AI can help track inventory levels, predict demand spikes, and improve shipment routes, thus minimizing waste and optimizing logistics. This can lead to reduced transportation costs, streamlined inventory management, and improved delivery times. Furthermore, AI-powered quality control mechanisms allow for real-time monitoring of production processes and ensure consistent product quality. AI is also set to automate repetitive tasks. This, coupled with better integrated systems, should result in cost reductions and improved overall efficiency.

AI is already improving overall operations and inventory management for many companies. The French retailer Carrefour, for instance, leverages AI to predict demand more accurately by key category, thereby ensuring efficient resource allocation, reducing excess inventory, and mitigating food waste. With the use of AI, the company said it optimized the shelf stacking in its bakeries and thereby saved 5 million croissants within a year. Beverage companies, such as Heineken and Coca-Cola European Partners, use AI to improve the route to market. Unilever uses AI to optimize its design and manufacturing processes through virtual simulations, automated workflows, and data-led decision making.

Case study: L'Oréal

Technology is transforming the way L'Oréal engages with its customers, the way it creates, and the way it spends its money. One example is BETiq (BET stands for beauty engagement touchpoints). It is an AI-powered engine that allows the company to find the best ways to allocate money, measuring and improving the return on advertisement and promotion (A&P) investments. The tool helps to focus on both short-term return on investment and on building brand equity for the long term. The company uses the tool in four countries today, and according to CEO Nicolas Hieronimus, it is generating "productivity increases of up to 10–15%" for some of the L'Oréal brands that have deployed it.

What are the new revenue opportunities based on AI?

The use of AI across the sector allows companies to better understand consumer behavior and to cater to changing consumer needs and preferences. We believe that this opens new revenue opportunities through better connectivity with consumers and better and faster product innovation. Personalized marketing initiatives, driven by AI algorithms, enable companies to target specific customer segments with more effective and tailored promotions and advertising campaigns. Moreover, better insights into consumer preferences and market trends help companies identify emerging opportunities and evolving consumer needs and come up with more personalized and more relevant product innovation. This is set to boost customer satisfaction and loyalty. For food retailers, we believe the main revenue opportunity of integrating AI lies in retail media. Retail media is linked to the ability to deal with huge amounts of first-party data that the grocers

collect (through loyalty programs) and can monetize through advertising or marketing campaigns.

Nestlé uses AI to analyze social media, online publications, and other web sources. This accelerates the creation of new ideas or trends that can be translated into product innovations. One example of such innovation is "Petivity," Nestlé's product ecosystem to help improve cat health. The ecosystem includes a smart litterbox that can track changes in urination, defecation, and weight patterns. Petivity can identify an increased risk of developing conditions such as kidney disease, urinary tract infections, or obesity and can then provide expert advice to pet owners, including recommendations for veterinarian visits when relevant. Walmart is using AI for customer search on its app, improving the customer experience and potentially generating higher sales. For an event or occasion, this allows customers to be given a customized solution consisting of multiple products, together saving the customer time and improving their overall experience. L'Oréal, too, has been investing in new technologies for many years. The company, for instance, developed an AI-powered make-up applicator designed specifically for consumers with limited mobility ("HAPTA") and an AI-powered beauty advisor that gives personalized advice to consumers ("Beauty Genius").

Conclusion – keeping up to stay relevant

Looking ahead, we believe that leveraging AI and digital tools broadly will be key to remain competitive in the consumer staples space. Doing so not only allows companies to develop better-resonating products, but also helps them produce products at a lower cost. We believe that early adopters and businesses with robust data collection and analysis practices stand to benefit most, while companies that are clinging to outdated methods risk falling behind. Companies that do not have strong digital capabilities will need to materially step-up reinvestments, in our view, which can curtail margin ambitions.

Energy

Authors: Rudolf Leemann, James Dobson (US strategist), Alexander Calvert (US strategist)

What does generative AI mean for the sector?

The energy sector should see a slight positive benefit from the adoption of generative AI across the global economy. Most of the benefits should come from operational and cost efficiencies across this complex industry (see third question below). Additionally, revenues could come indirectly from increasing energy demand for AI datacenters. Importantly, though the potential operational cost benefits for the energy industry may be widespread, the revenue benefits are likely to be regionally disparate driven by the location of AI datacenters. In the US, AI datacenters are already driving up electricity demand growth expectations. This is occurring in regions with available electricity supply and reasonable electricity prices. Some of this

electricity demand will be met with renewables, but some will also be met with fossil fuels, primarily natural gas.

What are the new AI revenue opportunities?

Most of the revenue opportunities from AI for the energy industry will likely be indirect, resulting from increasing energy demand for datacenters. This should prove true for the energy subsectors of exploration and production and integrated oil, as well as independent refining. According to the US Energy Information Administration (EIA), an internet search using AI can use up to 10 times the amount of electricity relative to a more traditional internet search.[20] AI's increasing demand for energy, primarily electricity, should drive more electricity generation using natural gas.

The energy equipment and services subsector could see greater revenue opportunity as AI takes on a bigger role in exploration and production services, which are partially or entirely outsourced by energy companies to field services companies. The largest of these companies are aggressively adding AI capabilities to their digital offerings, which could improve operational performance and add pricing power and revenue potential to the oil field services sector. These AI services could be applied to subsurface geologic analysis, well design, and production efficiencies, among other things.

Which areas may see productivity gains?

Given that the energy industry is complex—including the exploration, production, processing, and refining of raw hydrocarbons; heating and transforming hydrocarbons into useful products; and organizing complex logistics around global supply chains, intricate manufacturing systems, and customer delivery—we see many applications for AI that could improve operating efficiency and reduce operating costs for the energy industry. Improving efficiency and lowering costs are the primary benefits for the energy industry from AI, in our opinion.

What are the potential challenges of AI integration?

The energy industry is used to managing complexity and is embracing AI to help it. Hence, we think the challenges of AI integration for the energy industry are limited. Any challenges to AI integration would simply slow the capture and realization of potential operational efficiencies and cost benefits.

To what extent is AI disruptive for the sector?

The application of AI could help improve energy efficiency and advance the development of alternative sources of energy. Such applications will likely take several years to come about, but at the margin, AI could accelerate the evolution of the global energy system toward energy sources with lower carbon intensity.

Case study: Using supercomputers in energy

The energy industry has been an early adopter of leveraging computing power to manage the many complexities in producing energy, including their multi-faceted supply chains and intricate conversion and distribution systems. Some infrastructure investments can easily reach investment commitments of USD 5bn plus, such as offshore exploration and production, natural gas liquefaction (LNG) plants, or refineries expanding across large surfaces, sometimes square miles. Hence, managing the feasibility, build, and safe operations of this infrastructure has always involved extensive computer support. Geology sub-surface surveys to identify new resources and four-dimensional reservoir simulation are primarily computer-model based. Hence, some of the largest supercomputers are with energy-related entities.

One example is the "Frontier" supercomputer at the United States Department of Energy, which in May 2022 became the first machine to exceed 1 ExaFLOPS (i.e., more than 1 quintillion subtractions or additions per second). One of its applications is a study into the behavior of physical processes within rock structures, specifically to model multiphase flow processes, and to understand heterogeneous wettability (i.e., the surface energy between fluids and solids, which influences the productivity of oil fields). Another is Eni, which sees its facility "HPCx" regularly appearing in the top 10 list of global supercomputers. Although AI deploys a less centralized computing power application at the moment, the energy industry—as an early adopter in computing power in general—is embracing AI too.

Financials

Authors: Sacha Holderegger, Thomas Parmentier, Bradley Ball (US strategist)

What does generative AI mean for the sector?

Financials are likely one of the sectors with the largest cost opportunities, given that IT and personnel account for a major part of the overall cost base. Accenture estimates that more than two-thirds of workers in the US financial sector are deployed in functions that have a high potential for automation and augmentation. A reduction of 10% in bank personnel costs, for example, would on average improve return-on-equity by about 100 basis points, according to S&P Global Ratings (based on Global Top 200 rated banks). However, we do not expect AI to widely replace humans in the financial sector in the near term but rather support them by enhancing productivity.

Within the sector, we expect a wide discrepancy of impact between companies depending on scale, profitability, and their "digital" starting point, as companies will need to adjust their business models through controlled self-disruption of their legacy IT structure. Regionally, North American financials are leading in terms of AI adoption, while European and Japanese counterparts are ramping up investments. Artificial intelligence also holds the promise of facilitating credit access for the underbanked and the unbanked with no credit history, by capturing insights from alternative sources of data to generate credit scores for these individuals.

What are the new AI revenue opportunities?

Revenue opportunities result primarily from enhanced client interaction, new product development, and market expansion. We think this should especially benefit corporate and retail banks. With AI-driven chatbots and virtual assistants providing 24/7 support, customers can receive tailored financial advice and product recommendations to individual client needs, potentially increasing cross-selling and upselling opportunities. In the insurance sector, AI-driven analysis can help insurers assess risk, detect fraud, and reduce broker distribution fees. It could also help accelerate the development of innovative products and services by identifying emerging trends and customer needs through data analysis, and better serve underserved client segments at lower costs. However, competition could increase as some barriers to entry are lowered. This may cause some of the new volume benefits to be priced out as margins come under pressure.

Case studies: A brave new financial world

Europe: AI-driven robot-advisors are being deployed to provide personalized investment advice at a lower cost than traditional advisors. Société Générale, for example, launched Eliott, a callbot that manages around 4 million client conversations per year and over half of the conversations between the French online bank Boursorama and its 3 million clients. KBC and Natwest also have their own AI-powered virtual assistants: Kate and Cora.

Asia: There has been a rise of fintech robo-advisors, such as Endowus, Syfe, and Stashaway, providing affordable and tailored investment advice and options. AI is also a competitive advantage for early adopters in Southeast Asia's financial industry. For example, DBS reported economic value of SGD 180mn from its AI use cases in 2022, comprising a revenue uplift of SGD 150mn and SGD 30mn from cost avoidance and productivity gains.

US: A US start-up launched the first genAI-based tool to simplify and streamline the underwriting process and assess risk by collecting a myriad of data through multiple third-party data sources, spotting patterns, and summarizing the same into desired insurers' underwriting report format.

Which areas may see productivity gains?

We see productivity gains as the main benefit from adopting genAI in the financial sector. AI can automate routine and manual tasks such as transaction processing; it could also handle customer inquiries, offer better pricing, help with compliance checks, enhance decision-making processes, and streamline middle- and back-office operations. In risk management, AI models could enhance fraud detection and credit risk assessment by analyzing vast amounts of data in real-time, improve decision accuracy, reduce losses, and improve loss prevention for insurers. In compliance, AI could streamline the compliance process by monitoring transactions for suspicious activities and ensuring regulatory reports are accurate and submitted on time, reducing the risk of fines. Finally, a key productivity benefit of AI may be realized in the time freed up for financial firm managements to focus on more meaningful tasks.

Case studies: Reducing costs, improving satisfaction

Europe: Banks like Barclays, BBVA, Danske, and Société Générale utilize AI to analyze transaction patterns in real-time, identifying and preventing fraudulent activities more effectively. In risk management, Santander developed an AI-tool called Kairos that shows how a corporate client could be impacted by economic events, creating prediction patterns that enable it to make more informed investment and lending decisions.

Asia: Through hyper-personalized AI and data analytics, DBS has reduced the time needed for the financing application and approval process for its SME customers to one minute and one second, respectively. Even disbursements are now instant with no additional paperwork. In Japan, MUFG Bank estimates that with the introduction of ChatGPT into banking operations such as office work and sales, it will be able to reduce more than 220,000 working hours per month.

Insurance: AI can play a crucial role in risk identification and mitigation. By analyzing claims data from multiple sources and forms, historical claims, and external factors (e.g., weather patterns), genAI models can help P&C insurers identify areas prone to losses, which can aid in the development of risk mitigation strategies (e.g., recommending safety improvements, policy adjustments to reduce future losses).

What are the potential challenges and how may regulation impact AI integration?

There are significant challenges to AI integration, including the prevalence of legacy technology in many companies (mainframe legacy IT infrastructure and data silos), the ability and cultural change to move to a tech-led organization, as well as the need to comprehensively explain how certain outputs are generated by AI. Competition from disruptors (fintech) is a threat to the current incumbents, and new competitors can launch reasonably quickly as barriers to entry into certain markets are low (e.g., personal finance, motor insurance). A deep learning neural network like an LLM is not easily comprehensible, which could be viewed with suspicion (i.e., a black box) by clients, employees, and regulators and might pose a problem when questioned by a court. This could increase the risk of fines and regulatory problems for financial companies.

Case studies: Other use cases

Europe: ING is actively pursuing talent-building avenues by sponsoring AI education by collaborating with Dutch universities. Barclays provides AI training to senior management to raise awareness.

