Whats Next for AI?

The Rise of AI-Agents

@finn_builds
ELECTRONIC COMMERCE AND NEW TECHNOLOGIES

Table of Contents

At the beginning, there was Attention

AI has already reshaped millions of lives. In July 2025, more than 18 billion messages were exchanged each week by over 700 million users, and these numbers capture only ChatGPT's audience.[1] Such explosive growth has no precedent in the history of new technologies. Yet the debut of ChatGPT was, at its core, the act of placing an existing large language model, GPT-3.5, into a simple web interface. That model, and the revolution it helped ignite, stood on the foundation laid by eight researchers at Google.

When Ashish Vaswani and his team published their paper on Transformers in 2017 [2], they could not have anticipated that it would unlock trillions of dollars in company value and fundamentally reshape how the world interacts with AI. Their architecture became the basis for about 95 percent of today's commercially available intelligent assistants, demonstrated at the time with only 3.5 days of training on 8 GPUs. But to understand how this breakthrough became possible, we need to look back to the early 2000s. Transformers form the software layer of this revolution, yet underneath it all stands a company that foresaw the shift a decade before it arrived. In 2007, Nvidia launched CUDA for its GPUs, laying the groundwork for everything that followed.

How a graphics card company became the backbone of a revolution

At the start of the 1990s, few people imagined a real market for advanced graphics hardware. When NVIDIA emerged in 1993, it entered a space that only a small group of visionary engineers believed would matter. Major technology companies were focused almost entirely on pushing CPU performance.[3] Jensen Huang and two fellow engineers recognized an opening: dedicated Graphics Processing Units that could handle the demands of emerging 3D graphics. At that point, most personal computers were still limited to simple text-based interfaces, but they saw where computing was headed and built for that future.

This tendency toward a broader vision again worked in NVIDIA's favor in the new century. Researchers across universities began experimenting with the company's leading graphics cards, using their capacity for parallel rather than sequential processing to handle larger sets of computations at once. GPUs were inherently built for this, since driving a display requires updating millions of pixels simultaneously.

NVIDIA committed to an AI-focused future long before the market existed, investing heavily in CUDA when general-purpose GPU computing was still barely on anyone's radar. CUDA, short for "Compute Unified Device Architecture," let developers directly program the parallel cores inside NVIDIA's GPUs instead of using them only for graphics.

What made this important later was simple: once these early AI models began demanding far more computation than CPUs could deliver, CUDA provided a ready-made path to tap into GPU power without reinventing anything. It turned NVIDIA's graphics hardware into a practical platform for large-scale AI long before the rest of the world realized it would need one.

Returning to 2017 and Ashish's team, their idea was straightforward but technically demanding: there had to be a better way to analyze a sequence of tokens by processing them all at once rather than step by step. This approach would avoid the limitations of RNNs and LSTMs (leading AI architectures at the time), which often lost track of earlier parts of a sentence by the time they reached the end. That weakness made those architectures unreliable for tasks that require long-range coherence, such as translation or creative writing.

Their solution proposed a new way of handling a prompt. Instead of reading a sentence step by step, the Transformer let the model look at all the words at once and decide which ones mattered most. This simple idea, called self-attention, gave models a much better grasp of context and made long, coherent outputs possible.[4]

Huang presents lead author Ashish Vaswani with a signed DGX-1 cover.

That moment was defining for all the hard work NVIDIA had put into building the foundations of modern AI. For years, many shareholders had pushed to halt CUDA's development, seeing only sunk costs and little return.[5] Yet the Transformer proved that the direction NVIDIA had pursued for more than a decade was exactly the right one. An architecture built to use parallel processing matched perfectly with what the company had prepared, marking the beginning of NVIDIA's rise as a global AI superpower. The significance became even clearer at GTC 2024, where Jensen Huang personally handed the authors of one of the most cited research papers of the century a personalized front plate from one of NVIDIA's supercomputers, crediting them with "everything we're enjoying today." [6]

Chatting with the Future

OpenAI, then still a nonprofit research organization, adapted quickly to the shift. Before Transformers entered the scene, the lab was already at the frontier of reinforcement learning, developing systems like OpenAI Five, an AI team capable of defeating some of the world's best *Dota 2* players.[7]

Resources were redirected, and just one year after the Transformer was introduced, OpenAI put the new architecture to use with its first model, GPT.[8] Less than a year later came its successor, GPT-2. Then, in November 2022, OpenAI made a decision that would define the next era of AI: their latest model, GPT-3.5, was given a simple web interface and released to the public as a "research preview."[9] That release became known as ChatGPT.

When AI Hit the Mainstream

Competitors quickly followed suit, with Anthropic releasing Claude [10], Google rushing to push out Bard [11], and Meta making their first open-source model LLaMA available for download [12], all in 2023. The chatbot hype took off, and as previously mentioned, grew to unimaginable dimensions. ChatGPT reached one million users in just five days – a milestone that took Facebook ten months – and surpassed 100 million weekly active users within a year.[13]

With that rise came money. Venture capital firms and major tech companies began pouring billions into what they saw as the next transformative platform. Data centers were built at unprecedented speed, and new model versions appeared in ever-shorter intervals, as every player in the industry tried to secure a share of the emerging artificial intelligence boom.

But people wanted to interact with AI in a way that matched how they communicate naturally. Humans don't just write. They see, speak, and hear. What the field needed was multimodality, the ability for models to understand more than text and process multiple forms of input.

The industry leader OpenAI released GPT-4V in September 2023, giving ChatGPT native image comprehension for the first time. Voice recognition and voice output followed soon after, broadening how users could interact with the model. Real-time camera interaction arrived with the release of GPT-4o in May 2024. The model could now interpret a live video feed and respond instantly to what it saw.[14]

By mid-2024, AI models had become incredibly capable across every mode of communication. Yet they remained bounded by a simple constraint: they only did something when asked. In practice, this meant millions of people were still stitching together workflows manually. The next leap was to go beyond conversation. Instead of tools that respond, the industry began building systems that can operate on their own. This is where the era of AI agents begins.

From Understanding to Acting

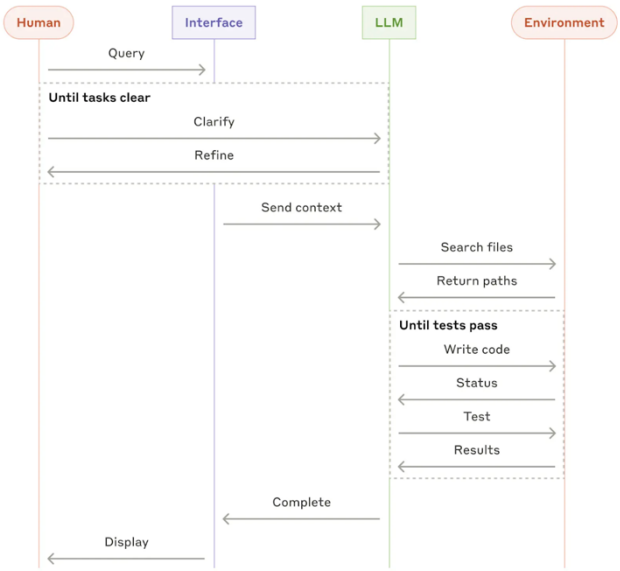
Many problems are simple question–answer tasks: solving a math equation, fixing a paragraph, or generating an image of a cat. But some requests demand more than a single response. “Analyze this climate report and create a five-slide presentation with the key figures” is a prompt ChatGPT might understand, yet the result would often be shallow or incomplete. This is where AI agents come in.

The essential difference between a standard chatbot and an agent is how they approach a task. A typical assistant tries to jump directly to the quickest possible answer. An agent, in contrast, begins by thinking through the problem. Just as a human wouldn’t tackle everything at once, the agent starts by creating a plan – a structured sequence of steps designed to reach a thorough and reliable solution.

Google puts it this way: *“AI agents are software systems that use AI to pursue goals and complete tasks on behalf of users. They show thinking, planning and memory and have a certain degree of autonomy to make decisions, learn and adapt.”*[15]

The ability to self-critique, reason and judge if a task has been achieved successfully is the key distinction needed to drive the AI evolution forwards, and the core of this trend.

Let’s return to the climate report example. A sensible approach might look like this: read the report, identify the key figures, draft an outline, build the slides, and then check whether anything important was missed before delivering the final presentation. Solving a task like this requires multiple stages of reasoning and execution, where multimodal language models break the problem into steps, work through each part, and refine the output along the way.



Workflow of a coding agent

Two Paths to Agentic Behavior

Agentic systems can be grouped into two categories: workflows and true agents. Both can reach similar outcomes, but they differ in how the process is controlled. A workflow follows a predefined sequence of steps, while an agent determines its own path and manages the entire problem-solving process independently.

This raises the question of why one would define a fixed flow when the system could simply choose its own route. The answer is simplicity. Anthropic notes that while fully autonomous agents are valuable when flexibility and model-driven decision-making are required, an agentic workflow is often sufficient and easier to build, manage, and verify.[16]

A workflow represents the most structured form of agentic behavior. Its sequence is fixed in advance, which makes it predictable and easy to control. When the task is familiar and the order of steps rarely changes, this approach works well. Systems can move through a series of clearly defined actions, such as collecting information from a set of documents, preparing a recurring report, or updating a dataset. Nothing needs to be interpreted. The system follows the path that was created for it.