Insurance: A large reinsurer now leverages AI to help its clients anticipate disasters and mitigate costs using satellite and aerial imagery to assess the damage severity for every property.

To what extent is AI disruptive for the sector?

The impact of AI on the financial sector is multifaceted, touching upon technological, regulatory, and operational aspects. We expect it to be highly

disruptive, capable of transforming business models, customer interactions, and internal operations. Large and capital-accretive business models are likely to have a competitive advantage when it comes to making the necessary investments in new technology. Conversely, companies that lack the ability to invest will lag in adopting AI and risk falling behind due to inefficiencies and outdated service models. This might drive further consolidation in the sector.

Healthcare

Authors: Lachlan Towart, Eric Potoker (US strategist)

What is the role of AI in drug research?

We explored the scope for AI to improve the drug development process in more detail in our report "TechGPT: Underpenetrated opportunities in healthtech and ASEAN new economy" (3 August 2023). In brief, two hoped-for goals are to shorten the drug discovery cycle and to improve the discovery-through-FDA-approval success rate. While biopharma companies are increasingly discussing generative AI's role in their R&D process, tangible signs that these goals are achievable remain limited.

DeepMind's AlphaFold has garnered some attention for its ability to predict the 3D structure of folded proteins, which could potentially help identify promising drug candidates to progress into clinical trials. If more molecules can be screened and eliminated earlier in the discovery process, it would lead to higher-quality molecules heading into the clinical trial funnel and presumably reaching the market. What that higher quality looks like—more broadly effective, cheaper, fewer side effects, or potentially more personalized—remains a series of guesses.

These approaches have yet to yield a significant increase in the number of viable drug candidates. The limited number of drugs advanced into clinical trials by companies with custom-built AI-driven drug discovery platforms have yet to generate published data with which to assess their clinical utility.

More prosaically, biopharma companies have adopted AI to drive efficiencies in their R&D operations—for example, speeding up patient recruitment processes or shortening the time required for regulatory interactions. But this does not represent a step-change in productivity, in our view.

If AI becomes a standard research tool in the pharma industry, we expect its costs to be absorbed within existing R&D budgets (broadly, 15–20% of sales for most established pharma companies). One potential implication is a shift in the benefit of scale in the industry. Currently, most observers agree that returns on pharma R&D do not scale well. But if generative AI can drive significant efficiency benefits, it would likely favor larger players with more budget to spend on technology. For example, Johnson & Johnson has said it has hired 6,000 employees for data analysis. However, given the 10–15-year drug development cycle, any changes for the industry, or even individual companies, will take years to play out or even to become visible in larger companies' pipelines.

Can AI open up new revenue opportunities in healthcare?

Even if a more efficient drug R&D process can be developed/enhanced with AI, it is unlikely to lead to significant new revenue opportunities at an industry level. While drug costs are only 10–15% of healthcare spending, the market's growth is generally constrained by payer budgets and pricing power is weak. AI is unlikely to change this.

Similarly, specific AI-linked opportunities in most other healthcare subsectors are also hard to identify. One area of promise is healthcare services. Today, healthcare provision is quite inefficient, care is delivered inconsistently across providers, and quality remains an issue. GenAI should facilitate better use of data to determine individualized patient diagnoses, provide treatment algorithms, improve quality of care and outcomes, and reduce cost (see our series of reports on Longer-Term Investments: Healthtech). Many companies have emerged to tackle these problems over recent years, although their presence among listed equities remains limited.

One question is, who will ultimately benefit? It is likely that over time any benefits from reduced cost would accrue to payers (government and employers), although some could filter down in the form of wages (lower employee insurance premiums), additional benefits, or even lower taxes (lower than they might otherwise be, not necessarily lower on an absolute basis) depending on the structure of healthcare funding in various countries.

Medtech companies have been incorporating technology, including AI, into their products for several years (e.g., reading diagnostic images). Thus far, the industry has seen little aggregate pricing benefit for its effort. The potential to collect more patient data and use genAI to interpret that data in real time could help to support providers' attempts to achieve faster and more accurate diagnoses, improving the utility of devices. Again, however, the extent to which these add-ons will drive incremental revenues for the industry as a whole, rather than driving competitive shifts within it, remains unclear.

Impact on pricing power/industry structure?

To the extent genAI eventually leads to more effective and more targeted therapies, the price per person treated could increase as specific medicines are better targeted for those patients who would benefit, while patients who do not benefit would be screened out before receiving that same medicine. The idea of value-based pricing—drug companies going "at-risk" for the benefits of their therapies on a patient or a population basis—could be facilitated with more personalized medicine, which itself could eventually be a result of more precise research tools and drug development, such as genAI. Such a change could lower the aggregate industry volume, as a higher per patient price is offset by lower volume.

The broader risk is if genAI truly democratizes biopharma R&D, reducing competitive moats, shortening product lifespans, and ultimately lowering terminal growth rates and values.

Barriers to adoption?

In terms of barriers to AI adoption in healthcare, it is already a highly regulated industry. In the US, for example, HIPAA limits use of patient generated data. Control over patient data remains an open question. In addition, regulators could increasingly view the aggregation of data from an anti-trust/ competition perspective, not wanting data to be aggregated and controlled by a small subset of industry players.

Conclusion

While overall the sector is unlikely to be an early beneficiary of generative AI, we think this is broadly understood by investors and see little current value ascribed to most large-cap healthcare stocks, and in particular to drug stocks, for their nascent AI efforts. For investors to ascribe durable platform value over and above the value of the drugs in the pipeline, companies must deliver proof that AI-driven R&D can deliver clinical trial outcomes superior to those of traditional research, in our view.

Case study: Johnson & Johnson

Diversified healthcare therapeutics company Johnson & Johnson (JNJ) has been among the leaders in AI spending within the healthcare sector. As reported late last year in the Wall Street Journal, JNJ has hired over 6,000 data scientists and digital specialists to leverage its internal data and drive machine learning and artificial intelligence efforts throughout the organization. As context, JNJ employs approximately 150,000 employees globally, and we estimate that those 6,000 hired employees represent approximately USD 1bn in annual investment by JNJ. Within this effort, JNJ leverages its own proprietary data, which it has amassed in an internal database named med.AI, containing over 3 petabytes of proprietary data. JNJ employees use that data throughout the entire enterprise.

JNJ cites multiple areas where this data can add value, both internally within JNJ's own product development efforts and externally to help JNJ's customers. Internally, JNJ hopes to eventually use its data to improve the success of its drug development efforts. While early days, JNJ hopes AI tools can shorten drug development timelines and eventually result in more effective drug therapies. JNJ also works with its hospital and physician customers to find superior treatment protocols—better use of diagnostic tools, shorter less invasive medical procedures, and more efficient management of provider settings.

Industrials

Authors: Alexander Stiehler, Nathaniel Gabriel (US strategist)

The industrials sector combines a wide range of industries, from airlines to automation equipment, ships to HR recruitment firms, and the aerospace and defense industry, to name a few. Therefore, the impact of AI will affect the industries within the sector differently. But a general conclusion is that the industrial sector comprises some of the most innovative industries and that AI will find its way in many business models, either supporting their topline with new product innovation or bottom line through reduced costs. Overall, we expect the impact on the sector to be positive. In our analysis below, we focus on the most relevant subsectors where AI could have the biggest impact.

Capital goods sector

In the capital goods sector, companies have made use of AI for many years to process reams of data generated from equipment with embedded sensors, predict maintenance intervals, operate "digital twins" of high-use assets, and optimize manufacturing processes. These applications have enabled higher factory uptime and reliability for customers, with the additional value shared between the user and equipment provider. We believe there is more runway for AI-driven data analytics to facilitate further improvements in customer outcomes, as well as develop new products with compelling value propositions. One example is John Deere's See & Spray technology, which is a crop sprayer that can attach to an autonomous tractor and uses mounted cameras and machine learning to spray herbicides precisely where weeds are detected, reportedly reducing herbicide use by 77%. Other examples include robots where embedded AI helps on several fronts; with generative AI, for instance, they can be more easily installed and adjusted for new applications with natural language instead of needing specialized programming skills. Like other applications, robots will also have predictive maintenance to avoid downtimes, which is very important in mass manufacturing. The embedded AI could also help robots make autonomous decisions and interact more naturally in their area.

The next leg of development is likely to streamline internal productivity in areas such as back-office tasks, paperwork, and product design. As with the advent of enterprise resource planning systems, we would expect new costs to set up and validate AI-related infrastructure, a learning curve to understand which tasks are best suited, a period of margin expansion, and ultimately, the benefits to be competed away through sector-wide adoption of tools.

One application where industrial firms benefit already today is data center infrastructure spending. Several industrial firms globally have exposure to this sector, where AI plays an important role for new data center spending. UBS expects 15–20% growth in data center equipment demand for electrical firms in 2024/25 and 10–15% in the medium term. The data center value chain is split into three parts: 1) the grey space, including power-related equipment (e.g., back-up/generator, uninterruptable power supply) and low-and-medium voltage equipment (e.g., switchgears, distribution transformers, filters); 2) the white space, otherwise known as the server room (e.g., PDUs, racks, cords, connectivity equipment); and 3) auxiliary, including the management (e.g., power, thermal, cable) and cooling (e.g., HVAC [heating, ventilation, and air conditioning], air, liquid).

In sum, we expect the impact of AI to be positive for the sector. We think it will create opportunities for higher sales through more value-added products for their customers, and on the cost side, it should help firms to improve processes.

Case study: Siemens

In February, Siemens announced a new version of their Senseye Predictive Maintenance solution, which will include generative AI functionality. According to Siemens, it makes the tool more conversational, improving human-machine interaction, and helps to make the whole predictive maintenance process more efficient, effective, and faster. The system combines machine learnings with generative AI. Until now, maintenance data was analyzed by machine learning algorithms and the user received notifications for individual cases. With generative AI, Siemens claims Senseye Predictive Maintenance can screen existing data for similar cases from the past and the solution at the time to support the new case. The company says it also enables interactive conversations between AI, the user, and maintenance experts. According to Siemens, the whole decision-making process becomes easier and more efficient. Another advantage is that even if employees leave a company, their knowledge (past solution to a case) remains in the firm and can be used for similar cases in the future. Siemens estimates that their predictive maintenance solution can improve downtime forecasting by up to 85%, reduce unplanned machine downtime by up to 50%, increase maintenance staff productivity by up to 55%, and reduce maintenance costs by up to 40%. [25]

Aerospace and defense (A&D)

The A&D sector belongs to the most innovative industries with high-tech applications across their product groups. We see several angles why AI should improve the offering and reduce costs. In the defense sector, for instance, we expect AI features to be included in drone programs and other autonomous technologies and to feature in testing, simulation, and training of combat groups (land, air, sea, space; e.g., BAE System's OdySSEy synthetic training environment) or pilots (commercial and defense). The importance of AI development is reflected in the US Department of Defense's spending on AI, which grew at a 95% CAGR from 2019–2023)—whereas the overall budget grew only 5%. For the commercial aerospace sector, partially or fully autonomous aircraft could be an option longer term, which would likely materially reduce labor costs for airlines. Other AI cost reduction initiatives include predictive analytics to optimize maintenance and repair. AI could also help with product innovation—e.g., to develop lighter components for aircraft with 3D printers, a tool that is already heavily used by manufacturers. We expect a small positive impact overall on topline and bottom line for the industry in the near term, but our view on the industry is very constructive over the medium term. We expect a strong pick-up in defense spending by NATO countries after years of underspending, and the commercial aerospace sector is still in the post-COVID recovery and ramp-up phase. Thus, we see attractive investment opportunities across the A&D sector.

Case study: War and AI

The US defense industry is investing heavily in AI, which may be better suited to identifying threats and adapting to the dynamic nature of a battlefield relative to human or algorithmically derived commands. Much of this work is classified, but one of the most visible applications is in autonomous aircraft, especially in the US Air Force's Skyborg campaign. The Next-Generation Air Dominance (NGAD) program involves the development of a sixth-generation fighter platform, which is widely expected to include a high-value manned fighter plus "Combat Collaborative Aircraft" (CCA)—autonomous wingmen manufactured at significantly less cost to fulfill a variety of supporting roles. These unmanned aircraft would carry onboard AI to detect risks, analyze, and transmit data quickly across domains or participate in combat without endangering service personnel. Though details are scant, the Boeing MQ-28 "Ghost Bat" is an uncrewed aircraft expected to utilize artificial intelligence to fly missions independently or alongside manned aircraft. The first vehicle was delivered to Australia in 2020.[26]

Professional services

The support services sector is one of the subsectors where companies could be disrupted most. Many business models are based on outsourcing activities of low-paid jobs. The average labor-cost-to-sales ratio is around 47% (ratio of the largest equity listed firms in Europe). AI will most likely lead to a material reduction of the workforce. AI will also have an impact on pricing. As cost savings may be shared with their customers, it could have a deflationary impact for this industry as other technical advancements had in the past. Companies that can't cut costs fast enough will likely feel the pressure on their operating margin.

Most disrupted could be the call-center industry. Already today, thousands of chatbots communicate with customers; AI will accelerate this trend as bots become smarter. The world market leader, Teleperformance, already uses AI tools to filter 95% of the content moderation. On the other hand, AI can also create new business opportunities—as their customers have higher incentives to outsource labor-intensive services to companies with professional AI chatbots and automated decision-making.

The credit bureau industry will also likely benefit from this development. The vast amounts of data collected could be used for even more automated decisions (e.g., credit scoring) or in their internal credit risk models. The staffing industry (recruitment platforms) may integrate AI models in their candidate-matching process and use it for job description writing. We expect that the overall impact for the industry will be positive. Notably, the industry has evolved in the past with new technological advancements.

Transportation

The sector contains several subsectors with different business models and dynamics. In the logistics industry, AI could become an important tool to improve costs through better fleet management (e.g., route planning and fleet optimization, both of which would save fuel costs), better matching of freight (supply/demand), and improving the efficiency of back-office tasks

(e.g., custom clearance). For parcel delivery, AI can streamline the flow and fluidity of packages delivered through sorting facilities and network hubs while reducing error and reloading rates. Among railroads, the technology can be used to optimize railcar and container switching as well as to screen equipment and prevent safety issues. Airlines can reduce costs significantly by optimizing flights on the fly as congestion or weather develops or through better forecasting capacity. Also, on the customer experience side, AI could help to better personalize the customer journey throughout the booking process and on the flight (e.g., more tailored entertainment).

In addition, a crucial part of any transportation business model is customer service. Shippers rely on real-time updates on the status, location, and delivery estimates of their cargo to facilitate their own operations. With AI and LLMs, shippers could proactively synthesize data regarding network performance and asset location to keep customers apprised of changes or to respond to inbound inquiries that, at present, require a human to investigate and reply.

We see the industry impact as mildly positive, as firms could achieve better pricing through a more dynamic process than in the past. Combined with the cost saving opportunities, we see the industry as among the longer-term beneficiaries of AI applications.

Construction sector

The construction sector has one main difference compared to the other industrial subsectors: Standardization is much more difficult, as residential and commercial dwellings are of different shapes and sizes. AI offers the potential to improve productivity, which hasn't improved much for the sector in the past decades due to the uniqueness of buildings. Building design and building information technology (BIM) are both opportunities, where the introduction of AI could improve the efficiency of architects and engineers; AI may also help for 3D modeling. Like the other subsectors, other cost-saving opportunities could be found via predictive maintenance of construction equipment and reduced back-office costs through implementation of AI tools and automated processes. In terms of revenues, we think AI will have an overall neutral impact on the short to medium term, with some cost saving potential and improved efficiency longer term.

Information technology

Authors: Kevin Dennean (US strategist), Achille Monnet

What does generative AI mean for the sector?

We believe generative AI is the tech theme of the decade and at the moment, information technology offers the best exposure to the trend. Despite long-term disruption risks for the lower-margin hardware and IT services industries within the IT sector, strong potential upside for the higher-margin software and semiconductor industries from genAI puts the sector in a sweet spot, in our view. In particular, we believe the US IT sector has a strong first-mover advantage in genAI with above-average growth prospects—hence our most preferred view at the moment. That being said, broadening AI trends

should eventually benefit companies outside the US—particularly companies exposed to the AI supply chain in Asia and Europe. In summary, AI should provide a strong tailwind for the global IT sector during the next few years, driving mid-single-digit percentage earnings growth prospects.

What are the new semiconductor AI opportunities?

The strong rally in semiconductors over the past year has made it the largest industry within global IT. For semiconductors, AI has pulled forward a significant expected investment cycle into computing power—especially spending on GPUs, whose parallel processing is a key enabler of AI training and inferencing. We continue to see a long runway for AI compute as hyperscalers, enterprises, and sovereigns all invest in their AI computing infrastructure—benefiting logic semiconductors, one of the largest sub-segments within semiconductors. Elsewhere, we see good secondary benefits accruing to leading-edge foundries, back-end packing, and other semiconductor fields like memory. Memory seems particularly interesting to us, as the advent of high-bandwidth memory (HBM), which is "semi-customized" to the base GPU die it sits atop, implies a slight market change. Finally, with industry fundamentals likely close to a bottom, we think the semiconductor equipment industry's growth is set for a strong rebound over the next few years fueled by AI-driven spending, including surging demand for the edge-AI computing segment.

What are the new software AI opportunities?

Artificial intelligence fundamentally changes software. It changes the entire value chain from how software is produced to how end users interact with software platforms. We expect significant value to accrue to the software layer of the computing stack and forecast a USD 395bn AI applications end market by 2027. Due to AI's ubiquitous applicability to software, we believe the applications layer of the stack should see more disruptive pressure than the infrastructure side. Foundational models from the largest tech incumbents serve as "AI foundries," and these models unlock new use cases and value propositions. Today, we group the opportunities in A16's three Cs framework (copilots, companionship, and creative). Moreover, data is a key differentiating advantage, as access to high-quality labeled data enables more effective foundational training or fine-tuning. We believe this advantages the largest platforms with the most entrenchment. Overall, we see software as a medium-term beneficiary of generative AI, driven by increased monetization and recent price increases in software.

What are the new hardware and IT services AI opportunities?

As software and semiconductor firms derive incremental value in the generative AI era, lower-margin industries like hardware and IT services are at risk of disruption in the long term. However, in the near term, the benefits outweigh the risks, in our view, as genAI could trigger a new hardware refresh cycle through AI edge-computing features in consumer applications like PCs and smartphones. While IT services companies are also at risk from genAI in the long term—as many functions like coding and testing can be automated—we still see near-term opportunities for IT companies from integrating genAI and from other consulting opportunities.

Case study: From retail to law

JD.com, a leading online retailer in China, is advancing its R&D through the JD Explore Academy by leveraging NVIDIA DGX SuperPOD for developing large-scale models akin to GPT-3. The program aims to enhance capabilities in smart retail, smart logistics, IoT, and various sectors, achieving an 8-fold speed increase over previous systems, according to the company. Their 5-billion-parameter model also benefits from the NVIDIA NeMo framework, which simplifies training with pre-set hyperparameters. This initiative is helping JD develop deep learning technologies that integrate data across diverse business domains, including retail and healthcare.

Klarna, a leading fintech company, developed an AI assistant that has held 2.3 million conversations so far, or two-thirds of Klarna's customer service chats. It is doing the equivalent work of 700 full-time agents and is on par with human agents in regard to customer satisfaction scores. The assistant is more accurate in errand resolution, leading to a 25% drop in repeat inquiries. Customers now resolve their errands in less than 2 mins, compared to 11 mins previously. It's available in 23 markets, 24/7 and communicates in more than 35 languages.

KPMG's Kymchat is a conversational AI assistant that was intended to heighten workplace productivity and collaboration. However, the adoption of Microsoft's Azure Cosmos DB for MongoDB vCore further elevated the search quality on KymChat from 50% to 91%, according to KPMG. By also leveraging the Azure App Service and Azure Container Registry, KymChat is able to reliably scale based on real-time data within the business, expanding on potential use cases without workload impairment.

DLA Piper, a global law firm pioneering technology in the legal sector, worked with C3 AI to create a first-of-kind generative AI application to streamline the analysis of complex legal agreements. In just three months, DLA Piper reportedly used C3 generative AI to reduce the attorney time it takes to create over 200 point due diligence analyses of limited partner agreements by 80%.

Luxury/retail

Authors: Thomas Parmentier, Sunny Mehra (US strategist)

What does generative AI mean for the sector?

Within the consumer sector, particularly in luxury, there are some hurdles to digitalization due to the focus on tailored client engagements, artisanal craftsmanship to create premium products, and the emotional connection associated with luxury goods. While genAI won't replace human relationships, it may play a role in enhancing the consumer journey and managing the supply chain. Its disruptive impact will likely be more pronounced in online retail due to higher competition and lower consumer fidelity. In the meantime, it can also help to improve traceability and authenticity, both of which are often appreciated by a new generation of consumers eager to know how, in what conditions, and where a product was made.

What are the new AI revenue opportunities?

We see opportunities throughout the luxury and retail value chain, primarily centered on enhancing the consumer experience. Leveraging individual profiles through AI agents and self-service alternatives offer the potential to meet the increasingly sophisticated needs of consumers. Despite their limitations, AI bots can have meaningful conversations, provide personalized fashion recommendations post-shop visits, train the sales team, and analyze online consumer behavior and real-time interactions. Generative AI holds promise for luxury brands, particularly in initiatives aiming at cultivating a special relationship with the top-end consumer through hyper-personalization and anticipating needs. Moreover, in the realm of retail, genAI could offer invaluable insights for trend forecasting. This could help brands keep pace in a world of "fast fashion" where trends change extremely quickly based on the flavor of the day and driven by social media. GenAI should further the ability of brands with more direct-to-consumer channels to accelerate their personalization, increase suggestive selling, and improve relationships with the customer. On the other hand, it could hinder mid-tier retailers that count on brands to drive their traffic (e.g., department stores). Very large retailers with reams of customer data will harness that information to make the customer experience more seamless, potentially gaining even more market share.

Case study: Buying the perfect hiking boots

A large e-commerce company has rolled out a beta version of a fashion assistant powered by ChatGPT. This innovative tool enables customers to explore the company assortment using their own language and preferences, enhancing the shopping process with a more intuitive approach. The fashion assistant can recommend suitable items for various occasions or settings. For instance, if a customer asks, "What should I wear for a hiking trip in the Rocky Mountains in October?" The company fashion assistant can understand the need for outdoor-appropriate clothing, consider the weather conditions in the Rocky Mountains during October, and suggest appropriate attire for the adventure.

Case study: Amazon's Rofus

In February, Amazon launched Rofus, a genAI-powered shopping assistant, in Amazon's mobile app. It's a way for consumers to discover products and is trained on Amazon's catalog of items, customer reviews, and broader interest. It offers questions to typical questions when searching for a product: "What should I look for when buying x," "What is better for this function: x or y product," "What product goes with this product," "What's a good gift for this occasion with these parameters," etc. It's a shopping assistant that gives buying advice. [27]

Which areas may see productivity gains?

Generative AI offers potential in several areas for consumer brands. First, cost control—for example, luxury brands can employ generative AI to pinpoint cost-saving opportunities across operations, such as refining marketing expenses by tailoring campaigns based on customer purchases and behaviors. It could improve operating expenses with a better managed

marketing budget and have less "shoot in the dark" expenses. Second, better forecasting—by leveraging genAI to analyze new trends, retailers can enhance demand prediction accuracy. Third, inventory optimization can help mitigate stock issues and fine-tune production. Fourth, AI can also help optimize resources in enhancing efficiency by minimizing waste in the supply chain and advancing sustainability efforts. Access to a large set of client databases will remain key. Most large retailers can use AI in this way as well, along with brands with large direct-to-consumer businesses.

Case study: Casio

Casio unveiled an AI-designed version of its renowned G-SHOCK watch. The exterior design process was made by utilizing generative AI to support human developers. Data amassed over 40 years of G-SHOCK development was inputted into the AI system to create a 3D model optimized for various factors such as structural strength, material characteristics, and manufacturing techniques.

What are the potential challenges of AI integration?

We see several challenges for the sector, especially around privacy. By analyzing extensive datasets, generative algorithms possess the capability to extract personal information from consumers. The sector collects and stores consumer data and without careful handling, genAI could compromise the privacy of this data, raising concerns about unauthorized access and regulators might want to intervene to limit their usage. To fully leverage the value of genAI and capitalize on new opportunities, the luxury sector must ensure it possesses the proper digital talents and foster a culture to act based on the insights gained. The disruptive nature of genAI could necessitate a new way of doing things, especially in an environment where consumer demands are evolving and the consumer journey is becoming more complicated with more touch points. Such changes might be difficult to push through in a conservative industry like luxury. The retail industry is more flexible in that regard, as it has less pricing power and must adapt faster.

To what extent is AI disruptive for the sector?

Generative AI should have a limited impact on the luxury industry. Any effect would likely be felt through an enhanced consumer journey, as the number of points of contact with a firm has exponentially increased with new generations (gen Z/millennials) and digitalization, and through reduced costs. AI can also help in verifying the authenticity of items and tracing through their supply chains. It could therefore be used to help to identify responsible sourcing or counterfeit goods. For the same reason, blockchain has been an area where brands have invested lately. We could see a bigger impact in online retail, as this industry does not have the same pricing power and barriers to entry as the luxury market. Digital laggards could see lower revenues, as they would lag the latest trends in a world moving faster every day. The large retailers should be able to employ AI to optimize costs faster than the smaller players. Thus, it will likely benefit the bigger players with the resources and direct customer data—potentially further concentrating market shares in various parts of retail.