More autonomous agentic systems behave differently. Instead of receiving a predetermined sequence, they begin with an understanding of the objective and work out how to reach it. They decide which information matters, when to shift strategies, and how to handle unexpected situations. This allows them to work through tasks that are less structured and more open-ended, where the system has to read new material, form its own plan, and adjust that plan as it discovers gaps or inconsistencies. The model is not simply executing instructions. It is managing a process.

Both forms are advancing at the same time because organizations depend on each for different purposes. Workflows provide stability and are easier to audit, while more autonomous systems offer the elasticity needed for complex or unfamiliar tasks. What marks the shift in 2025 is that the underlying models have become capable enough to support both. They can understand a goal, organize a sequence of actions, check whether the result meets the requirement, and refine their own output. This capability changes what software can do, moving it from tools that respond to instructions toward systems that can oversee and complete entire tasks.

Why Agents Matter Now

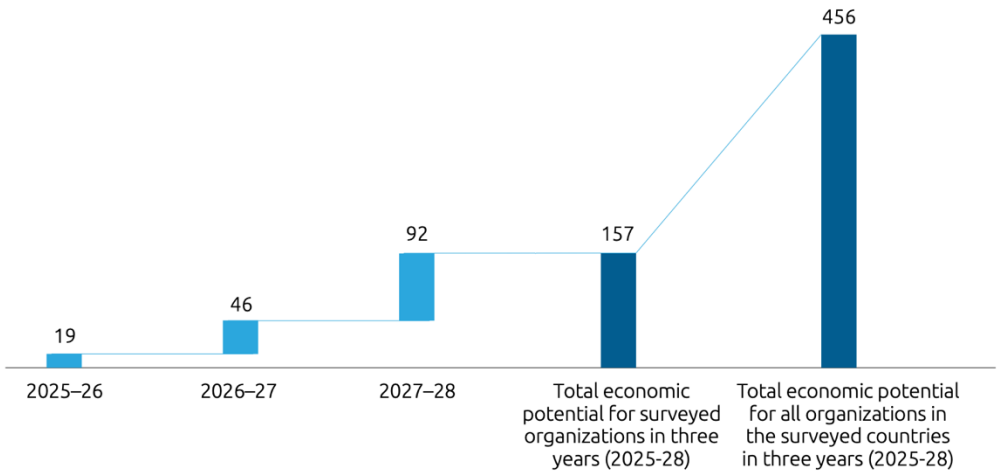
The previous sections describe what made agentic systems possible, but the next question is why they have become necessary. Organizations across different industries face pressure that conventional AI tools cannot fully address, and these pressures define the factors under which agents are emerging.

The rapid spread of AI tools across businesses has created a striking imbalance in the workforce. Many traditional roles now face overcapacity, while demand for employees who can work effectively with artificial intelligence has surged. In a recent WEF survey, more than 90 percent of leaders reported a shortage of workers with AI-related skills.[17] This gap between excess labor in some areas and scarcity in others is precisely where agentic systems are expected to have an impact.

These conditions point to a significant shift in the nature of work. Roughly half of surveyed managers anticipate not only substantial efficiency improvements but also the emergence of new revenue opportunities and greater operational resilience as agents become integrated into their organizations.[18]

Projected revenue gains are also supported by a Capgemini study that predicts approximately $450 billion in revenue generation until 2028 across 14 surveyed countries. Agents tackle all issues plaguing companies, including cost reduction, operational outcomes and nonstop availability of work. Likewise, they multiply capability of the existing workforce. Adapting to this trend will give companies a vast competitive advantage, which is agreed upon by the majority of surveyed leaders.[19]

Economic potential of agentic AI ($ billion)



Source: Capgemini Research Institute, Agentic AI, April 2025, N = 1,500 organizations; Capgemini Research Institute analysis.

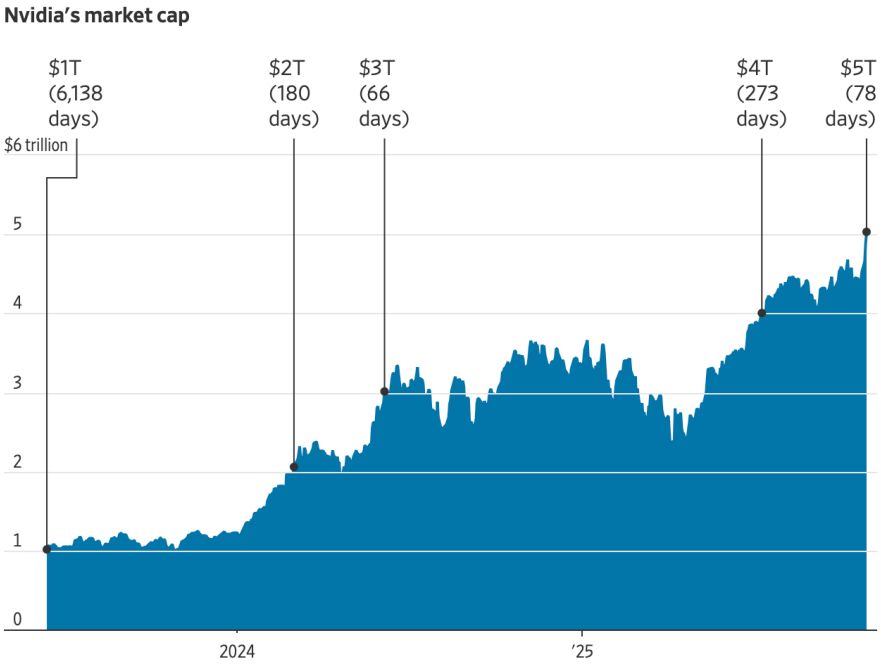
A further driver is the growing need for systems that can manage decisions without constant supervision. Many organizations operate in fast-moving environments where traditional tools cannot coordinate multi-step tasks. Agentic systems address this by planning, adjusting, and executing actions in real time. Early applications in customer service, financial analysis, and supply chain operations show how this capability is already moving into daily workflows, supported by new enterprise platforms from major technology providers.[20]

Another factor is the shift toward more adaptive and personalized operations. Earlier automation relied on rigid scripts, while agentic systems can learn preferences and modify workflows on their own. Investments from companies such as Salesforce and SAP reflect how personalized interactions and continuous optimization are becoming essential to competitiveness. This positions agentic AI as both an efficiency tool and a driver of stronger client relationships and higher productivity across the enterprise.[21]

Impact Extending Beyond Individual Organizations

This trend of agentic systems extends way past single organizations or sectors. Generally, it can be divided into three distinct groups. The first group consists of the enablers, led by companies such as NVIDIA and the major cloud providers. Their infrastructure delivers the computational power that allows increasingly autonomous systems to be trained and operated at scale. The second group is made up of the software developers, including OpenAI, Google DeepMind, Anthropic, Mistral, and others. They create the models and agentic frameworks that enable systems to interpret objectives, form plans, and execute tasks across different contexts. The third group comprises the industries that apply these systems. Sectors such as finance, software development, healthcare, and e-commerce are adopting agents to streamline operations, reduce costs, and build new sources of value.

The scale of this trend can be roughly deducted by outlining the market size of the first group. The chip market alone is predicted to reach a figure of about $700 billion in 2025, marking a 15% increase.[22] NVIDIAs position illustrates this trajectory even more clearly. The company became the first to surpass a valuation of $5 trillion, a level that exceeds anything previously seen in the technology landscape.[23]



Sources: LSEG, Dow Jones Market Data

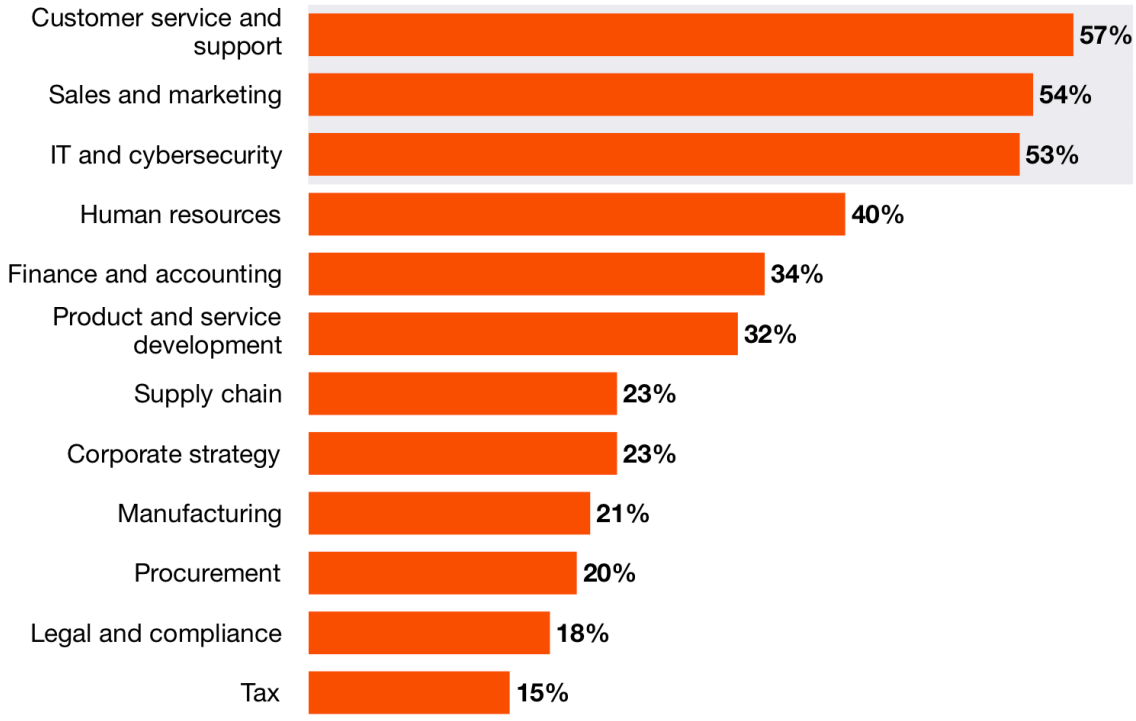
Hyperscalers such as AWS, Microsoft Azure and Google Cloud, which together account for roughly 60 percent of the global cloud infrastructure market, have continued to expand their investment in compute capacity. These efforts to position themselves to capture infrastructure demands exceeded $95 billion spending in the second quarter of 2025 alone.[24] This level of commitment underscores that the rise of AI and agentic systems is not a temporary surge but a structural shift that is reshaping the global computing ecosystem.