Materials

Authors: Nathaniel Gabriel (US strategist), Rudolf Leemann

The materials sector is comprised primarily of companies that produce and process raw or intermediate materials. On the whole, we think the impact of AI will be marginal for revenues and modestly positive for cost productivity, though benefits could be competed away relatively quickly in commoditized end markets.

Chemicals

We think there will be a neutral effect broadly on the chemicals subsector. Several chemical companies already use AI to streamline prototyping and production processes, but its sophistication and use cases could improve further. This may decrease research and manufacturing costs for a time, but structural gains could be limited as industry players jockey for market share. The largest participants are most likely to benefit from their greater scale and lower cost bases.

Downstream companies, such as those in agricultural or consumer chemicals, could utilize AI to identify promising chemical compounds. Nonetheless, this is more likely to shift market share than move the needle on aggregate demand, and we expect little impact on industry revenues.

Case study: Dow

Chemical company Dow is using AI to more rapidly identify molecules with desirable properties, accelerating the pace of R&D and helping to meet customers' unique needs. In one example, Dow partnered with a division of the American Chemical Society to develop a system to screen through more than 200 million compounds and hone in on ideal candidates in minutes rather than weeks. The integration of natural language processing tools added new functionality to the existing database catalog. In another case, Dow's coatings business has begun leveraging an AI tool to predict the durability and corrosion resistance of a theoretical coating, improving formulations and time to market. This tool, called Paint Vision, won an Innovation award from ICIS and Artificial Intelligence Excellence award from the Business Intelligence Group in 2023. [28]

Metals & mining

We expect the impact on the metals and mining space to be largely neutral, as the commoditized nature of the product likely limits structural industry improvement. In the interim, however, some companies may benefit from new tools to identify and model commodity resources or advances in autonomous machinery. In one example, copper miner Freeport McMoRan is working with Caterpillar to fully automate haulage equipment at its Arizona mine, which it hopes will reduce idle time and safety risks to personnel.

That said, the requisite investments in computing infrastructure to facilitate the use of AI may accelerate demand for metals such as copper and aluminum. Moreover, development of these resources has been relatively subdued, and the industry is likely underinvesting, which could support a stronger upcycle as supply catches up to future demand.

Case study: Freeport McMoRan

Copper mining giant Freeport McMoRan has been collaborating with Caterpillar to equip 33 resource hauling trucks with autonomous systems at a mine in Arizona for over three years. These vehicles will use Caterpillar's MineStar Command solution, which allows both remote control and complete autonomy. The company believes this can reduce equipment idling time by 10,000 hours per year, meaningfully reducing greenhouse gas emissions. Safety is also expected to improve, as fewer employees will need to enter the open pit mine. The company is excited about the value proposition given the high costs of labor in the US and rural location of the mine.

Separately, at its copper and gold site in Indonesia, Freeport is employing autonomous rockbreakers to protect employees from hazards in the underground mine. [29]

Real estate

Authors: Thomas Veraguth, Jonathan Woloshin (US strategist)

From a landlord perspective, opportunities abound for the use of AI in real estate. But (un)known risks and cost considerations make it likely that implementation will be gradual.

Beneficiaries, beyond the obvious

Big beneficiaries of the AI boom are data centers. AI needs a high capacity of data to improve processes. This generates more data, which is mostly saved on a cloud and consume the computing power of the data centers. The demand for the data center service should therefore increase steadily. Furthermore, data centers are very challenging to manage, as there are only a few sufficiently qualified personnel who are familiar with this area. Likewise, the creation of a data center is very difficult and was recently hindered by delivery delays and regulation or the delivery of building permits.

Also, AI is increasingly being used in data centers for efficient operations and for power and cooling systems. However, the technology is not advanced enough yet to replace staff, and most operators believe it will take years before AI allows data centers to reduce staffing levels.

Overall, it is too early—beyond the obvious with data centers and towers—to determine which real estate subsectors—industrial & logistics, self-storage, single family, apartment, student housing, healthcare and medical properties, shopping centers, or retail and offices—will benefit the most. Real estate activities will likely be only indirectly impacted over the long term, and it is difficult to understand how AI may at all significantly change things. Within real estate, investment costs and supply and demand will still matter much more than any proptech/AI applications, in our view.

Companies (landlords and or tenants) that must manage rapid flows of products or customers (logistics, retail, self-storage) may try to more rapidly implement new AI solutions than those that face more stable

businesses (residential, healthcare, offices). Hence, AI applications and the implementation will likely evolve differently depending on subsectors and specific needs.

More about future opportunities:

Promotion

Other applications include writing marketing letters or real estate prospects. By entering the basic data, an AI system can produce a comprehensive text aimed at convincing prospective buyers of the property. Another advantage of AI is that it knows the relevant keywords and incorporates them appropriately into the text. In this way, the real estate company receives more views of the advertised properties. However, this may lead to conflicts with copyrights and responsibilities of the authors.

In concurrence or addition to the existing "proptech"

Very significant for the real estate industry is the dynamic between blockchain technologies and artificial intelligence. One possible area of application is smart contracts in facility management. These can help property management companies automate administrative activities while maintaining transparency across all processes and legal entities. These processes have been in place for a few years and are known as "proptech." AI will therefore likely need to be integrated into proptech applications already in place within property management. This ends up as being a question of comparative costs and advantages.

This is another reason why AI applications are likely only to progress gradually in the existing systems that have proven to increase productivity. We expect companies to continue to use proptech and AI to grow the business rather than to shrink manpower. CoStar—a supplier of real estate information and platforms for brokers—is a good example of a company that uses technology to gain market share. This shows that companies that rapidly/continuously explore opportunities within proptech/AI could have a market advantage under the condition that these innovations help save costs and grow the business.

Landlords have increased and improved their offerings by building data analysis systems for years, and they will likely continue to do so in the years ahead. Companies are already facing issues over private data, copyrights, and sources of information. For the real estate sector, we hence expect the gap between fantasy and reality to grow—not to diminish.

Will office demand grow in geographies that enjoy proptech/AI boom?

The demand for offices in suburbs where AI technologies are developed is already increasing because the AI development teams need more space. As we have highlighted, it is so far unlikely in the foreseeable future that AI will reduce space demand by tenants. It remains to be seen, but according to the companies we are in contact with, there is no real evidence that AI may accelerate the work-from-home trend and thus reduce office demand in the foreseeable future.

Case study: A successful use case

Royal London Asset Management, a leading UK investment firm, experienced significant improvements in HVAC (heating, ventilator, air conditioning) operations and energy efficiency in an 11,600-square-meter commercial office building. By implementing JLL's AI-powered Hank technology, the firm has reached a return on investment of 708% and energy savings of 59%. HVAC operations and maintenance keep mechanical systems working at peak performance during the life of the building.

Impact on jobs

We do not expect to see any major effects on labor in the years ahead. We believe companies will look at their market share and use proptech/AI to grow their business, requiring more staff (not less) in the initial stage at least. The evidence from the companies seems to confirm this. It will also take time to better understand all the implications around copy/property rights and in regard to responsibilities/liabilities in the business operations. And ultimately, AI must find its way as an extension of existing proptech solutions in terms of efficiency (benefit/cost relation).

Market predictions, companies' strategy

Overall, big companies are set to implement AI in their operations or strategic planning. Prologis, as an example, may develop tools for forecasting demand flows within markets to adapt their footprint in strong regional markets. We understand that not only landlords but also tenants will optimize their location based on the analysis of their customers' base. It remains to be seen if AI alone will really bring enough added value compared to existing systems. This is why landlords and tenants are using AI—more and more—in addition to existing systems, not to replace them. We hence expect AI to be an evolution rather than a revolution.

Utilities

Authors: Carsten Schlufter, James Dobson (US strategist), Alexander Calvert (US strategist)

What does generative AI mean for the sector?

The rising use of AI and datacenters should drive a significant increase in demand for electricity to power AI computing and facility cooling, among other things. AI data centers require large amounts of electricity, in some cases three to eight times the amount of a traditional datacenter. Across a market like the US, this could add 2–5% to annual electricity demand growth but will vary by region[30]. Though 2–5% sounds like a small impact, this is already driving electricity demand growth in a sector that has not seen material annual demand growth in well over three decades. This should boost capital investment for additional electricity generation, transmission, and distribution infrastructure. However, given the regulated nature of the

electric utility industry, the potential benefit from rising demand will likely be more limited.

Utilities are often regulated based on an allowed return on investment. Therefore, rising capital spending should drive a modest earnings benefit from AI thanks to increased revenues from adding new infrastructure to serve rising demand from AI datacenters.

Unregulated generation companies should benefit more directly from rising prices for electricity and potential demand and contractual supply opportunities from AI datacenters. Given rising demand, limited supply, and the long lead times for additional baseload electricity generating capacity, AI datacenter developers and operators have an incentive to aggressively pursue contractual access to electricity supply.

What are the new AI revenue opportunities?

Unregulated generators are likely to see the most significant benefit in the electricity industry. By selling electricity into the wholesale market, unregulated generators could see higher prices and greater demand for their production. Though largely focused on existing generation that is dispatchable (nuclear- and natural gas-driven electricity supply, for example), AI datacenters are also likely to accelerate additional renewable energy infrastructure development. AI datacenter developers appear keen to offset some of the electricity demand with renewable power. This should drive development opportunities for renewable power developers.

For regulated utilities, the advancement of AI should drive additional electricity sales, but the benefit to earnings should come from the investment in infrastructure to serve the new electricity demand. Regulated utilities will also likely benefit from accelerating development of renewable energy. Though a more coincident supply (sun and wind power), renewable energy supply can partially offset growing demand. A global trend toward decarbonization and renewable energy expansion, which is expected to intensify in the coming years, should be positive for utilities' revenue growth. However, the growth in electricity demand and the expansion of renewable energies also require additional transmission investment. Therefore, companies operating in the transmission and distribution of electricity should also benefit from AI growth. However, it is likely to be limited to the countries or regions in which data centers are built.

Which areas may see productivity gains?

The impact of AI on utility productivity is limited but could aid in more efficient grid operations and dispatch. Preventative maintenance and data analysis are also likely to be areas of focus. We expect utilities to continue looking for ways to improve operations through AI.

The impact of AI on the profit margins of utilities is likely to be limited. Since many utilities are regulated on a rate of return methodology, lower operating costs and the financial benefits of more efficient operations would likely be passed on to customers. However, the lower operating costs increase the potential for capital investment, which drives earnings growth for utilities. This should, over time, drive utility productivity gains through the use of AI.

Since regulated utilities operate as monopolies and implementation must be approved by regulators, they are likely to be late adopters of AI.

Case study: Powering datacenters

In March 2024, Amazon announced an agreement to purchase land directly adjacent to the Susquehanna nuclear plant in Pennsylvania, which is owned by Talen Energy. The real estate purchase from Talen included existing plans and infrastructure for a hyperscale datacenter campus and a multi-year contract for up to 960 gigawatts (GW) of electricity capacity directly from the adjacent nuclear plant. The purchase power agreement could last up to 30 years. The land and power would allow Amazon to develop several large datacenters. The electricity purchase price was not disclosed but was likely a premium to the wholesale price and a modest discount to the retail price.

Case study: Go inspect, Sparky!

US-based energy company Avangrid, which is a member of the Spanish Iberdrola Group, recently announced a pilot project with Levatas and Boston Dynamics to advance substation inspections using AI. This project should help the company to be more efficient, with the goal of increasing reliability. It will deploy a mobile robot called Spot, modeled after a dog, to complete visual and thermal inspections at two substations.

According to Avangrid, the pilot project will take place at two substations and test a variety of AI models, developed by Levatas, to read analog gauges, record thermal images, and detect damaged equipment. To do so, the robot dog—nicknamed Sparky—is outfitted with a camera. The project will also test how quickly and accurately the robot can detect and read several of the substation's analog gauges. According to the company, the project will also test the robot's ability to inspect transformers, circuit breakers, and capacitor banks. [31]

Chapter 5: Beyond borders: AI's resonance in jobs, inflation, and sustainable development

Authors: Paul Donovan, Dean A. Turner, Yifan Hu, Brian Rose (US economist)

The latest iteration of artificial intelligence closely follows most of the economic patterns of previous technological developments. However, the context in which artificial intelligence is taking place—changing demographics and an evolution from output to impact economics—may add some peculiarities to the ways in which economies react.