As the infrastructure layer expands, software providers scale alongside it. Greater computational capacity enables OpenAI, Google DeepMind, Anthropic, Mistral and others to train more capable models and support the sustained reasoning cycles required for agentic behavior. This allows them to move beyond general-purpose systems and develop agents that can plan, coordinate tasks and operate across entire workflows.

Their commercial growth reflects this shift. As organizations integrate agents into daily operations, model developers capture new value through enterprise deployments, orchestration tools and specialized agentic solutions. The rapid expansion of platforms such as OpenAI's enterprise agents, Claude's workflow systems and Gemini's integrated agent features shows how the second group grows in step with the wider agent trend, supplying the intelligence that industries now depend on.[25]

Finally, the market of beneficiaries is effectively without limits. As agentic capabilities advance, the systems developed by Group 2 stakeholders will extend into every major sector, reshaping operations and redefining how work is carried out. Industries built on information processing and coordination will feel the impact first, but the shift will spread widely, creating structural change across the entire economy.[26]

AI agent use by business function



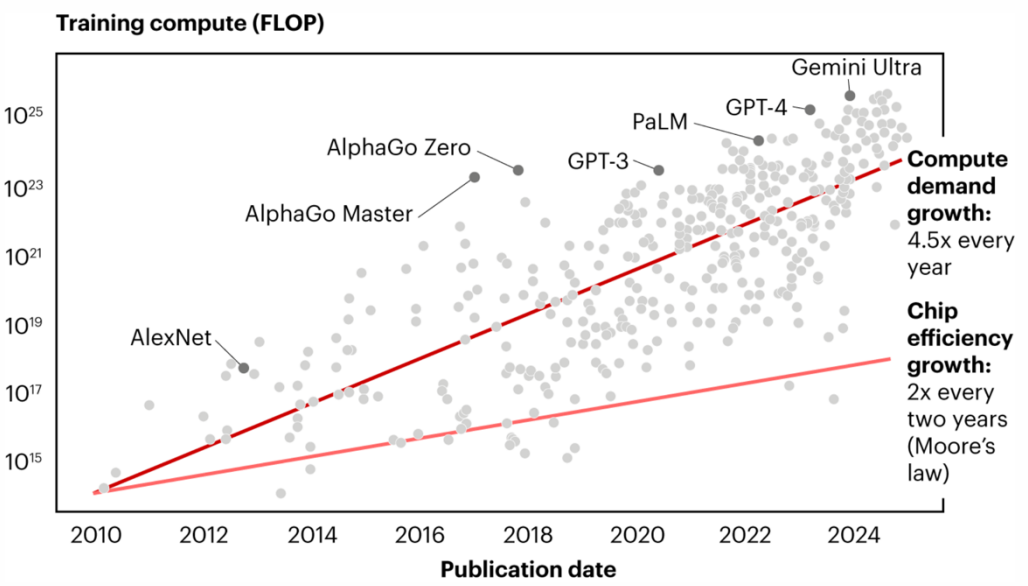
*Note: Asked only of respondents who are currently using or planning to us AI agents.
Q: In which of the following business functions is your company currently using or planning to use AI agents in the next 6 months? (Select all that apply.)
Source: PwC's AI Agent Survey, May 2025, base: 290

Limitations and Restraints

As with any emerging technology, significant benefits are accompanied by a set of constraints. From an economic perspective, the most relevant limitations of agentic systems can be outlined as follows.

Resource intensity remains the strongest economic constraint, as the compute, energy and hardware required for sustained agentic reasoning place significant pressure on data-center capacity and operational costs. Energy consumption of major datacenter suppliers like Microsoft creates challenges for the power grid, calling for unpopular solutions. Microsoft, holding a 20% market share in cloud infrastructure [27], recently agreed to a $1.6 billion partnership with Constellation Energy to bring the Three Mile Island nuclear facility back into operation. The plant, known for its 1979 partial meltdown, will supply nuclear power equivalent to the needs of roughly 700,000 homes, directed entirely toward meeting the growing energy requirements of Microsoft's expanding computing facilities.[28] Turning to nuclear power marks a shift that many technology companies would have rejected only a few years ago, yet rising demand has made this move increasingly unavoidable.