Productivity

Nearly all technological change is introduced because it increases economic efficiency. There is no point investing in new technology if that technology does not allow the economy to do more with less. The importance

of efficiency may change over time, changing the desire to adopt new technologies—the concept of a steam engine was understood during the Roman Empire, but was not an (economically) efficient production method in a society with plentiful slave labor and limited fuel. By the eighteenth century, the potential for efficiency gains from steam power in a country like the United Kingdom with relatively abundant coal supplies was significant—and productivity increased.

The principle of introducing technology is therefore that it will enhance productivity. There are two caveats that apply to this: whether established business practices adapt to the new technology; and whether the resources for the new technology are available to allow productivity to be maximized.

Sometimes technological changes are small—economists call this iterative innovation. With such changes technology will increase productivity, but it is unlikely to be significant. Very often the change is an improvement on existing technology—despite what manufacturers claim, upgrading to the next model of smartphone is unlikely to significantly change the user's economic potential. More dramatic changes are classed as dynamic innovation, and the productivity gains from these are likely to be large.

Dynamic innovation tends to require greater upheaval in the economy, because the new technology changes working practices. This can limit the productivity gains in the short term. The productivity of the second industrial revolution with its use of electric power was famously hindered by the legacy of steam power from the first industrial revolution. Initially, electric engines were substituted for steam engines, with a single engine powering multiple machines in a single factory. The productivity gains from this were minimal. The real productivity gains of the electric motor only started to become significant when this centralized power method was abandoned, and smaller individual motors powering specific pieces of equipment became the norm. Tearing up the old business model to start again required both confidence and innovative thinking and took years to implement.

Artificial intelligence seems to be a mix of iterative and dynamic innovation. Some of the changes and productivity gains from artificial intelligence seem to be relatively small. However, there is the potential for more dramatic improvements in productivity in some areas of the economy if established business practices change.

The second caveat is the assumption that the necessary resources to make use of the new technology will be readily available. The Romans' lack of fuel denied them the productivity gains of the steam engine. During the 1950s, a lack of male computer programmers meant that Britain's significant competitive advantage in computing failed to enhance national productivity (there were many trained programmers from wartime computer operations but these programmers, being women, were not considered employable in peacetime).

Resource availability is a significant issue for productivity associated with artificial intelligence, which relies on access to data. Companies are already rushing to protect their data from being accessed by large language models. This deliberately creates inefficiencies to protect proprietary information.

The consequence of data restrictions for productivity is not clear. It may be that a two-tier system develops: a broad-based artificial intelligence based on public but inferior data; and a superior artificial intelligence using proprietary information that has a limited user base. The question is whether attempts to protect proprietary information might undermine economic productivity that existed before the growth of artificial intelligence.

Growth

If artificial intelligence increases productivity, it naturally should lift economic growth. This is where the context for artificial intelligence becomes significant. Efficiency should lead to improved standards of living, but growth (in the output economic sense of growth in gross domestic product) is not the same thing as living standards.

Different sectors of the economy will experience different growth reactions from increased efficiency. The straightforward impact from efficiency is the potential for higher output with the same or fewer inputs. But technology also has the potential to lower barriers to entry for different sectors, allowing output to rise by diverting resources to that sector. The advent of music streaming (replacing the purchase of music in a physical format) significantly lowered barriers to entry. The removal of cost barriers also lessened the artificial barrier of "gatekeepers" (record labels deciding what music was commercially viable). With limited expense and a reasonable social media following, any musician can now make an impact where previously they would be excluded from the marketplace. The result has been an increase in the volume, diversity, and quality of music readily available.

Almost half of officially reported global GDP is currently generated in countries with declining populations. The output economics of GDP is a simple combination of the number of workers and how productive those workers are. If productivity is rising as a result of artificial intelligence, but the number of workers is falling, the role of artificial intelligence may be to limit the slowing of GDP growth rather than accelerating GDP growth.

Greater productivity or efficiency also does not have to increase output. The global sustainability crisis implies that the priority for efficiency is to lower inputs required to achieve current standards of living. This type of improvement is not captured by GDP. The severity of the sustainability crisis, its perception by consumers, and shifts in the relative costs of inputs are likely to tilt the balance between increasing output and reducing inputs.

Inflation

If efficiency means that we can do more with less, then economic costs should decline if artificial intelligence increases productivity. Lower costs in the production chain will mean lower consumer prices or increased profit margins—but if the latter, then either competition or (potentially) regulation should push some of the lower costs through to the consumer. The net effect of any technology-led efficiency is normally disinflationary for an economy.

Beneath that overall disinflation story, technological change tends to create a lot of relative price shifts. Some prices are likely to rise as the economic restructuring wrought by technological change alters patterns of demand. Thus, in the first industrial revolution, the automation of spinning reduced the price of thread, but then increased the demand for and price of weaving (as there was an abundance of thread but—until automation—a limited capacity for weaving that thread into cloth). Similarly, the need to build factories increased the demand for and price of bricks, again until automation brought the price back down. Other prices, seemingly unrelated to the new technology, will fall. In the 1920s, the rise of the automobile in the United States slowed demand for shoes and boots.

The wider adoption of artificial intelligence will likely increase some prices, in absolute or relative terms. This is most obvious in the demand for the technological infrastructure required to run artificial intelligence, but the shifting demand patterns caused by changing barriers to entry mean that many other prices are likely to fluctuate. Energy demand might rise to power the new technology. That would raise the price of energy as other prices decline. Of course, those prices that do rise may then incentivize additional technological advances, eventually bringing down prices again. The pattern of short-lived price booms on demand reversing as new technologies with new efficiencies have responded has played out time after time. The macro disinflation trend disguises micro economic price turmoil.

Artificial intelligence has the potential to offer a specific important change for pricing. For at least the past century, retailers have been at an information disadvantage. Urbanization and large-scale retail meant that shop owners lost personal knowledge of their customers. Before the mid-nineteenth century, a shop owner would have intimate knowledge of their customers—and that meant that they would be able to form an idea of the maximum price that each individual customer would pay. This allowed for personalized pricing —two customers being charged different prices for the same item. The farm laborer paid less than the local estate owner for the same good. Once urbanization and large-scale stores took over, customer anonymity grew, and shops started charging a single price to all their customers (the law of one price).

The ability of artificial intelligence to manage large amounts of data has the potential to remove that customer anonymity, if combined with some basic identification system (cookies with online retail, or a customer loyalty card). Personalized pricing—if allowed to develop—means that more of the benefits of productivity are collected in retailers' profit margins. It also makes the concept of inflation, as we understand it, increasingly difficult to measure, because no two consumers are certain to be charged the same price.

It should be noted that while personalized pricing is most controversial at a consumer level, there is no reason why it cannot occur higher up the supply chain in producer prices.

Employment

All technological change brings with it a fear that machines will take over and leave humanity unemployed and unemployable. From the smashing of machines in the first industrial revolution, to computer manufacturers

advertising how their technology could replace lazy staff in the 1970s, the threat of mass unemployment has always been bound up with technological change. And yet, as we embark on the fourth industrial revolution, the number of people with jobs in the global economy is at an all-time high.

It is perfectly true that technological change will lead to specific jobs being lost. This is either because fewer people can produce the same output, or because people are simply not required to fulfil certain roles. It is best to think of a job as a series of tasks that are performed. If over half of those tasks could be taken over by artificial intelligence, then the job is likely to be lost. If fewer than half the tasks can be automated, then the job may change (which may still be traumatizing)—but someone will be employed.

Technology also creates new jobs. There are three ways in which this happens: Directly, through efficiency, and indirectly through shifting demand patterns. Any new technology directly creates new jobs. The mechanics of the first industrial revolution, the electricians of the second industrial revolution, and the computer programmers of the third industrial revolution all directly owed their employment to the introduction of new technology. Without that technology, the jobs would not have existed. Artificial intelligence is no different—it will require people trained to ask the right questions, and computers will have to be manufactured for artificial intelligence to run on.

Efficiency can also create new jobs as things become cheaper to do. In a world where media meant newspapers, there were very few "influencers" because it was expensive to communicate with large numbers of people. In modern times, being a social media influencer has become a viable career for some.

Those jobs only came into being as the introduction of new technology significantly reduced the cost of communicating with millions of people. As AI lowers the cost of accessing, processing, or presenting information, jobs will be created to exploit that.

Because technological change like artificial intelligence is revolutionary, the impact that it has on society will also disrupt the labor market indirectly through changing demand patterns. For example, if artificial intelligence reduces the time taken to perform certain tasks, it may increase the amount of leisure time some people have. Increased leisure time means more jobs in leisure-related industries. Artificial intelligence could increase the number of people employed making gardening tools, filming travel vlogs, or working as personal trainers as they service the increased demand for leisure. This has happened before.

In 1930, the British economist Keynes famously predicted that by 2030, everyone would be working 15 hours a week and no more. This prediction was not as wrong as it might appear—while time spent in paid work has not fallen as dramatically as Keynes predicted, the time spent on unpaid household chores has collapsed as a result of technology. The resulting increase in leisure time was more or less what Keynes forecast and has generated huge numbers of jobs in the leisure industry.

This means that at a macroeconomic level, artificial intelligence is unlikely to create mass unemployment in the long term. Existing jobs are destroyed, new jobs are created, and some jobs that already exist will grow in number. The

problem is that full balance is only likely to be achieved in the long term, and in the short term, the imbalance in the labor market may be more painful.

The potential for an imbalance between the skills of people losing their jobs and the skills required in the newly created jobs is made worse by the accelerated speed with which new technologies are adopted. The pandemic already sped up some of the changes of the fourth industrial revolution, and artificial intelligence has the potential to further accelerate change. Rapid change gives little time for existing workers to adapt or retrain, and risks people losing income, social status, or becoming long-term unemployed.

The short-term pain in labor markets could have serious consequences. Social status is often closely tied to a person's employment, and historically when the social status associated with a job declines—or disappears—there has been a tendency to look for a scapegoat to blame. Typically, the scapegoat will be an identified minority in society, who will of course have had nothing to do with the relative loss of social status felt by the disgruntled worker. Faced with complex challenges, people tend to crave simplicity and believing that all their problems can be blamed on a religious, ethnic, or sexuality minority —or the ever-popular groups of immigrants or foreigners—is a seductively simple and comforting explanation. Scapegoat economics can quickly move to a culture of prejudice and prejudice politics, offering ineffective but simple "solutions." Political leaders curry favor with the disgruntled population by promising to restrict the actions of a group in a society, as can already be seen in several political systems around the world.

The economic damage of prejudice is potentially devastating. In a world of rapid change, flexibility is required. The mantra of the right person, the right job, and the right time is the hallmark of economic success.

Prejudice prevents the right people being employed, by using irrational criteria to restrict (or worse, dehumanize) key workers. To offer an example, artificial intelligence will not produce effective economic results if the person with the best ability to ask the right question is prevented by prejudice from being employed. As the earlier computer programmer example demonstrates, the gender prejudice of British society 70 years ago had a similarly devastating impact on productivity.

Changes to the shape and structure of the labor market are likely to be the key economic challenge posed by artificial intelligence. In a direct sense, artificial intelligence should be productivity enhancing to a greater or lesser degree depending on the balance of iterative and dynamic innovation it inspires. That suggests improving living standards, appropriately measured, and at least some macroeconomic disinflation. But while in the long term there should still be enough jobs to occupy the working population, the revolution in social status may bring about a more polarized society. A polarized society could slow the introduction of new technologies, resulting in less efficiency and overwhelming some of the direct economic benefits that artificial intelligence could produce.

Regional perspective:

US

The US has the potential to benefit from the development of AI by more than other countries. Many of the world's largest tech companies are headquartered in the US, and it also boasts leading academic institutions producing AI-related research. These institutions are able to lure some of the best and brightest students from overseas who have historically played major roles in the development of the tech industry. The US also has the most developed and robust venture capital industry, which should lead to continued funding of startups. This has been an important advantage historically, as new technologies tend to require significant private sector investment.

The regulatory environment is also likely to remain more favorable in the US compared to other advanced economies. For example, some local governments have approved the use of robotaxis, providing a valuable opportunity for developers to gain experience and collect real-world data. More generally, rules on how private data can be exploited for profit are relatively lenient, a key factor in improving AI models.

Regarding the physical manifestation of AI in devices including smartphones and robots, the US is likely to provide a lot of intellectual capital in product design. However, we would not expect a big boost to domestic manufacturing output, which has been on a sideways trend for the past decade.

In our view, AI could boost productivity growth. We estimate that for some jobs, this could be around a 0.5% per annum increase over the medium term, perhaps even more if there are breakthroughs in some key technologies. However, there could be a negative impact on total hours worked if AI disrupts labor markets, potentially reducing the overall impact on GDP growth.

Europe

Europe, like other economic blocs, has the potential to reap the benefits of AI, which could result in higher economic efficiency. Given the troubling demographic trends in Europe, where working-age populations are shrinking in several key economies, any increase productivity growth through better economic efficiency as AI technologies spread through the economy may merely mean that the trend rate of economic growth falls at a slower pace, rather than increase it from its current levels.