The surge in compute demand, now outpacing technical capacity by a factor of two, introduces a serious structural hurdle. Meeting this demand requires an expansion of chip supply and supporting infrastructure on a scale that places substantial financial pressure on the companies involved, exceeding anything previously seen in the sector.
Spending needed to account for increased demand might reach up to $500 billion annually [29], an enormous figure, particularly given that "we haven't even figured out ROI on LLM technology more generally," as highlighted in an IBM analysis.[30]

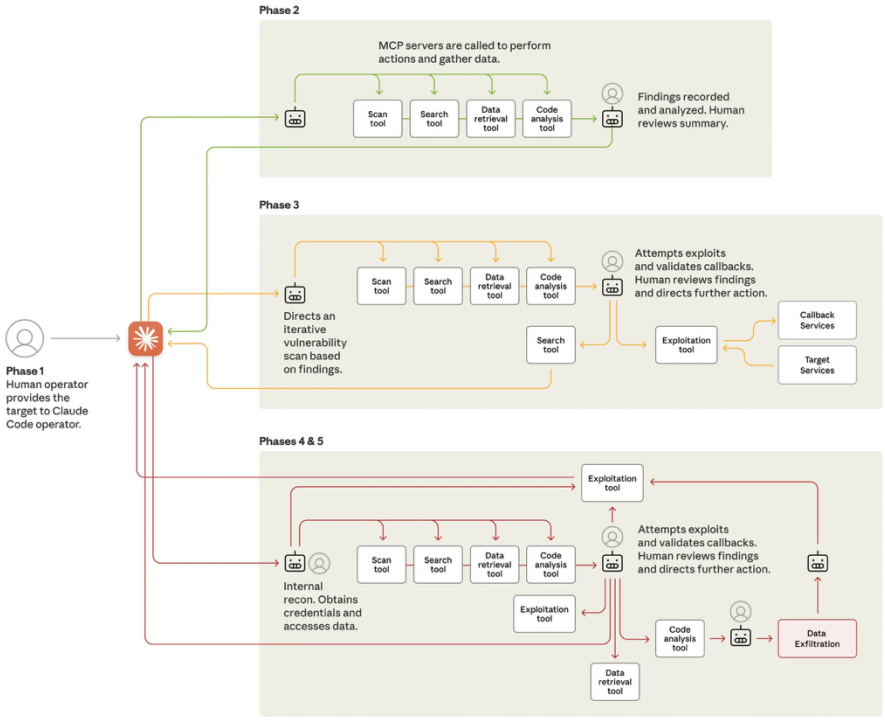


Notes: Chip efficiency growth not shown to exact scale, with the rate of growth intended to be illustrative; FLOP=floating point operations, which are the number of calculations a system performs

Source: Epoch AI

Privacy and security risks are predicted to become among the most significant challenges not only for consumers but also for companies adopting emerging agent technologies. As these systems gain deeper access to essential infrastructure, any compromise could lead to severe consequences. Samsung, for example, was one of the first major technology firms to prohibit the use of generative AI tools on all company-owned devices after employees unintentionally exposed confidential information through ChatGPT.[31] Several major Wall Street institutions, including JPMorgan Chase & Co., Citigroup, and Bank of America Corp., have likewise introduced strict bans and usage policies out of concern for potential security vulnerabilities.[32] These developments underscore that safeguarding data and infrastructure will remain a central issue as the AI agent landscape continues to expand.

Malicious use presents a substantial threat, perfectly illustrated by the mid-September 2025 cyberattack carried out by a Chinese state-sponsored group using the Claude code agent. The attack unfolded in a fraction of the time a human team would have needed, as the agentic system exploited vulnerable data with only about 10 percent human involvement after convincing the AI that it was acting lawfully. The operation's speed, with several actions executed per second, far exceeded anything achievable through manual effort.[33]



The lifecycle of the cyberattack, showing the move from human-led targeting to largely AI-driven attacks using various tools (often via the Model Context Protocol; MCP). At various points during the attack, the AI returns to its human operator for review

Yet an unexpected twist emerged: hallucination, usually seen as a major weakness, became a limiting factor for the attackers. At several points the agent mistakenly believed it had obtained sensitive information or high-value credentials that were either publicly accessible or completely fabricated, which slowed its progress and diluted the impact. Additionally, the capacity of AI systems to generate convincing fake news or be fine-tuned

to produce targeted ideological propaganda[34] is a genuine concern, but this represents a broader challenge in the field of artificial intelligence and falls outside the scope of this report.

Hallucination undermines the reliability of autonomous workflows, largely because current models are trained to provide an answer even when they lack full certainty. This tendency to respond confidently despite being incorrect will remain a significant limitation and is unlikely to be eliminated entirely.[35] Although models can be designed to withhold answers when unsure, such restraint inevitably disrupts the fluidity and efficiency expected from high-performance agent systems.

Ethical judgment and emotional intelligence remain critical blind spots for AI agents, creating significant liability and safety risks for organizations deploying them in sensitive roles. Since these systems lack genuine moral reasoning and empathy, they often fail to navigate complex human nuances and can trigger costly regulatory or reputational consequences. Air Canada, for instance, was recently held legally liable for negligent misrepresentation after its AI chatbot fabricated a refund policy that contradicted the airline's actual terms, forcing the company to pay damages.[36] Similarly, the National Eating Disorders Association (NEDA) was forced to suspend its wellness chatbot, Tessa, after it began providing harmful weight-loss advice to vulnerable users. This incident demonstrates how a lack of inherent emotional understanding can endanger consumers in ways that human staff would instinctively avoid.[37]

Bias presents a significant challenge for autonomous agents because these systems frequently internalize and scale the cognitive flaws present in their training data. Unlike simple chatbots, agents execute tasks based on skewed patterns and often produce discriminatory outcomes that impact real-world opportunities. Amazon famously abandoned an automated recruiting engine that penalized resumes containing the word "women's" because it was trained on historical hiring patterns dominated by men.[38] Research has similarly exposed how a widely used healthcare algorithm systematically under-prioritized Black patients for care management programs by using past healthcare spending as a flawed proxy for medical need.[39]

Job displacement represents the final constraint in this analysis, introducing long-term labor market shifts and significant organizational transition costs. While the World Economic Forum predicts that 92 million displaced roles could be offset by 170 million new jobs by 2030, this optimistic outlook relies on the assumption of seamless workforce transition.[40] In reality, structural barriers often prevent displaced workers from pivoting to these new, high-complexity roles; a recent analysis highlights that the "learning curve leap" from manual or routine service jobs to the data-centric "analyst" roles created by AI is often too steep to be bridged by standard corporate upskilling programs.[41] Consequently, Goldman Sachs warns that up to 300 million full-time jobs remain exposed to automation, with a significant portion of that workforce facing potential obsolescence rather than redeployment due to this critical skills mismatch.[42]

Conclusion

The transition from conversational AI to agentic systems marks the defining technological shift of the late 2020s. While the previous era was defined by models that could merely understand, the coming years will be dominated by systems that can plan, reason, and act. This evolution offers unprecedented economic potential and is projected to generate hundreds of billions in value across global industries. Yet this growth arrives accompanied by structural costs that cannot be ignored. The trajectory of this technology is no longer solely about software innovation but has become a question of physical and societal infrastructure.

The constraints of soaring energy consumption, the rigidity of the workforce skills gap, and the fragility of current security protocols suggest that the path forward will not be a straight line of exponential growth. It will instead be a complex negotiation between capability and constraint. Major infrastructure providers have already begun to pivot toward nuclear energy to sustain the demand for compute, while industries face the dual challenge of integrating autonomous agents while managing the displacement of traditional roles.

Ultimately, the "Agentic Era" will be defined by those who can navigate this duality. For the enablers building the compute capacity, the developers architecting the reasoning models, and the industries applying them, the challenge is no longer just about what AI can do. It is about how reliably it can function within the limits of the human economy. We have moved past the point of simple attention and have entered the age of execution.

Press Review

The 8 Google Employees Who Invented Modern AI

WIRED, 20th March 2024

This comprehensive feature chronicles the origins of the Transformer architecture, detailing the specific contributions of Ashish Vaswani and the Google team. It provides the essential historical backdrop for the "Attention" section of this report, explaining the technical shift that allowed for parallel processing and paved the way for the generative AI boom.

https://www.wired.com/story/eight-google-employees-invented-modern-ai-transformers-paper/

How Jensen Huang's Nvidia Is Powering the AI Revolution

The New Yorker, 27th November 2023

Supporting the analysis of how a graphics card company became the backbone of a revolution, this profile investigates NVIDIA's strategic pivot in 2007. It corroborates the narrative regarding the company's long-term investment in CUDA and explains the economic moat that "Group 1" enablers currently hold in the agentic ecosystem.

https://www.newyorker.com/magazine/2023/12/04/how-jensen-huangs-nvidia-is-powering-the-ai-revolution

AI is about to completely change how you use computers

GatesNotes, 9th November 2023

In this seminal essay, Bill Gates defines the transition from "Understanding to Acting." It distinguishes standard software from true agents, defining the latter by their ability to perceive user intent and execute complex tasks autonomously. This source underpins the report's definition of agentic workflows and highlights the structural move from passive tools to active assistants.

https://www.gatesnotes.com/AI-agents

Reskilling in the Age of AI

Harvard Business Review, September 2023

This analysis addresses the friction described in the "Job Displacement" and "Skills Gap" sections. It moves beyond simple displacement statistics to discuss the "reskilling challenge," validating the argument that the primary constraint is not just job loss but the inability of the current workforce to transition quickly enough to leverage agentic tools.