Thus far, Europe has lagged others in terms of development of AI technologies, with the US and to a lesser extent China leading the way. There is no single explanation as to why this is, but one could be the fragmented nature of Europe's markets. Unlike in the US, European companies operate in a region that lacks a level regulatory playing field, which can result in more bureaucracy. Moreover, although the total size of the potential consumer market looks not so different from the US (perhaps even larger depending on the definition), the lack of a common language, cultures, and legal environment makes addressing this market more costly. Moreover, Europe's lack of a unified capital market makes funding and developing new research comparatively more costly. This does not mean that innovation will not happen—just that the hurdles are sometimes higher.

Another factor that may determine the speed at which AI technologies are deployed in the workplace is the structure of employment on the continent. Higher levels of government spending in Europe mean that a greater share of jobs (roughly one in four) is in the government sector than in the US (closer to one in eight). Private sector companies tend to have a greater economic incentive to introduce productivity-enhancing technologies to boost profits, something that is rarely a focus of the state sector, which could result in a delayed adoption across the economy than in other regions. Nevertheless, as new technologies become mainstream, the wealth of data held by the government sector may eventually mean that they can utilize the AI more effectively.

Where Europe is likely to lead is in regulation. Europe is the first region to introduce legislation around AI with the hope of encouraging innovation while protecting consumer rights. Time will tell if this delicate balance is achieved, but a bigger question is how far such legislation will extend beyond Europe's borders. General Data Protection Regulation legislation is a good example of where Europe's rulebooks can extend beyond borders, which it could dampen some competitive disadvantage that Europe's innovators are under. AI legislation could have comparable results, but either way, it seems that Europe's desire to focus on regulation and protections will place some constraints on innovation and, perhaps, implementation in the near term.

China

Over the last decade, China has been quick to embrace the development and adoption of AI, particularly in consumer-facing industries with key players using new technologies to personalize the end-user experience. Key factors driving these developments are China's market size and the regulatory environment, both of which have been conducive to the adoption and development of AI. A supportive funding environment for research and development from both the government and private sector has also been instrumental.

The Chinese government has laid out ambitious plans to lead the world in AI by 2030, as it believes it can facilitate the transition to a more consumption-led economic model. This should mean that the funding and regulatory environment remains supportive, and these are likely to be complemented by policies that encourage AI adoption across various industries.

Notwithstanding Beijing's ambitions for AI, China is facing growing headwinds from the US where the government is leading a multinational effort to tighten curbs on Chinese companies' access to the frontier technology and talent. Moreover, although data privacy concerns haven't been a major consideration for policymakers yet and there haven't been significant concerns raised by the majority of citizens, this may change in the future. In addition, the government's desire to censor content could also slow progress in the development and implementation of AI technologies. A greater emphasis on boosting indigenous technology in bottle-necked industries (i.e., chips) and basic research could lead to fragmentation of the global AI infrastructure, resulting in less economic inefficiency that affects all countries.

What does AI mean for sustainable development and the SDGs

Authors: Amanda Gu, Antonia Sariyska

The United Nations created 17 Sustainable Development Goals (SDGs) in 2015, calling for action on a range of topics—social, economic, and environmental sustainability—through 169 underlying targets by 2030. Despite an increase in SDG investments, the UN finds the annual SDG investment gap has widened to USD 4tr, from the previous funding gap estimate of USD 2.5tr in 2015. This increase is partly attributed to the effects of various global challenges, including the COVID-19 pandemic and the food and energy crises. To address this shortfall, working to achieve the SDGs has gained much attention in public and private domains, and investors and businesses are tapping into opportunities driven by sustainability objectives globally.

As an enabling technology that helps deliver sustainable solutions and products, AI can create new pockets of opportunity to support sustainable development initiatives. A study finds that AI may enable almost 80% of targets across all SDGs through technological advancement to overcome present limitations[32]. This is significantly more than the 35% of targets that can be negatively impacted by AI because of varying cultural values, wealth dispersion, and potential overexploitation of resources in given areas to fuel increased productivity.

Recognizing the interwoven nature of the SDGs, we highlight areas where AI integration is progressing, and while still in its nascency, how AI has the potential to help advance the SDGs. We discuss some of these applications throughout our report but find it important to summarize the most critical positive applications here.

Climate change: AI can help accelerate the discovery and development of green materials, resulting in reduced carbon emissions and decarbonization costs. Enabling technologies and digitalization can drive power grid optimization, climate modeling, smart agriculture, and data center energy use optimization to ensure effective and tracked progress on the decarbonization journey. This is coupled with improved applications in waste management and the broader circular economy.

Healthcare and education: Using AI in diagnosis may reduce treatment costs by up to 50% and improve health outcomes by 40%, according to Harvard School of Public Health[33]. Appropriate application of AI can address affordability and improve access to healthcare. Large language models combined with superior computing power can accelerate drug discovery and bring down drug innovation costs. AI's diffusion in education is happening, through the adoption of AI chatbots and other AI-integrated tools for educators and students. ChatGPT has generated considerable interest among students, with the potential to enrich their learning experiences. Schools and government institutions need to be prepared for such structural change,

for example through adjusting education curriculums, reconsidering business models, and tailoring to students' needs.

Labor: Technology breakthroughs historically have improved workers' productivity and boosted economic growth. AI can reshape industries, shifting from labor-intensive to capital-intensive processes. Its benefits can be fully maximized with responsibilities such as reskilling employees, embracing experimentation, and adapting to new forms of working. Currently, 40% of global employment is exposed to AI, according to an IMF analysis[34]—of which developed markets (DMs) are more exposed (about 60% employment) and better poised to reap AI benefits than emerging markets (EMs). Older workers are potentially less adaptable to new technologies. Structural differences between DMs and EMs and between younger and older workers indicate varying levels of exposure to the benefits and pitfalls of AI. Seizing AI's benefits unequally may result in a widening of the digital divide and global income disparity. Successful AI integration across sectors and verticals warrants holistic considerations of regulatory policies and best-practice guidance. At a company level, we believe that companies can differentiate through harnessing changes and creating a diverse, balanced workforce for the AI evolution to unfold. Governments and supervisory bodies are key to updating regulatory frameworks and supporting labor reallocation while safeguarding those adversely affected.

We think AI will continue to increase transparency and accountability, both of which are critical to sustainable development, and specifically to SDG 16 on Peace, security and strong institutions. However, this needs to come with adequate governance and regulation, given geopolitical tensions, the rise of misinformation (and the so-called "deep fakes"), and the vulnerability of digital systems.

As with every technological development, we also need to remain conscious of the inherent differences in the speed of deployment, particularly across developed, emerging, and frontier markets. Many new applications require vast upfront capital expenditure, which might lead to further economic inequality between regions and economic systems. The role of public-private partnerships and development finance is particularly critical, in a similar fashion, as it is essential to critical sustainable infrastructure deployment such as renewables, internet broadband, and financial systems. We think private investors have the opportunity to facilitate some of that critical financing through partnering with multilateral development banks (MDBs), such as the World Bank and its affiliates, by investing in MDB bonds, whose proceeds facilitate broader sustainable development and the adoption of new technologies in underserved regions.

Chapter 6: Geopolitics and regulations

Geopolitics: US-China rivalry and the new global tech order

The global stage is witnessing a shifting landscape, with a splintering of investment and supply chain routes. The once unipolar order is transforming into a multipolar landscape, underpinned by the emergence of China as a challenger to the long-standing dominance of the US. The last few years

have seen a significant deterioration in their relationship, and a long-term resolution seems challenging given their domestic political environments, in our view.

This structural shift brings nuanced and profound long-term implications across various areas. Given the deep economic and trade connections between the US and China, a complete decoupling seems unrealistic. However, both sides are likely to continue distancing their relationship on several strategic issues, with the technology sector taking center stage. As both nations vie for technological supremacy, a complex web of tensions unfolds, presenting unique challenges and opportunities for technology investors that extends beyond the US and China.

The tit-for-tat escalation in global tech

The epicenter of this technological rivalry lies in three key hotspots: semiconductors, artificial intelligence, and quantum computing. These intertwined, cutting-edge fields hold immense economic and strategic potential, making them prime targets for national investments, restrictions, and security considerations.

The Biden administration has openly pursued coordination restrictions on exports, aiming to limit China's access to cutting-edge chip technology and advanced semiconductor equipment. These measures, alongside others—like curbs on US investment in Chinese high-tech ventures, the CHIPS Act, and provisions in the Inflation Reduction Act that aim to lessen reliance on China's green technology—have heightened concerns about escalating tensions.

In response, China has significantly expanded its export restrictions on key materials used in chip production such as gallium, germanium, and certain types of graphite, banning technology exports for making rare earth magnets and for extracting and separating these materials. Additionally, China restricted its domestic companies from procuring certain semiconductor products from US memory maker Micron.

Although China remains a significant hub for end-product manufacturing, there's a notable trend among Chinese and international companies toward diversifying manufacturing investments, particularly in Southeast Asia and India, along with reshoring efforts back to their home countries (the US, Europe, and other north Asian countries). In terms of international relations, the US, Japan, South Korea, and Taiwan formed a consultative alliance dubbed "Chip 4" to foster coordination and cooperation within the semiconductor industries, posing even larger development challenges for mainland China.

Despite mainland China's efforts to achieve self-sufficiency in chip production, exemplified by the development of domestically produced 5-nanometer chips for Huawei smartphones, there are still challenges on the production yields and unit economics. The latter was mainly due to the use of deep-ultraviolet (DUV) lithography, an older generation equipment compared to the more cost-efficient extreme-ultraviolet (EUV) lithography machine. This highlights challenges behind its high-end chip manufacturing, design, and equipment production.

Conclusion: The US-China tech rivalry is centered around three key areas: semiconductors, artificial intelligence, and quantum computing. Despite mainland China's efforts to achieve self-sufficiency in chip production, there are still challenges given its restricted access to advanced semiconductor equipment.

The geopolitics of AI

AI is poised to be economically disruptive. While certain jobs may become more valued, others are at risk of becoming obsolete. Such periods of economic upheaval often breed social tensions. Those experiencing rising status and income may not voice complaints, but individuals witnessing a decline in their income and status may seek someone to blame, especially if they perceive their neighbors as ascending while they are descending.

The cause of this relative social change is intricate, yet people tend to seek simplicity, which can lead to scapegoat economics. This phenomenon involves wrongly identifying a societal group as the cause of a declining social status, potentially fueling prejudice politics. Geopolitically, throughout history, scapegoat economics often targeted foreigners, as they are perceived as outsiders, thus fostering a particular brand of prejudice politics known as economic nationalism.

In an increasingly polarized world, we anticipate AI will exacerbate geopolitical tensions, as evidenced by recent action such as the US imposing export controls on advanced chip shipments. In October 2023, the US government expanded the export control list to include less advanced chips like the L40S and A800, in addition to the restrictions already imposed on the H100 and A100 in the previous year. In response, NVIDIA plans to launch new chips such as the H20, L20 and L2 to meet the new requirements, with the H20 slated for mass production in 2Q24. These new chips include NVIDIA's latest features for AI-related computation, though its total processing performance (TPP) and performance density are below the compute restriction map.

Conclusion: Given that the generative AI industry is still in its early stages, further escalation is plausible during periods of geopolitical flare-ups. We advise investors to manage these risks accordingly.

Quantum computing – will geopolitics hinder the growth of quantum computing?

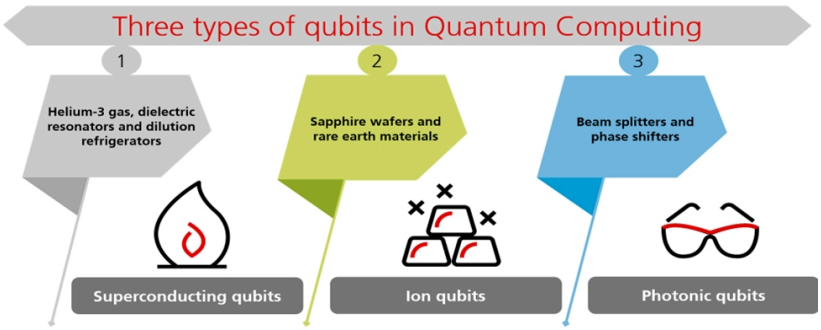
Before addressing this question, it is imperative to understand the difference between traditional computing and quantum computing. Unlike traditional computing, which is developed around the laws of mathematics, quantum computing is built on physics. With traditional computing, information is stored in bits with two states: 0 or 1. With quantum computing, information is stored in qubits (or quantum bits) that can be in any state between 0 and 1. In other words, rather than just being either 0 or 1, qubits can be in what's called "superposition," when they're both 0 and 1 at the same time, or somewhere on a spectrum between the two.

Here's an easy way to think about it: When you flip a coin, it will either land heads or tails. But if you spin it, the coin can be both heads and tails while

it's moving. Binary computing is the flipping of a coin, and superposition the spinning coin; the latter is particularly useful for simulations. While current quantum systems with only a few hundreds of qubits lack the power to perform complex tasks compared to an ideal requirement of a few thousand qubits, recent progress suggests we may reach that stage in a few years.

Though current application remains niche, recent developments suggest it could become the next focal points of geopolitics, akin to the recent attention on advanced chips. Media reports highlight China's increasing involvement in quantum computing and the potential for trade restrictions, underscoring the importance of understanding the supply chains and major players in this field.

Figure 29 - Segments of quantum computing



Source: UBS, as of 2024

Firstly, there are superconducting qubits, which are mainly used by leading companies like Google and IBM. A superconductor is a material that changes from a normal state when it is cooled to a superconducting state where there is essentially no resistance to the flow of direct electrical current. The advantages of superconducting qubits are faster computation and better integration potential with current circuit processes, while disadvantages include quick decoherence and the need for a very cold environment for them to work. The key components of the superconducting qubit supply chain include Helium-3 gas (mostly from North America and Russia), dielectric resonators (mostly from Germany and Japan), and dilution refrigerators (mostly from Finland, the Netherlands, and the United Kingdom).

The second type is trapped ion qubits, which are mainly used by companies like Ion Q, Alpine Quantum Technologies, and reportedly, Foxconn, the leading electronics assembly company in the world. This technology traps ions using magnetic fields, with the key advantages being stability and the ability to operate at room temperature; the major disadvantage is that trapped ion qubits are considerably slower compared to superconducting qubits. The key components of the trapped ions qubit supply chain include sapphire wafers (mostly from Japan and Russia) and rare earth materials (mostly form China and as byproducts from nuclear reactors).

The third type is photonic qubits, which are mainly used by companies like Xanadu. This technology uses particles of light to carry and process information and manipulate photons with mirrors, beam splitters, and phase shifters. The key advantages include the ability to operate at room

temperature, whereas the main disadvantages include the lack of scalability. The key components of the supply chain include beam splitters and phase shifters (mostly from the US).

Figure 30 - Traditional computing vs. quantum computing

Traditional computing	Quantum computing
Use electric circuits which are in a single state at any given point in time: on or off	Uses quantum circuits that can be in more than one state at any given point of time
Runs at normal temperatures	Runs at extremely low temperatures
Information storage is based on voltage	Information storage is based on direction of electron spin
Processing of information is carried out by logic gates in a sequential basis	Processing of information is carried out by quantum logic gates in a parallel basis
Conventional bits store limited amount of information and consume more energy	Quantum bits can storage an enormous amount of information and use less energy
Circuits interface condition is stable	Circuits are increadbly sensitive to interface
Results are specifically defined, limited by algorithm design	Due to superposition and entanglement, answers are probabilistic in nature
No restriction on copying or measuring singnals	Encryption has high degree of restrictions on copying and measuring signals

Source: UBS

Conclusion: Despite limited applications in quantum computing so far, we believe it could become the next focal point of geopolitics, underscoring the importance of understanding the supply chains and major players in this field.

What's next?

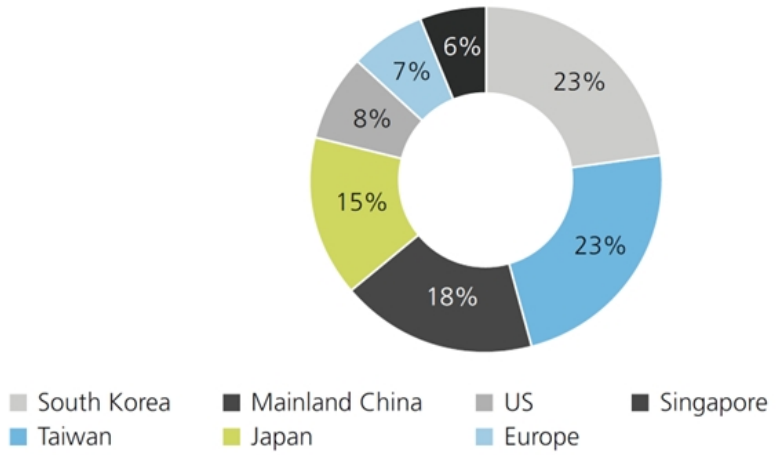
Following escalations over the past few quarters, efforts have been made to resume bilateral dialogue and defuse the tensions between the US and China. More conversations could usher in more clarity on evolving attitudes between the two sides. Though smaller conflicts could still emerge—such as the recent ban on materials used in chip production by both sides—we think the growing number of engagements with high-level dialogue is positive for bilateral relations. Collaboration will likely continue out of necessity, but we also expect structural tech tensions to persist in the coming years, with the potential to turn outright adversarial at times. Increased semiconductor capacity will likely lead to greater competition, a modest increase in industry cyclicality, and lower industry utilization rates, posing downside risks to industry profit margins.

Our base case is for the current restrictions on advanced semiconductor shipments—both for end-products and semiconductor equipment—to remain in place. The scope of restrictions may broaden, especially to other leading-edge technologies in quantum computing and AI. That said, we expect limited additional restrictions to be imposed on China's mature semiconductor industry and other low value-added hardware manufacturing segments. In fact, restrictions here could lead to adverse implications for mainland China's broader semiconductor industry given its reliance on mature fabrication nodes. On the upside, better bilateral dialogue may lead to the lifting of enacted restrictions, particularly in advanced semiconductor shipments for AI-related chips and certain segments of advanced semiconductor equipment. Should this occur, we think a rerating of global chip equipment makers and China's chipmakers could be likely.

Conclusion: Given the US and China's domestic political environments, we believe a long-term resolution is challenging. The current restrictions on

advanced chips and semiconductor equipment will likely remain in place. That said, a complete decoupling is unlikely given their deep economic and trade ties. US-China collaboration in low value-added hardware will remain in place, in our view, despite the gradual diversification of supply chains beyond mainland China.

Figure 31 - Global semiconductor installed wafer capacity per region
Market share, in %



Source: Company data, UBS

Regulations

Regulations have always been and will always be a risk for the tech sector—more so for AI today, as we expect regulations to evolve quickly in the years ahead. Moreover, with many general elections scheduled in 2024, we should see further heated debates around AI regulations.

That said, we welcome regulations at the earlier stages of the industry's evolution, as that should drive more orderly growth. Damage can be done when they are introduced during the later stages of development, which happened in a few industries like education and fintech.

From a regional perspective, China is currently the most advanced in terms of AI regulations, drafting enforcement on the technology as early as 2021. While we have yet to see its very own comprehensive AI Act, as opposed to its European Union peers, we have seen broader regulations and guidelines being disseminated in 2023 targeted specifically on generative AI services known as the "Interim Measures." This guideline is applicable to all generative AI service providers that leverage the technology to provide services to the broader public in China. Key principles with regard to content and data management include the upholding of core socialist values, non-discrimination, intellectual property rights, lawful rights of individuals, and algorithm transparency. Any violations will subject the service providers to be penalized in accordance with the relevant laws, including the Cybersecurity and Data Security Law. External vendors outside China that are providing

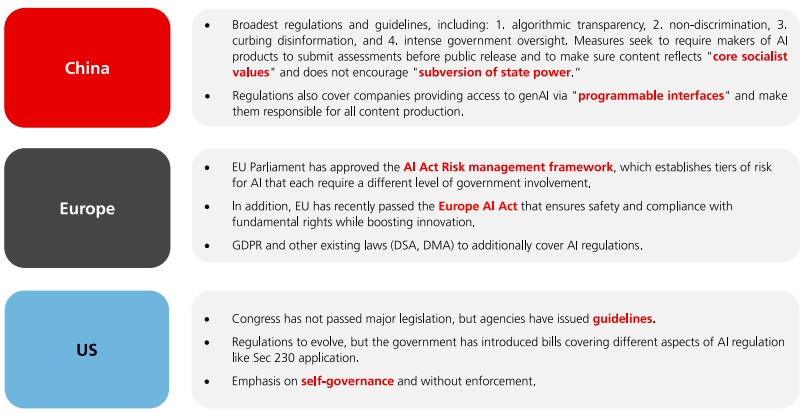
services to the nation are likewise subjected to similar requirements, which highlights the degree of extra-territorial effect of the Interim Measures.

In Europe, the European Parliament recently approved the Artificial Intelligence Act that aims to make AI systems safe, ethical and respectful of fundamental rights, while also promoting innovation. This Act resembles Europe's General Data Protection Regulation of 2016 but is tailored specifically for the artificial intelligence domain. The Act categorizes applications based on risk, with high-risk uses like hiring decisions facing strict rules. Overall, it emphasizes trustworthy AI development and use, potentially setting a global standard for AI governance. Greater levels of scrutiny and obligations are placed on companies utilizing or designing AI within the EU space, with penalties ranging from 3% to 7% of annual global turnover (with a cap of EUR 15–35mn) depending on the severity of the violations. The European Commission has introduced a tiered risk-based approach to assess the various magnitude of possible systemic risks, with the extreme end of the spectrum including violations such as utilizing AI-enabled manipulative techniques, or social scoring for public and private purposes.

On the other hand, we expect the US administration to try to strike a balance, since regulation around AI can stifle innovation and erode the US's (and its major tech platforms') significant first-mover advantage. Currently, we have yet to see any wide-ranging AI law being passed, but an Executive Order regarding the development and use of AI was set forth by the Biden administration in late 2023. "Accountable" development and deployment of AI remains the key focus, and the definition of AI remains rather broad, not solely defined by neural-networks-related systems or generative AI. With a greater emphasis on drafting comprehensive regulations by governing bodies, we expect the Executive Order to be the beginning of further regulatory initiatives within the US.

Conclusion: Regulations are worth monitoring as a risk, including export controls, but an excessive correction due to geopolitics or regulations could present a buying opportunity. Underlying demand trends for AI should continue to be solid in the foreseeable future, in our view.

AI regulation outlook



Source: Morgan Stanley, Bloomberg Intelligence and UBS

End notes

[1] Source: https://www.businessinsider.com/google-researchers-openai-chatgpt-to-reveal-its-training-data-study-2023-12#:~:text=The%20AI%20model%20powering%20ChatGPT,or%20570%20GB%2C%20of%20data.

[2] Source: https://sh-tsang.medium.com/brief-review-mmlu-measuring-massive-multitask-language-understanding-7b18e7cbbeab

[3] Source: https://rpc.cfainstitute.org/en/research/financial-analysts-journal/2023/long-term-shareholder-returns-evidence-from-64000-global-stocks

[4] Source: https://a16z.com/who-owns-the-generative-ai-platform/

[5] Source: https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds

[6] Source: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction

[7] Source: International Labor Organization (ILO), UBS estimates, as of May 2024

[8] Source: https://resources.github.com/copilot-for-business/

[9] Source: https://www.bcg.com/publications/2023/how-generative-ai-transforms-customer-service

[10] Source: https://www.coatue.com/insights

[11] Source: https://www.reuters.com/technology/microsoft-openai-planning-100-billion-data-center-project-information-reports-2024-03-29/

[12] Source: https://qz.com/openai-will-make-2-billion-revenue-soon-tech-ai-chatgpt-1851247985#:~:text=We%20may%20earn%20a%20commission%20from%20links%20on%20this%20page.&text=Just%20seven%20years%20after%20being,knowledge%20told%20the%20Financial%20Times.

[13] Source: https://analyticsindiamag.com/github-copilot-is-all-gain-no-pain-for-microsoft/

[14] Source: https://visualstudiomagazine.com/Articles/2024/02/05/copilot-numbers.aspx

[15] Source: https://marketingreport.one/news/global-digital-advertising-market-to-reach-667-billion-in-2024-report.html

[16] Source: https://finance.yahoo.com/news/global-call-centers-business-analysis-090800828.html

[17] Source: "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models", Eloundou et al (2023).

[18] Source: "AI Ethics in Business", Gartner (2021), accessed at: https://www.gartner.com/peer-community/oneminuteinsights/ai-ethics-business-hvj

[19] Source: "Tracking Clean Energy Progress", International Energy Agency (2023), accessed at: https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks

[20] Source:CB Insights

[21] Source: Sinovation Ventures Slide Deck, February 2024

[22] Source: sunyan.substack.com "The Economics of Large Language Models", Jan 21, 2023

[23] Source: https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024.pdf

[24] Source: International Energy Agency: Electricity 2024 – Analysis and forecast to 2026. Accessed on https://www.iea.org/reports/electricity-2024 as of 19 April 2024.

[25] Source: Siemens press release: Generative artificial intelligence takes Siemens' predictive maintenance solution to the next level, as of 5 February 2024.
And Siemens webpage: Predictive maintenance at scale. Accessed on: https://www.siemens.com/global/en/products/services/digital-enterprise-services/analytics-artificial-intelligence-services/predictive-services/senseye-predictive-maintenance.html, as of 27 February 2024.

[26] Source: USAF website: https://www.airforce.com/experience-the-air-force/airmen-stories/inside-air-force-innovation/project-skyborg

[27] Source: https://www.aboutamazon.com/news/retail/amazon-rufus

[28] Sources: https://www.cio.com/article/475786/dow-turns-to-ai-to-accelerate-chemical-search.htmlhttps://corporate.dow.com/en-us/news/press-releases/dow-wins-2023-ai-excellence-award.html#:~:text=MIDLAND%2C%20Michigan.,opens%20in%20a%20new%20tab).

[29] Source: https://www.prnewswire.com/news-releases/freeport-mcmoran-to-convert-fleet-of-cat-793-trucks-at-its-bagdad-mine-in-arizona-to-autonomous-haulage-using-cat-minestar-command-for-hauling-301972583.html

[30] Source: based on Goldman Sachs research

[31] Source: Avangrid Press release: Avangrid Pilots Mobile Robot Dog to Advance Substation Inspections with Artificial Intelligence, as of 08 February 2024. https://www.avangrid.com/w/avangrid-pilots-mobile-robot-dog-to-advance-substation-inspections-with-ai

[32] "The role of artificial intelligence in achieving the Sustainable Development Goals", Vinuesa et al, 2020.

[33] Source: https://www.hsph.harvard.edu/ecpe/programs/ai-for-health-care-concepts-and-applications/

[34] Source: International Monetary Fund, Cazzaniga et al, Gen-AI: Artificial Intelligence and the Future of Work, 2024

Non-Traditional Assets

Non-traditional asset classes are alternative investments that include hedge funds, private equity, real estate, and managed futures (collectively, alternative investments). Interests of alternative investment funds are sold only to qualified investors, and only by means of offering documents that include information about the risks, performance and expenses of alternative investment funds, and which clients are urged to read carefully before subscribing and retain. An investment in an alternative investment fund is speculative and involves significant risks. Specifically, these investments (1) are not mutual funds and are not subject to the same regulatory requirements as mutual funds; (2) may have performance that is volatile, and investors may lose all or a substantial amount of their investment; (3) may engage in leverage and other speculative investment practices that may increase the risk of investment loss; (4) are long-term, illiquid investments, there is generally no secondary market for the interests of a fund, and none is expected to develop; (5) interests of alternative investment funds typically will be illiquid and subject to restrictions on transfer; (6) may not be required to provide periodic pricing or valuation information to investors; (7) generally involve complex tax strategies and there may be delays in distributing tax information to investors; (8) are subject to high fees, including management fees and other fees and expenses, all of which will reduce profits.

Interests in alternative investment funds are not deposits or obligations of, or guaranteed or endorsed by, any bank or other insured depository institution, and are not federally insured by the Federal Deposit Insurance Corporation, the Federal Reserve Board, or any other governmental agency. Prospective investors should understand these risks and have the financial ability and willingness to accept them for an extended period of time before making an investment in an alternative investment fund and should consider an alternative investment fund as a supplement to an overall investment program.

In addition to the risks that apply to alternative investments generally, the following are additional risks related to an investment in these strategies:

- Hedge Fund Risk: There are risks specifically associated with investing in hedge funds, which may include risks associated with investing in short sales, options, small-cap stocks, "junk bonds," derivatives, distressed securities, non-U.S. securities and illiquid investments.
- Managed Futures: There are risks specifically associated with investing in managed futures programs. For example, not all managers focus on all strategies at all times, and managed futures strategies may have material directional elements.
- Real Estate: There are risks specifically associated with investing in real estate products and real estate investment trusts. They involve risks associated with debt, adverse changes in general economic or local market conditions, changes in governmental, tax, real estate and zoning laws or regulations, risks associated with capital calls and, for some real estate products, the risks associated with the ability to qualify for favorable treatment under the federal tax laws.

- Private Equity: There are risks specifically associated with investing in private equity. Capital calls can be made on short notice, and the failure to meet capital calls can result in significant adverse consequences including, but not limited to, a total loss of investment.
- Foreign Exchange/Currency Risk: Investors in securities of issuers located outside of the United States should be aware that even for securities denominated in U.S. dollars, changes in the exchange rate between the U.S. dollar and the issuer's "home" currency can have unexpected effects on the market value and liquidity of those securities. Those securities may also be affected by other risks (such as political, economic or regulatory changes) that may not be readily known to a U.S. investor.

Appendix

Risk information

UBS Chief Investment Office's ("CIO") investment views are prepared and published by the Global Wealth Management business of UBS Switzerland AG (regulated by FINMA in Switzerland) or its affiliates ("UBS"), part of UBS Group AG ("UBS Group"). UBS Group includes former Credit Suisse AG, its subsidiaries, branches and affiliates. Additional disclaimer relevant to Credit Suisse Wealth Management follows at the end of this section.

The investment views have been prepared in accordance with legal requirements designed to promote the **independence of investment research**.

Generic investment research – Risk information:

This publication is **for your information only** and is not intended as an offer, or a solicitation of an offer, to buy or sell any investment or other specific product. The analysis contained herein does not constitute a personal recommendation or take into account the particular investment objectives, investment strategies, financial situation and needs of any specific recipient. It is based on numerous assumptions. Different assumptions could result in materially different results. Certain services and products are subject to legal restrictions and cannot be offered worldwide on an unrestricted basis and/or may not be eligible for sale to all investors. All information and opinions expressed in this document were obtained from sources believed to be reliable and in good faith, but no representation or warranty, express or implied, is made as to its accuracy or completeness (other than disclosures relating to UBS). All information and opinions as well as any forecasts, estimates and market prices indicated are current as of the date of this report, and are subject to change without notice. Opinions expressed herein may differ or be contrary to those expressed by other business areas or divisions of UBS as a result of using different assumptions and/or criteria. UBS may utilise artificial intelligence tools ("AI Tools") in the preparation of this document. Notwithstanding any such use of AI Tools, this document has undergone human review.

In no circumstances may this document or any of the information (including any forecast, value, index or other calculated amount ("Values")) be used for any of the following purposes (i) valuation or accounting purposes; (ii) to determine the amounts due or payable, the price or the value of any financial instrument or financial contract; or (iii) to measure the performance of any financial instrument including, without limitation, for the purpose of tracking the return or performance of any Value or of defining the asset allocation of portfolio or of computing performance fees. By receiving this document and the information you will be deemed to represent and warrant to UBS that you will not use this document or otherwise rely on any of the information for any of the above purposes. UBS and any of its directors or employees may be entitled at any time to hold long or short positions in investment instruments referred to herein, carry out transactions involving relevant investment instruments in the capacity of principal or agent, or provide any other services or have officers, who serve as directors, either to/for the issuer, the investment instrument itself or to/for any company commercially or financially affiliated to such issuers. At any time, investment decisions (including whether to buy, sell or hold securities) made by UBS and its employees may differ from or be contrary to the opinions expressed in UBS research publications. Some investments may not be readily realizable since the market in the securities is illiquid and therefore valuing the investment and identifying the risk to which you are exposed may be difficult to quantify. UBS relies on information barriers to control the flow of information contained in one or more areas within UBS, into other areas, units, divisions or affiliates of UBS. Futures and options trading is not suitable for every investor as there is a substantial risk of loss, and losses in excess of an initial investment may occur. Past performance of an investment is no guarantee for its future performance. Additional information will be made available upon request. Some investments may be subject to sudden and large falls in value and on realization you may receive back less than you invested or may be required to pay more. Changes in foreign exchange rates may have an adverse effect on the price, value or income of an investment. The analyst(s) responsible for the preparation of this report may interact with trading desk personnel, sales personnel and other constituencies for the purpose of gathering, synthesizing and interpreting market information.

Different areas, groups, and personnel within UBS Group may produce and distribute separate research products **independently of each other**. For example, research publications from **CIO** are produced by UBS Global Wealth Management. **UBS Global Research** is produced by UBS Investment Bank. **Research methodologies and rating systems of each separate research organization may differ**, for example, in terms of investment recommendations, investment horizon, model assumptions, and valuation methods. As a consequence, except for certain economic forecasts (for which UBS CIO and UBS Global Research may collaborate), investment recommendations, ratings, price targets, and valuations provided by each of the separate research organizations may be different, or inconsistent. You should refer to each relevant research product for the details as to their methodologies and rating system. Not all clients may have access to all products from every organization. Each research product is subject to the policies and procedures of the organization that produces it.

The compensation of the analyst(s) who prepared this report is determined exclusively by research management and senior management (not including investment banking). Analyst compensation is not based on investment banking, sales and trading or principal trading revenues, however, compensation may relate to the revenues of UBS Group as a whole, of which investment banking, sales and trading and principal trading are a part.

Tax treatment depends on the individual circumstances and may be subject to change in the future. UBS does not provide legal or tax advice and makes no representations as to the tax treatment of assets or the investment returns thereon both in

general or with reference to specific client's circumstances and needs. We are of necessity unable to take into account the particular investment objectives, financial situation and needs of our individual clients and we would recommend that you take financial and/or tax advice as to the implications (including tax) of investing in any of the products mentioned herein.
This material may not be reproduced or copies circulated without prior authority of UBS. Unless otherwise agreed in writing UBS expressly prohibits the distribution and transfer of this material to third parties for any reason. UBS accepts no liability whatsoever for any claims or lawsuits from any third parties arising from the use or distribution of this material. This report is for distribution only under such circumstances as may be permitted by applicable law. For information on the ways in which CIO manages conflicts and maintains independence of its investment views and publication offering, and research and rating methodologies, please visit www.ubs.com/research-methodology. Additional information on the relevant authors of this publication and other CIO publication(s) referenced in this report; and copies of any past reports on this topic; are available upon request from your client advisor.
Important Information About Sustainable Investing Strategies: Sustainable investing strategies aim to consider and incorporate environmental, social and governance (ESG) factors into investment process and portfolio construction. Strategies across geographies approach ESG analysis and incorporate the findings in a variety of ways. Incorporating ESG factors or Sustainable Investing considerations may inhibit UBS's ability to participate in or to advise on certain investment opportunities that otherwise would be consistent with the Client's investment objectives. The returns on a portfolio incorporating ESG factors or Sustainable Investing considerations may be lower or higher than portfolios where ESG factors, exclusions, or other sustainability issues are not considered by UBS, and the investment opportunities available to such portfolios may differ.
External Asset Managers / External Financial Consultants: In case this research or publication is provided to an External Asset Manager or an External Financial Consultant, UBS expressly prohibits that it is redistributed by the External Asset Manager or the External Financial Consultant and is made available to their clients and/or third parties.
USA: Distributed to US persons only by UBS Financial Services Inc. or UBS Securities LLC, subsidiaries of UBS AG. UBS Switzerland AG, UBS Europe SE, UBS Bank, S.A., UBS Brasil Administradora de Valores Mobiliarios Ltda, UBS Asesores Mexico, S.A. de C.V., UBS SuMi TRUST Wealth Management Co., Ltd., UBS Wealth Management Israel Ltd and UBS Menkul Degerler AS are affiliates of UBS AG. **UBS Financial Services Inc. accepts responsibility for the content of a report prepared by a non-US affiliate when it distributes reports to US persons. All transactions by a US person in the securities mentioned in this report should be effected through a US-registered broker dealer affiliated with UBS, and not through a non-US affiliate. The contents of this report have not been and will not be approved by any securities or investment authority in the United States or elsewhere. UBS Financial Services Inc. is not acting as a municipal advisor to any municipal entity or obligated person within the meaning of Section 15B of the Securities Exchange Act (the "Municipal Advisor Rule") and the opinions or views contained herein are not intended to be, and do not constitute, advice within the meaning of the Municipal Advisor Rule.**
For country information, please visit ubs.com/cio-country-disclaimer-gr or ask your client advisor for the full disclaimer.
Additional Disclaimer relevant to Credit Suisse Wealth Management
You receive this document in your capacity as a client of Credit Suisse Wealth Management. Your personal data will be processed in accordance with the Credit Suisse privacy statement accessible at your domicile through the official Credit Suisse website . In order to provide you with marketing materials concerning our products and services, UBS Group AG and its subsidiaries may process your basic personal data (i.e. contact details such as name, e-mail address) until you notify us that you no longer wish to receive them. You can optout from receiving these materials at any time by informing your Relationship Manager.
Except as otherwise specified herein and/or depending on the local Credit Suisse entity from which you are receiving this report, this report is distributed by UBS Switzerland AG, authorised and regulated by the Swiss Financial Market Supervisory Authority (FINMA).
Version A/2025. CIO82652744