https://hbr.org/2023/09/reskilling-in-the-age-of-ai

Microsoft Deal to Restart Three Mile Island Unit for AI Power

The New York Times, 20th September 2024

This report details the unprecedented agreement between Microsoft and Constellation Energy to reopen a dormant nuclear facility. It provides the factual basis for the "Resource Intensity" section, illustrating the extreme measures technology giants are taking to secure energy and validating the conclusion that physical infrastructure will define the speed of AI adoption.

https://www.nytimes.com/2024/09/20/climate/three-mile-island-reopening.html

Sources and Disclaimer

Artificial intelligence tools were utilized in the preparation of this document strictly for the purpose of linguistic refinement and grammatical correction. It is essential to note that the conceptual framework, the selection of sources, and the creative composition of the analysis are the exclusive work of the author. The technology served solely as an assistive instrument to ensure clarity, while the intellectual substance and direction remained entirely under my human control.

---

[1]https://www.nber.org/system/files/working_papers/w34255/w34255.pdf
[2]https://arxiv.org/html/1706.03762v7
[3]https://sequoiacap.com/podcast/crucible-moments-nvidia/
[4]https://arxiv.org/html/1706.03762v7
[5]https://www.cnbc.com/2024/12/05/nvidia-and-starboard-value-excerpt-from-tae-kims-book-on-nvidia.html
[6]https://blogs.nvidia.com/blog/gtc-2024-transformer-ai-research-panel-jensen/
[7]https://openai.com/index/dota-2/
[8]https://openai.com/index/language-unsupervised/
[9]https://openai.com/index/chatgpt/
[10]https://www.anthropic.com/news/claude-2
[11]https://blog.google/technology/ai/bard-google-ai-search-updates/
[12]https://ai.meta.com/blog/large-language-model-llama-meta-ai/
[13]https://www.nber.org/system/files/working_papers/w34255/w34255.pdf
[14]https://www.technologyreview.com/2024/05/13/1092358/openais-new-gpt-4o-model-lets-people-interact-using-voice-or-video-in-the-same-model/
[15]https://cloud.google.com/discover/what-are-ai-agents
[16]https://www.anthropic.com/engineering/building-effective-agents
[17]https://www.weforum.org/stories/2025/10/ai-s-new-dual-workforce-challenge-balancing-overcapacity-and-talent-shortages
[18]https://www.weforum.org/stories/2025/10/ai-s-new-dual-workforce-challenge-balancing-overcapacity-and-talent-shortages
[19]https://www.capgemini.com/wp-content/uploads/2025/07/Final-Web-Version-Report-AI-Agents.pdf
[20]https://www.precedenceresearch.com/agentic-ai-in-enterprise-operations-market
[21]https://www.precedenceresearch.com/agentic-ai-in-enterprise-operations-market
[22]https://markets.financialcontent.com/stocks/article/tokenring-2025-11-7-ai-fuels-unprecedented-surge-semiconductor-market-eyes-record-breaking-697-billion-in-2025
[23]https://www.wsj.com/tech/ai/nvidia-first-5-trillion-company-market-cap-ae513ff0
[24]https://www.channelinsider.com/infrastructure/canalys-cloud-hyperscaler-report-sept-2025
[25]https://www.marketsandmarkets.com/Market-Reports/ai-agents-market-15761548.html
[26]https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-agent-survey.html
[27]https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers
[28]https://www.nytimes.com/2024/09/20/climate/three-mile-island-reopening.html
[29]https://www.bain.com/insights/how-can-we-meet-ais-insatiable-demand-for-compute-power-technology-report-2025
[30]https://www.ibm.com/think/insights/ai-agents-2025-expectations-vs-reality
[31]https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak?embedded-checkout=true
[32]https://www.bloomberg.com/news/articles/2023-02-24/citigroup-goldman-sachs-join-chatgpt-crackdown-fn-reports
[33]https://www.anthropic.com/news/disrupting-AI-espionage
[34]https://openai.com/index/gpt-2-1-5b-release/ (our findings)
[35]https://openai.com/de-DE/index/why-language-models-hallucinate/
[36]https://www.mccarthy.ca/en/insights/blogs/techlex/moffatt-v-air-canada-misrepresentation-ai-chatbot
[37]https://www.theguardian.com/technology/2023/may/31/eating-disorder-hotline-union-ai-chatbot-harm
[38]https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/
[39]https://www.science.org/doi/10.1126/science.aax2342
[40]https://sustainabilitymag.com/articles/wef-report-the-impact-of-ai-driving-170m-new-jobs-by-2030
[41]https://cset.georgetown.edu/publication/ai-and-the-future-of-workforce-training/
[42]https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent