Deloitte. Insights

TMT Predictions 2026: The gap narrows, but persists

Deloitte predicts 2026 will see the gap between the promise and reality of AI narrow, as further movements towards getting it to scale are made

ARTICLE • 9-MIN READ • 17 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

In 2026, Deloitte predicts the roar around artificial intelligence will be getting quieter—and smarter—as the sometimes unglamorous, high-impact work of making AI usable at scale continues to get underway. The gap between promise and reality will narrow but not disappear: Progress will come less from headline-grabbing new models and more from fundamentals. That more practical focus matters because tech, media, and telecom's growing importance is not just about chips and code—it's about how every other industry uses those TMT capabilities for its own growth, efficiencies, and innovation.

AI helps drive cross-industry transformation

In 2026 and beyond, it looks like we have moved from "software is eating the world" to "TMT is eating the world," led by AI—especially agentic AI. In the United States, spending on AI data centers currently accounts for almost all gross domestic product growth in the first half of the year.[1] In 2008, about 19% of the S&P 500's market value was in tech stocks, and TMT now makes up almost 53% of market capitalization.[2] Things could change, but at this rate, TMT is poised to not merely become larger than any other industry, but larger than all other industries *combined*—both in terms of value and contribution to economic growth. One reason for that is that other industries use TMT—tech and telecom specifically—to power their own AI innovations, and TMT happens to be the hardware, software, and services provider in the AI gold rush.

That said, other industries play critical roles. In both TMT Predictions 2025 and again in 2026, we have pulled in specialists from other Deloitte research centers and industries: energy, mining, and chemicals, manufacturing and construction, defense and aerospace, government and public services, and life sciences and health care. It takes some serious cross-industry collaboration to properly predict generative AI and agentic AI trends and implications.

Of our 13 topics for 2026, over half follow an AI theme. At a high level, we're seeing a narrative around making AI scale. New foundational models, or even shiny new enterprise agentic applications, continue to impress—but translating those beyond pilots and trials requires work that's typically considered less exciting, like data hygiene, integration into existing workflows, governance, new pricing models, and regulatory compliance. Those may be less glamorous than press releases about AI beating humans on a science test, but they will likely be more useful in the near term.

Gen AI and agentic AI are driving a lot of things that are very much here and now, but we also have an eye on the future. Although Deloitte predicts that robotics and drones will be slow but steady growers over the next year or two, the emergence of "physical AI" models is poised to transform both industries with massive acceleration in growth and usefulness. Meanwhile, newer forms of media, like short-form

vertical video series, appear to be crossing over from Asia to the rest of the world. And while the spread of gen AI–created images on social media may be exciting, it may also stimulate regulation.

A quick look at our 13 topics for 2026

Gen AI inside existing search engines overtakes standalone gen AI

Gen AI, possibly one of the most consequential technologies of our decade, may see its user base widen faster through its incorporation into existing mainstream digital applications than through its usage on a standalone basis

Deloitte predicts that in 2026 and beyond, more people will use gen AI when it's embedded within an existing application—such as a search engine—than those using a standalone gen AI tool. In terms of daily use, accessing gen AI within a search engine (when a search yields a synthesis of results) will be 300% more common than using any standalone gen AI tool. Standalone gen AI may require skill in prompt engineering and persistence, whereas passive gen AI is less overt and the experience more familiar; as such, demand is greater because it's more accessible. Going forward, standalone gen AI app owners will likely face a choice between embedding their tools' capabilities within another application or remaining a standalone interface.

Why AI's next phase will likely demand more computational power, not less

The world is moving from just training gen AI models to using them at scale. Many believe this means more consumer edge computing and less data center computing. Neither is likely to happen in 2026.

Deloitte predicts that "inference"—running AI models—will account for two-thirds of all AI computing power by 2026. Despite forecasts to the contrary, most inference will still take place in new data centers worth nearly half a trillion dollars and in on-premises enterprise servers using costly, power-intensive AI chips worth over $200 billion, rather than at the edge on inexpensive, lower-powered chips. There will be billions of dollars' worth of specialized chips optimized for inference, but they'll sit in data centers and enterprise servers as well, and some will use as much or even more power than general-purpose AI chips do.

Unlocking exponential value with AI agent orchestration

Autonomous AI agents may be transformational, but orchestration can be key for intelligent automation. Open-source and proprietary communication protocols will compete to lead the way.

On average, market estimates suggest that the autonomous AI agent market could reach $8.5 billion by 2026 and $35 billion by 2030. Deloitte predicts that if enterprises orchestrate agents better and thoughtfully address the associated challenges and risks, this market projection could increase by 15% to 30%—or as high as $45 billion by 2030. In 2026, businesses will likely work on their readiness to orchestrate agents with a specific degree of autonomy. Also, multi-agent systems will likely work for those businesses that focus on agent interoperability and management and redesign their workflows and talent effectively.

AI for industrial robotics, humanoid robots, and drones

Can more powerful AI models and chips catalyze what has been a relatively stagnant industry?

Deloitte predicts that the global cumulative installed capacity of industrial robots could reach 5.5 million by 2026, but annual new robot sales have stalled at just over half a million units since 2021. We could see an inflection point by 2030, with annual new robot shipments doubling from current levels to reach one million a year, driven by the following growth catalysts: (i) labor shortages in specialized industrial applications in developed countries and (ii) exponential advancements in computing power and the emergence of specialized foundational AI models. Robots can permeate multiple industries and applications, including autonomous drones, but unless the broader technology, AI, and robotics ecosystem addresses bottlenecks related to data quality, integration, and cybersecurity, the market for industrial robots may remain at its current level of relatively modest annual growth.

SaaS meets AI agents: Transforming budgets, customer experience, and workforce dynamics

As AI agents pervade the SaaS market, how businesses experience and leverage software will likely change —shifting business models, capabilities, and expectations

AI continues to disrupt the software as a service market. As agentic AI capabilities mature and vendors build out their platforms to create, integrate, and orchestrate AI agents, how organizations use and spend on SaaS could shift dramatically. In 2026, SaaS applications will likely become more intelligent, personalized, and autonomous, evolving toward a federation of real-time workflow services that can learn. Traditional pricing could shift away from seat-based and subscription licensing toward a more hybrid approach that blends consumption- and outcome-based models. In the longer term, some are even suggesting that sufficiently advanced agentic AI could replace existing enterprise SaaS. All these shifts will increase the complexity around financial planning, operations, ecosystem management, and value measurement.

New technologies and familiar challenges could make semiconductor supply chains more fragile

With escalating trade restrictions on critical next-gen AI chip technologies, leaders should adapt quickly to make supply chains more resilient

Making the most advanced chips has, for a long time, meant navigating fragile supply chains, but the stakes are much higher now. Extreme ultraviolet lithography has been restricted for years, but Deloitte predicts that in 2026, certain other advanced technologies and software tools that enable advanced AI models will become supply chain chokepoints. Many of these high-tech processes and materials rely on a handful of suppliers whose dominance in key regions has prompted governments to impose trade barriers to protect strategic interests and reduce dependency, underscoring the critical role they play in the global semiconductor supply chain.

Tiny episodes, massive appeal: Short-form serials are gaining viewers and empowering independent studios

From independent creators to major platforms, micro-series are helping redefine how viewers connect with and consume content worldwide

Micro-series—scripted video series told in bite-sized, mobile-first episodes—are reshaping global viewing habits. Micro-series apps now generate billions in revenue, with the United States leading growth. In 2026, Deloitte predicts that the revenue growth of in-app micro-series will more than double, reaching $7.8 billion. Deloitte also predicts that the United States will account for half of global revenue in 2025, but its share will decline to 40% as other markets convert more views and downloads into cash. Micro-

dramas blend short-form convenience with serialized storytelling, appealing to fragmented, mobile audiences. Uplifted by new technologies, independent creators are building studios that are lean and nimble, potentially challenging larger and more traditional studios.

Video podcasts dominate: Opportunity for brands, competition for traditional video

Podcasting is becoming a video-first, multilingual medium with booming reach that may help brands reach global audiences while occupying a larger share of viewers' screen time

Video podcasts (vodcasts) are transforming audience engagement by blending audio storytelling with visual appeal and may be competing directly with TV and streaming platforms. Deloitte predicts that annual global podcast and vodcast advertising revenues will reach approximately $5 billion in 2026—a nearly 20% year-over-year rise. Emerging markets such as India, Nigeria, and Brazil are fueling this growth through localized and multilingual content. Overcoming challenges related to discoverability, monetization, and scalability will likely be key to sustained growth.

A new era of self-reliance: Navigating technology sovereignty

Countries and regional blocs are racing to build out their own sovereign tech and AI infrastructures. What are the implications, and how can global businesses prepare?

As the global geopolitical environment becomes increasingly complex and uncertain, businesses and policymakers are urging their countries and regions to take greater direct control of their digital infrastructure, especially those parts related to AI. The desire for sovereignty is not new, but the shift toward technology sovereignty will likely quicken in 2026. Over the next decade, significant investment will flow into cloud computing, semiconductors, data centers, AI models, connectivity, and satellite communication efforts. In an interconnected world, total sovereignty is unlikely to be achieved by any country or region, but many are aiming to become at least more sovereign.

Generative AI video is perfect for social media, but could disrupt social media companies

Approaching Hollywood quality, the latest gen AI video models appear to be supercharging independent video but could provoke a stronger regulatory response against social video platforms

Generative video could empower independent creators and boost platforms' ad revenues—but it also risks overwhelming audiences, eroding authenticity, and fueling misinformation, likely intensifying regulatory scrutiny. Deloitte predicts that in 2026, generative video could provoke a regulatory response in the United States, potentially driving broader age verification in more states, refreshing federal challenges to Section 230 protections established in 1996 under the Communications Decency Act, and requiring labeling for AI-generated content published on social platforms. Success will likely hinge on balancing innovation with moderation, as unchecked generative video could disrupt business models, accelerate misinformation, and further fragment society's shared sense of reality.

Public media partnerships with streaming giants could be a model for making traditional TV sustainable

Public service broadcasters are publishing to social platforms, co-producing with streamers, and forming partnerships with the largest video distributors. They can offer lessons to for-profit US media companies.

Public service broadcasters (PSBs) are adapting to the pressures facing many traditional networks by coproducing with streamers, promoting content on social platforms, and experimenting with staggered releases. These strategies help extend reach, attract younger audiences, and inject local content into global platforms. In 2025, there was an acceleration, with three notable deals between broadcasters and streamers in just a few months. In 2026, Deloitte predicts another handful of broadcaster-and-streamer deals. Further, we also expect to see more coproductions and other initiatives—once again led by PSBs. Their adaptability can offer lessons for US broadcasters and niche studios facing similar disruption from streaming and social video. However, PSBs should be careful when navigating for-profit relationships that could threaten their mandates to represent the public.

Next-gen satellite internet is transforming pricing, capacity, and regulation worldwide

Satellite connectivity sees direct-to-device growth but often faces monetization hurdles, while low-Earth-orbit data expansion and tech advancements help reshape deployment and resilience, and create regulation complexities

Deloitte predicted spending on direct-to-device (D2D) network infrastructure—mainly satellites—at $3 billion in 2024, but it reached around $4 billion and is expected to rise to between $6 billion and $8 billion by 2026. Deloitte also predicts that around 1,000 D2D satellites will provide low-bandwidth connectivity services (SOS, text, and voice) in areas that may lack terrestrial cell coverage, with some D2D networks aspiring to provide higher-speed services. Adoption and willingness to pay for D2D remain uncertain, meaning monetization and business models for D2D are still unclear. We further predict that the number of communications satellites in low Earth orbit will reach between 15,000 and 18,000 satellites, connecting over 15 million global subscribers by the end of 2026. Another trend for 2026 in low Earth orbit will be new entrants that may disrupt emerging-market telcos with low-cost monthly broadband services, rather than partnering with terrestrial telcos as some other satellite providers are.

Gifts beat gigabits: Some mobile users rank rewards over network upgrades

Some consumers in developed markets struggle to perceive improvements in network performance. Telecom companies should consider more creative offerings to increase market share.

Deloitte predicts that in 2026, mobile operator reward schemes may matter to consumers in developed markets as much as—or even more than—network performance. In the medium term (the next five years through 2030), there is a reasonable probability that no new fundamentally revolutionary devices connecting to mobile networks will emerge, nor will there be any transformative applications running on these networks. Over the remainder of the decade, as network upgrades continue, non-network benefits may become increasingly critical to attract users or suppress churn. Such perks may be more tangible to consumers than network infrastructure upgrades.

BY

Gillian Crossan
Global

Tim Bottke
Germany

Girija Krishnamurthy
Global

Deb Bhattacharjee
United States

Jody McDermott
Canada

ENDNOTES

1. Nick Lichtenberg, "**Without data centers, GDP growth was 0.1% in the first half of 2025, Harvard economist says**," *Fortune*, Oct. 7, 2025.
2. Deloitte analysis of historical S&P500 data. As of December 31, 2008, technology weighting was 15.27% and communications services was 3.83%, for a combined TMT total of 19.1%. As of October 31, 2025, information technology weighting was 35.02%, communications services was 10.94%, and two consumer discretionary stocks that are generally considered tech stocks have a combined weighting of 6.68%, for a total of 52.64%.

ACKNOWLEDGMENTS

We wish to thank **Duncan Stewart**, **Jeff Loucks**, and **Paul Lee**, plus the entire team, for their work on the *TMT Predictions* report.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

Gen AI inside existing search engines overtakes standalone gen AI

Gen AI, possibly one of the most consequential technologies of our decade, may see its user base widen faster through its incorporation into existing mainstream digital applications than through its usage on a standalone basis

ARTICLE • 7-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

Deloitte predicts that the generative artificial intelligence user base in 2026 will surge, with the expansion mostly attributable to existing applications that incorporate gen AI capabilities. Deloitte also predicts that more people will use gen AI when it's within an existing application than those using a standalone gen AI tool. In short, passive usage will exceed proactive, explicit usage in 2026 and beyond.

Deloitte's forecast is that daily usage of gen AI within search—that is, when a search yields a synthesis of results—will be 300% more common than usage of any standalone gen AI tool with any focus: text, audio, image, video, code, or multimodal.[1] We forecast that in 2026, across developed markets, about 29% of adults will initiate one or more searches every day with results that incorporate a gen AI summary. This compares to 10% using any standalone gen AI app. We further predict that in 2027, daily usage of both search modalities will rise, but the 3:1 ratio will remain: Forty percent will use search overviews daily, versus 13% for any standalone gen AI app. Our forecast focuses on a single passive application for ease of comparison (figure 1).

Deloitte further predicts that passive usage of gen AI inside other applications will grow fastest among groups that are currently relatively low adopters, especially those in older age brackets.

Passive vs. standalone gen AI

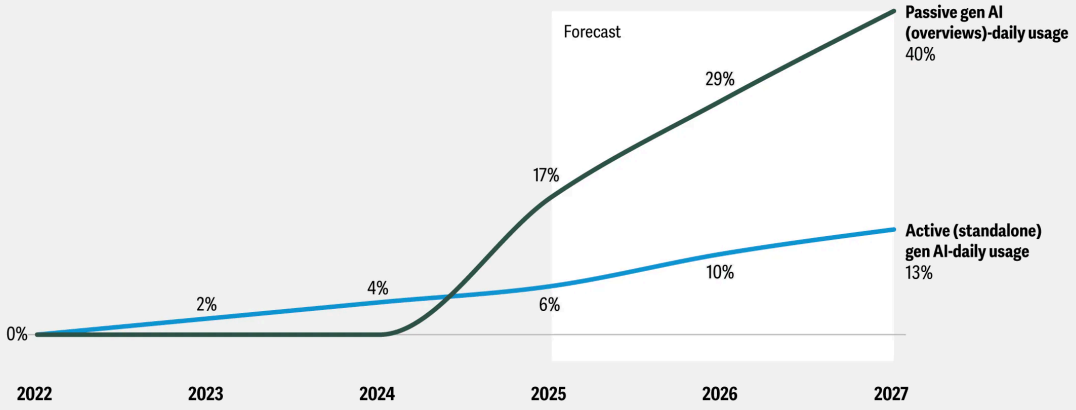
Common examples of where gen AI technology will be used passively include search, e-commerce, social media, and online news. This usage of gen AI inside existing apps contrasts with what we may term the "traditional" usage of standalone gen AI apps, such as ChatGPT or Gemini, which users open on their devices and use specifically to create an output—be it text, image, code, or another type.

With passive gen AI use, the technology is an embedded, essential but not overt capability within another application. The user is not explicitly using gen AI, but this technology is core to the experience. For example, gen AI may be used to synthesize numerous responses from a search; to summarize thousands of individual product reviews; or to create content disseminated via social media or online news.

Figure 1

Passive search summaries see higher daily usage than any standalone tool

Percentage of those who report daily usage of passive gen AI search overviews vs. any standalone gen AI tool, 2023–2027



Note: 2023–2025 data, weighted base. All respondents aged 16–75 years; 2023 (4,150), 2024 (4,150), 2025 (4,150).

Source: Deloitte forecasts based on Deloitte Digital Consumer Trends, UK, 2025.

Deloitte. Insights | deloitteinsights.com

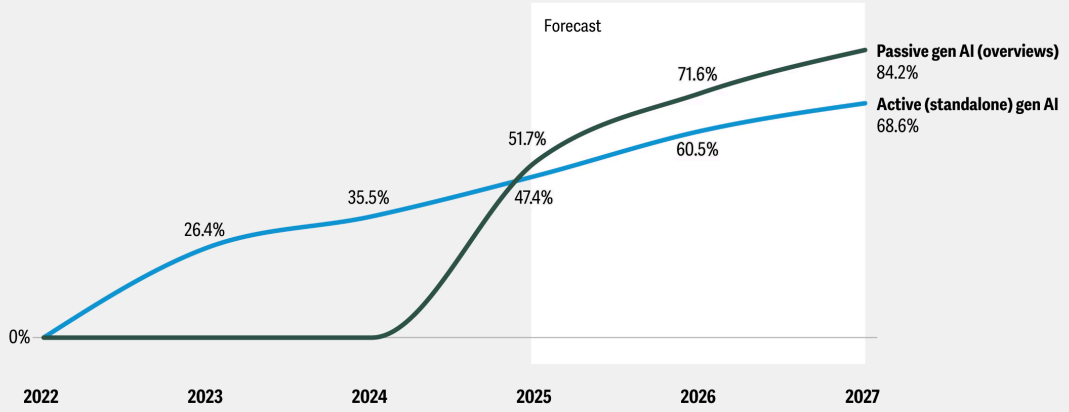
Deloitte estimates that comparing the usage of *all* passive gen AI apps relative to *all* standalone, proactively used apps would show the former already being notably more popular in 2025 as well. In the UK market, where emerging products are often early to launch, Deloitte UK's research found that as of mid-2025, about three-quarters of respondents had ever used one of four types of passive gen AI applications[2]—notably ahead of the 47% of respondents who had ever used any dedicated, standalone gen AI app.

Another metric for emerging applications is comparing usage at any time in its relatively short history. Passive gen AI applications were first launched in the United States in May 2024 with the introduction of search summaries,[3] and rollout into additional markets was announced in November of that year[4]—almost two years after the launch of the first popular standalone gen AI apps in late 2022. Despite standalone apps having this lead, we forecast that by mid-2026, more adults will have generated a search overview (72%) than those who have used a standalone gen AI tool at any time (61%) (figure 2).

Figure 2

More report having ever used passive gen AI search summaries than having ever used standalone gen AI tools

Percentage of those who report ever using passive Gen AI search overviews vs. any standalone gen AI tool, 2023–2027



Note: Weighted base. All respondents aged 16–75 years; 2023 (4,150), 2024 (4,150), 2025 (4,150).

Source: Deloitte forecasts based on Digital Consumer Trends, UK, 2025.

Deloitte. Insights | deloitteinsights.com

The prediction implies that gen AI, as a fundamental process within an existing mainstream application, will be significantly more pervasive and ubiquitous than as a standalone destination. If our prediction is correct, this does not imply that standalone gen AI, *per se,* is not useful; rather, it indicates that this technology, when integrated into an application that is already mainstream, is likely to be far more commonly used and, as such, may deliver greater overall utility. There is, of course, a read-through on the medium-term penetration of dedicated gen AI: Will it ultimately become as popular as online services like social media or search? Or will it plateau at about a fifth of all web users who use any dedicated tool daily?

What can we learn from user preferences for passive search?

Search, social media, and e-commerce are already among the most frequently used digital applications. There are over 15 billion searches undertaken every day. On average, users spend over two hours on social media daily.[5] E-commerce sales in Q1 2025 alone in the United States totaled $300 billion.[6] Users may be more likely to use gen AI capabilities within a familiar search tool rather than search within an unfamiliar, novel gen AI chatbot.

In 2026, questions about gen AI's impact on the viability of the search business model may continue, but there may also be questions about the impact of gen AI-enhanced search on the popularity of standalone gen AI tools such as ChatGPT or Synthesia.[7] According to Deloitte's research, the most common workplace application for gen AI is search. It may be that some users who currently search within standalone gen AI apps move back to mainstream search applications.

The pace at which passive gen AI has overtaken standalone gen AI is impressive and, perhaps, predictable. Standalone gen AI is both presented and perceived as novel and relatively experimental. It requires skill and persistence: a disappointing outcome may result from a poor prompt rather than a flawed model, and the remedy is to re-prompt.[8] It may be the user's prompt-engineering skills that are blamed rather than the product. Passive gen AI should be lower friction if it's an incremental capability that is seamlessly integrated into an existing mainstream digital application, be it search engine, e-commerce site, social media app, or office productivity tool. There is rarely a need to try again. The technology is less overt, the experience more familiar, and, as such, the demand is greater because it's more accessible. The application of gen AI to create a summary of search results automates and completes a task that many users otherwise would have done manually—that is, click on and read multiple links to

formulate a personal summary, a chore that is eliminated by the AI summary. The integration of gen AI into an existing application is akin to one-touch checkout, including payment, integrated into e-commerce sites, or facial-recognition authentication incorporated into a consumer banking app.

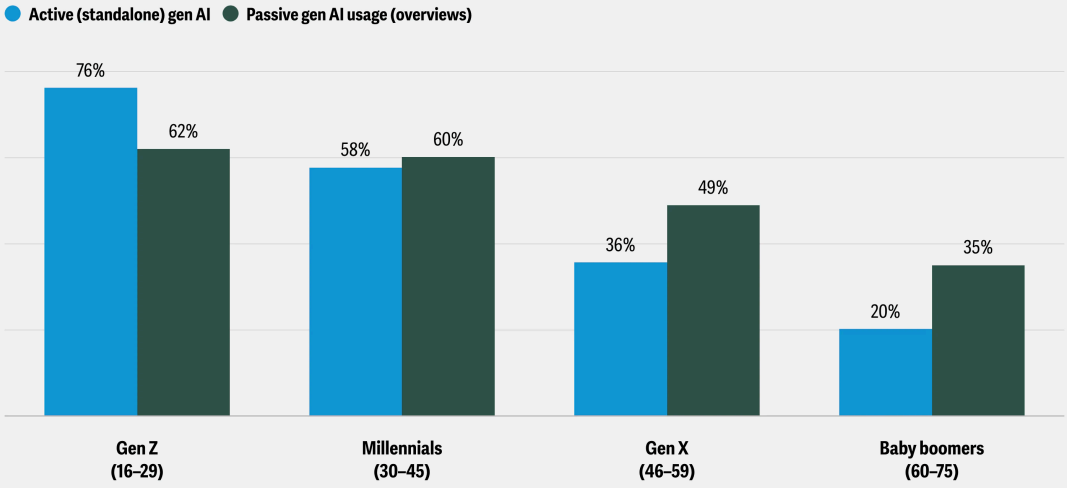
Adoption trends across generations

Passive AI's accessibility is evident in the rapid adoption of search summaries among older age groups, who may be less inclined to master new standalone tools. As of mid-2025, boomers were hesitant about standalone gen AI. Deloitte's research found that only 20% of boomers had ever used any generative AI tool—despite an awareness rate of 58%. By contrast, almost four times as many (76%) members of Generation Z had used a gen AI tool in 2025 (figure 3). However, adoption of search overviews was 75% higher among boomers—at 35%—relative to any standalone tool.

Deloitte forecasts that passive gen AI usage among boomers will grow at a faster rate than standalone gen AI, with adoption reaching 49% for search overviews in 2026 and 59% in 2027—the latter markedly higher than the 32% usage of standalone gen AI (figure 4).

Figure 3

Older generations in 2025 use passive gen AI search more than they use any standalone gen AI tool for any purpose



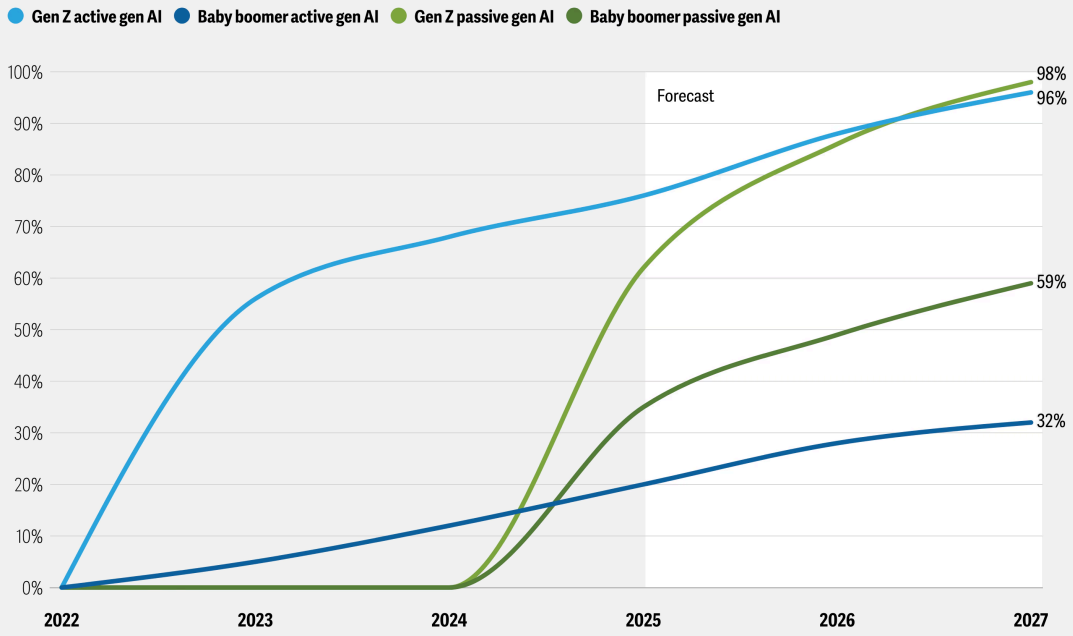
Note: Weighted base. All respondents aged 16–75 years; 2025 (4,150).

Source: Deloitte forecasts based on Digital Consumer Trends, UK, 2025.

Figure 4

Year-one passive gen AI search overview usage outpaced year-one standalone usage among both baby boomers and Gen Z

Percentage of those in both boomer and Gen Z cohorts who report ever using standalone gen AI or passive gen AI search overviews



Note: Weighted base. All respondents aged 16–75 years; 2023 (4,150), 2024 (4,150), 2025 (4,150).

Source: Deloitte forecasts based on Deloitte Digital Consumer Trends, UK, 2025.

Deloitte. Insights | deloitteinsights.com

Bottom line: Passive AI usage has market implications for gen AI

Gen AI is one of the most important technologies of its time, but its fullest potential may only be realized when it's deployed additionally as a discreet, yet integral, capability within existing, mainstream applications.

Many of today's most important technologies began as standalone capabilities, often within dedicated devices. It was not long ago that GPS, or sat-nav referred to a physical appliance so useful that users took it on work trips and vacations. This functionality was then integrated into smartphones and their applications. Now, satellite navigation is integrated into myriad applications beyond route finding—its usage is vital, ubiquitous, and largely in the background.

Gen AI often improves existing applications, even if it may not make them perfect. It can summarize search results, and while it may introduce errors when doing so, in many cases this may not matter. Further, users may trade the simplification of the search process enabled by gen AI for the errors that may be introduced by the technology's inherently probabilistic approach.

For the many standalone gen AI app owners in the market, a core question to address in 2026 will be to consider choosing between focusing on embedding their tools' capabilities within another application or to remain as a standalone interface—the latter approach generating higher revenue per user but potentially lower adoption. A few players will be able to do both, but for the remainder, a choice may need to be made.

Paul Lee
United Kingdom

Ben Stanton
United Kingdom

Gillian Crossan
Global

Girija Krishnamurthy
Global

Tim Bottke
Germany

Steve Fineberg
United States

ENDNOTES

1. Deloitte's forecast is based on multiple sources, including its proprietary research undertaken as part of Deloitte's Digital Consumer Trends survey, fielded in April and May 2025, and also in 2023 and 2024. This longitudinal data set provides a trajectory for the adoption of standalone gen AI apps. Our proprietary data set includes surveys conducted in multiple developed markets globally. Additionally, we have considered multiple other data points, including Alphabet's reporting on the volume of AI Overviews, which had a monthly usage base of over two billion as of July 2025. See: Alphabet, "Alphabet announces second quarter 2025 results," July 23, 2025.
2. Paul Lee and Ben Stanton, "Digital Consumer Trends 2025, UK edition," Deloitte, June 2025.
3. Elizabeth Reid, "Generative AI in search: Let Google do the searching for you," Google, May 14, 2024.
4. Hema Budaraju, "New ways to connect to the web with AI Overviews," Google, Aug. 15, 2024.
5. Josh Howarth, "Worldwide Daily Social Media Usage (New 2025 Data)," Exploding Topics, June 23, 2025.
6. United States Census Bureau, "Quarterly retail e-commerce sales," press release, Aug. 19, 2025.
7. Danny Goodwin, "Google search is 373x bigger than ChatGPT search," *Search Engine Land*, March 11, 2025.
8. Kara Kennedy, "Poor prompts lead to misleading research," AI Literacy Institute, Aug. 19, 2024; Ulster University, "Generative artificial intelligence (Gen AI): Prompt engineering," Oct. 23, 2025; Haringun Nur Adha, "You made a specific prompt but the results are disappointing? Maybe you're using ChatGPT wrong," Medium, Sept. 16, 2025.

ACKNOWLEDGMENTS

The authors would like to thank **George Johnston**, **Pedro Barros**, **Stephen Hipkiss**, **Lorraine Barnes**, **Nick Seeber**, **Chris Arkenberg**, **Duncan Stewart**, **Susanne Hupfer**, **Debapratim De**, **Cornelia Calugar Pop**, **Eytan Hallside**, **Jonas Malmlund**, **Steve McMullen**, **Ian Stewart**, **Rupert Darbyshire**, **Andy Cowen**, and **Ralf Esser** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

Why AI's next phase will likely demand more computational power, not less

The world is moving from just training gen AI models to using them at scale. Many believe this means more consumer edge computing and less data center computing. Neither is likely to happen in 2026.

ARTICLE • 11-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

It's widely expected that generative AI computing will shift in 2026, from mainly being about *training* models on very large amounts of data to *using* those models to help think about and answer enterprise and consumer questions, prompts, and tasks—a process known as "inference." Many speculate that such a shift in computational workload—or "compute"—would mean that the AI ecosystem would need special chips optimized for inference only, and that these (possibly much cheaper) chips might be deployed on edge devices outside of the massive data centers where most AI chips are currently located and might even mean we need fewer, smaller, or at least different data centers, and spend less.

Deloitte sees things somewhat differently. Inference workloads will indeed be the hot new thing in 2026, accounting for roughly two-thirds of all compute (up from a third in 2023 and half in 2025),[1] and the market for inference-optimized chips will grow to over US$50 billion in 2026. But Deloitte also predicts that a majority of the computations will still be performed on cutting-edge, expensive, power-hungry AI chips worth US$200 billion or more, which will still mainly sit in large data centers valued at US$400 billion or more, or on-prem enterprise solutions worth US$50 billion that use the same chips and racks as data centers, rather than on chips used in edge devices. Meaning, we likely will need all the data centers and enterprise on-prem AI factories that are currently being planned and all the electricity that these facilities will need.

The ever-growing computational demands of AI

While the growth in demand for training compute on new models has likely slowed (it is likely still growing, but at lower rates than in 2023 and 2024),[2] AI models continue to evolve through advanced techniques that can improve them after training. These methods, combined with the sheer volume of inference queries, likely mean that computational demand will increase, not decrease. Put another way, it is expected that, even though the chips used for compute are becoming more efficient every year, thanks to Moore's Law, the demand for compute is rising even faster at four to five times per year out to 2030.[3]

It's true: Compute demand growth is slowing for initial training

A 2020 paper showed that bigger models, trained on more data, and using more advanced AI-accelerating chips produced better results consistently: This was gen AI's first scaling law.[4] By 2022 and 2023, training models had grown from one billion parameters to 100 billion to one trillion.[5]

Two issues began to emerge in 2024. There wasn't an infinite amount of training data out there, and ever larger models were showing diminishing returns: Ten times more training data might produce a "state-of-the-art" AI model that was only slightly better than the previous version, or perhaps not even better at all.[6] At the same time, smaller and more efficient AI models looked like they might be able to produce truly state-of-the-art AI models using less data, less time, less money, and fewer chips.[7]

If growth in training slowed, then AI computing would become increasingly about inference. Asking a large language model (LLM) to summarize a document (one example of inference) takes only a tiny fraction of the compute capacity needed to train that model. However, the logic went, as billions of consumers and enterprise workers made more of those requests and more frequently, all that inference would add up, shifting the overall compute workload from training to inference. Some of those requests could be processed on consumer and enterprise devices such as smartphones and personal computers, and, as Deloitte correctly predicted in 2024, hundreds of millions of PCs and smartphones with on-device AI-accelerating chips were sold in 2025.[8] Also, since inference is less computationally intensive than training, perhaps special inference-optimized chips could be used inside data centers. These chips are cheaper and use less energy per inference than the superpowered AI chips needed for scaling training, and might not require as much co-packaged expensive memory.[9]

All of that is happening in 2025 and will likely continue in 2026. Deloitte surveys in 2025 found that, both in the United States and globally, more consumers are using gen AI, and more are doing it daily.[10] Edge devices such as PCs and smartphones increasingly have onboard AI accelerators. A number of inference-optimized chips (application-specific integrated circuits, or ASICs) have been designed, manufactured, and are being deployed in data centers and some edge devices. The list includes, but is not limited to, chips from Meta, Google, Amazon, Intel, AMD, Qualcomm, Groq, SambaNova, Cerebras, and Graphcore, some of which are based on a Broadcom package solution, with the designer providing the processing core.[11] Although sales figures for all these different chips are not publicly available, Deloitte believes that 2025 revenues for these chips are over US$20 billion collectively and will reach US$50 billion or more in 2026.[12]

Then why do we still need power-hungry chips costing US$30,000 each or more—US$400 billion or so by 2028 in aggregate[13]—in data centers that will cost an estimated US$400 billion in 2026 alone, rising to a potential trillion dollars annually in the same year?[14]

AI model training is more complex than it used to be

The point of the first scaling law was to produce "better" AI models, and it worked very well, at least for a few years. This initial form of scaling used to be called training, but is now called "pre-training," producing foundational models.

It turns out there are two more ways to make even better models: One is "post-training" scaling, which involves various techniques such as fine-tuning, pruning, quantization, distillation, reinforcement learning from both human feedback and increasingly from AI feedback, and synthetic data augmentation.[15] The other is test-time scaling, or long thinking, in which the models reason their way through the inference process after they have been asked a question using a variety of techniques, such as chain-of-thought prompting, sampling with majority voting, search, and even some post-training techniques.[16] This allows for more accuracy, with more choices, better sources, and fewer hallucinations.[17]

New power-hungry AI techniques will likely outpace efficiency gains

First, post-training scaling and test-time scaling appear to be the new normal: Most AI companies now use them to make AI models better in various ways.[18]

Second, they're both AI compute hogs. It's estimated that post-training in aggregate uses 30 times the compute needed to train the original foundational model, while long thinking uses more than 100 times the compute of a simple inference like asking an AI to summarize an email.[19]

Third, since both of these scaling techniques are widely used and computing resource–intensive, there are implications for AI data centers, the locations and power needs of those AI data centers, the chips that go into AI data centers (and other places where AI needs to be performed), last year's AI chips, edge devices, and more.

A brief refresher on the chips Deloitte predicted would be needed in AI data centers in 2025 and beyond

Data centers have existed for decades. In fact, there are tens of millions of square feet of data centers globally, and tens of billions of dollars of semiconductor components have been sold annually to fill those data centers for years.[20] But the new AI data centers, and the new semiconductors that enable them, are often radically different from yesterday's data centers and semiconductors. Night and day.

The next generation of AI data centers is likely to cost hundreds of billions of dollars annually to build and consume hundreds of gigawatts of power. In most of these facilities, the cooling will likely be different from previous generations of data centers, the power supplies and voltages will likely be different, the internal communications technologies will likely be different, and the very floors will likely need to be thicker to support denser and heavier server racks. Perhaps most importantly, instead of having central processing unit–centric servers with memory close by, newer AI server racks are mainly made up of specialized chips called graphics processing units, or GPUs,[21] which often have specialized high-bandwidth memory (HBM), tightly integrated with the GPUs, and special central processing units (CPUs) to orchestrate the vast AI compute workloads. Many components are unique to the needs and scale of this newer generation of AI data centers.[22]

As recently as 2006, high-end GPUs were thought to be for gaming computers and boxes, not data centers.[23] The tasks of most data centers were well met by CPUs, which were largely serial processors, where tasks were executed in order. Some high-performance computers, or "supercomputers," have special chips in them, which are called "massively parallel processors," that execute hundreds of tasks simultaneously. These chips, however, were often tens or hundreds of times more expensive than gaming GPUs or data center CPUs.

In 2009, scientists noted that gaming GPUs were also parallel processors and tried running machine learning models on high-end GPUs—the exact same GPUs as were found in gaming computers.[24] They worked well, and within a few years, specially optimized GPUs (slightly different from the gaming versions) were being used in some data centers and some on-premises devices to perform machine learning AI.[25] But the market was measured in single-digit billions of dollars annually as recently as 2018.[26]

In 2022, the development of LLMs for generative AI required even further specialized GPUs, and often required them to be integrated in the same advanced package with a relatively new and specialized kind of memory: HBM.[27] These GPU plus HBM components also required a device to coordinate and orchestrate data flows. Optimized special CPUs (different from the CPUs in computers, smartphones, or data centers, although similar in their core architecture) were also an important part of the generative AI data center, along with multiple other, perhaps equally critical, components. In 2025, almost all the top 500 supercomputers in the world have a similar mix of GPUs, special memory, and CPUs.[28] In a way, the megascale AI data centers that are being built could be described as a version of specialized supercomputers.

Bottom line: What more compute demand could mean for the AI ecosystem

Businesses and executives should prepare for a future where compute demand, especially in big data centers and enterprise AI factories, continues to rise, driven in part by post-training and test-time scaling. There will likely be growth in inference-optimized chips and in edge processing, but there will still be a need to invest in hyperscale data centers and enterprise AI boxes. "Optimized for inference" doesn't necessarily mean less power: One recent product optimized for inference pre-fill compute avoids using HBM and uses GDDR7 instead, but each rack needs 370 kW, which is almost three times the power density of the training version from the same supplier.[29]

AI data centers: AI data center capital expenditure for 2026 is expected to be US$400 billion to US$450 billion globally,[30] with over half of that spending being the chips inside devices (US$250 billion to US$300 billion)[31] and the rest being everything else (land, construction, power, permitting, and more). It's further predicted that AI data center capex will rise to US$1 trillion in 2028,[32] with AI chips being over US$400 billion in that year.[33] Although pre-training growth is slowing, and compute is shifting from training to inference, the compute demands from post-training scaling and test-time scaling, and increased usage suggest that the world likely needs a lot of data centers, and the ramp from US$300 billion to US$400 billion in 2025 to roughly US$1 trillion in 2028 is directionally realistic.

Location of AI data centers: Pre-training a 100-trillion-parameter LLM could take weeks and can be incredibly sensitive to small interruptions. The failure of a key component or an excessively high latency handoff between processors could lead to the loss of all the work thus far and require a fresh start. Most foundational model pre-training has been co-located, with all the servers and racks inside a single building or campus. However, increasingly, AI compute loads are able to be done in different data centers across the United States, or even around the globe.[34] Further, there will likely be a range from gigawatt-scale data centers to smaller-scale inferencing data centers where fully trained models can be deployed, which will tend to be closer to metro locations to help reduce latency. This helps set up a growing demand for **sovereign AI solutions** (each country or region having its own domestically located and even locally operated AI compute capacity) as well as enterprise edge on-premises solutions as part of the hybrid cloud.[35]

Power demand for AI data centers: At a high level, more AI data centers that are doing all three kinds of scaling are still going to need a lot of power. But the ability of both post-training and test-time scaling can be relatively "interruptible" compared to pre-training, which needs to be done all in one training run. That helps allow AI companies to participate in demand response programs, where they can shift tasks to different data center locations or slow down processor clock speeds, reducing demand during peak times.[36] It's estimated that increasing this kind of flexible load could allow large new data centers to help maintain grid reliability and affordability.[37]

That AI training and inference can be distributed means that data centers don't need to all be in one state or one county, but can be spread more evenly around the world, distributing electrical demand.

Chips in AI data centers: Some may have viewed the AI chip market as a zero-sum game. The view was often something along the lines of: "Sure, I needed to spend tens of thousands of dollars for advanced GPUs co-packaged with HBM for pre-training my foundational models, but as we shift computing to inference, maybe I can use cheaper chips that are optimized for inference and have less HBM."

Instead of the chip market being an "either-or," it looks like it will be a "both-and." There's likely to be considerable growth in inference-only or inference-optimized chips, but at the same time, the kind of chips typically best suited for foundational model pre-training, post-training, and test-time scaling (which are a mix of training and inference compute) remain the big, powerful, energy-hungry GPUs with HBM

that cost tens of thousands of dollars each. For those buying the chips, they may be even more expensive in 2026, with leading-edge process wafers expected to cost 50% more.[38]

Edge AI in consumer or enterprise devices such as smartphones and PCs: As mentioned earlier, hundreds of millions of smartphones and PCs are being shipped and purchased with neural processing units (NPUs):[39] dedicated chips or portions of the CPU chip (worth a few dollars or tens of dollars for the NPU portion) that are optimized for processing AI inference tasks with reasonable power consumption.

However, NPUs are only powerful enough for the kind of one-shot inference discussed earlier ("summarize this email," etc.). Therefore, Deloitte predicts that almost all AI computing performed in 2026 will be done mainly in the kind of giant AI data centers being planned, or on relatively expensive high-end AI servers owned by enterprises, not on PCs and smartphones. At least for now, in the hyper-growth, land-rush phase we seem to be in, a cost-optimized hybrid architecture does not appear to be a priority for vendors or enterprises. Further, things like test-time scaling can be overkill for the vast majority of consumer use cases, and even most enterprise on-device use cases. One day, computers and smartphones may have a much bigger role to play, but it won't likely be in 2026.

More recently, one AI company introduced a gen AI model that can reason and that runs locally on PCs. It's unclear how well it works, what impact it has on battery life, or how many PC users will want to use AI locally, rather than through the cloud.[40]

Edge AI and the enterprise using on-prem solutions: The very powerful, power-hungry GPU plus HBM plus coordinating CPU trays that are typically going into giant AI data centers around the world are also available to enterprises that want to pursue an on-prem, hybrid, more resilient approach to gen AI computing, especially for post-training. Driven by concerns around cost, intellectual property ownership, sovereignty, resilience, and customization, enterprises can spend US$300,000 to US$500,000 on a box with about eight GPUs (and HBM and CPUs) that can perform a certain level of AI training and inference.[41] Or they can spend US$3 million to US$5 million on a rack with up to 72 cutting-edge GPUs (and HBM and CPUs) that do more.[42] Or they can even spend tens of millions on multiple racks that do more still.[43] Deloitte predicts that this on-prem hybrid enterprise market will be worth over US$50 billion in 2026.

Edge AI for robots, drones, and autonomous vehicles: Still comparatively small in 2026, there are several use cases that can require inference in real time and on device. These range from drones and robots to self-driving cars. These currently span a wide variety of chips: Most drones have relatively primitive and low-powered AI inference chips,[44] while most self-driving vehicles are using GPU chip solutions that are only slightly less powerful than those found in data centers.[45] This non-AI factory market is likely still fairly small (under US$5 billion in 2026)[46] but could become **much larger, especially if the robot market takes off, which could happen, but likely after 2030.**[47]

We're still in the early days of AI. As of summer 2025, the growth in the need for AI compute (and therefore the need for more data centers, enterprise on-prem solutions, and more high-powered AI chips, whether for pre-training, post-training, test-time scaling, and inferencing) is very high, even in spite of constant attempts to make the algorithms more efficient.[48] At some point, it's possible that new techniques could see a breakthrough, and improved AI models could run well on cheaper chips, needing fewer data centers and less power. But that won't be in 2026.

BY

Duncan Stewart
Canada

Jeroen Kusters
United States

Deb Bhattacharjee
United States

Arpan Tiwari
United States

Girija Krishnamurthy
Global

Karthik Ramachandran
India

ENDNOTES

1. Rodrigo Liang, "Scaling AI without breaking the grid: The path to sustainable innovation," World Economic Forum, Jan. 3, 2025.

2. Michelle Weaver, "Big debates: The AI evolution," Morgan Stanley, Jan. 10, 2025.

3. Josh You and David Owen, "How much power will frontier AI training demand in 2030?" Epoch.AI, Aug. 11, 2025.

4. Jared Kaplan et al., "Scaling laws for neural language models," OpenAI, Jan. 23, 2020.

5. Amazon Web Services, "What are foundation models?" accessed Sept. 19, 2025.

6. Ashu Garg, "Has AI scaling hit a limit?" Foundation Capital, Nov. 27, 2024.

7. Aixin Liu et al., "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437* (2024).

8. Chris Arkenberg, Duncan Stewart, Gillian Crossan & Kevin Westcott, "On-device generative AI could make smartphones more exciting—if they can deliver on the promise," *Deloitte Insights*, Nov. 19, 2024; IDC Media Center, "Worldwide smartphone market forecast to grow 1% in 2025, driven by accelerated 3.9% iOS growth, according to IDC," Aug. 27, 2025; Gartner, Inc., "Gartner says artificial intelligence ("AI") PCs will represent 31 percent of worldwide PC market by the end of 2025," press release, Aug. 28, 2025.

9. Amazon Web Services, "AWS Inferentia," accessed Sept. 19, 2025.

10. Paul Lee and Clare Mortimer, "How citizens use devices and AI: what government needs to know," Deloitte UK, Aug. 29, 2025; Steve Feinberg, et al., "In the gen AI economy, consumers want innovation they can trust: Deloitte's 2025 Connected Consumer Survey," Deloitte, Sept. 25, 2025.

11. Wylie Wong, "Data center chips in 2024: Top trends and releases," Data Center Knowledge, April 11, 2024; Reen Singh, "AI inference chips latest rankings: Who leads the race?" Uvation, July 11, 2025; Broadcom Inc., "3.5D XDSiP AI Accelerator Platform," accessed Oct. 23, 2025.

12. Deloitte Consulting LLP performed an analysis of the data center market, including a rough bill of materials for the various components, and market sizes. This analysis is due to be published in December 2025.

13. Skye Jacobs, "NVIDIA Blackwell server cabinets could cost somewhere around $2 to $3 million each," TechSpot, July 28, 2024.

14. Beth McKenna, "2 key things from AMD's earnings call that investors should know," The Motley Fool, Feb. 1, 2024; Dell'Oro Group, "AI infrastructure spending sustains strong growth momentum," press release, Feb. 5, 2025.

15. Kari Briski, "How scaling laws drive smarter, more powerful AI," NVIDIA, Feb. 12, 2025.

16. Ibid.

17. Jonathan Farrington, "What is chain of thought prompting – AI prompt engineering," Silicon Dales, July 24, 2025.

18. Briski, "How scaling laws drive smarter, more powerful AI."

19. Ibid.

20. "Data centers: Computing risks and opportunities for U.S. real estate," S&P Global, Oct. 22, 2024; Equinix, Inc., "Form 10-K: Annual report for fiscal year ended Dec. 31, 2023," Feb. 16, 2024; Digital Realty Trust, Inc. and Digital Realty Trust, L.P., "Form 10-K: Annual report for fiscal year ended Dec. 31, 2023," Feb. 23, 2024.

21. Shubham Sharma, "Going beyond GPUs: The evolving landscape of AI chips and accelerators," VentureBeat, Sept. 26, 2024.

22. Deloitte Consulting LLP performed an analysis of the data center market, including a rough bill of materials for the various components, and market sizes. This analysis is due to be published in December 2025.

23. Eric Reed, "History of NVIDIA: Company and stock," SmartAsset, May 22, 2024.

24. Rajat Raina, Anand Madhavan, and Andrew Y. Ng, "Large-scale deep unsupervised learning using graphics processors," In Proceedings of the 26th Annual International Conference on Machine Learning, 2009.

25. NVDIA, "NVIDIA delivers massive performance leap for deep learning, HPC applications with NVIDIA Tesla P100 accelerators," press release, April 5, 2016.

26. Hannah Wilson, "NVIDIA facts and statistics (2025)," Investing.com, Aug. 28, 2025.

27. Hannah Wilson, "NVIDIA facts and statistics (2025)," Investing.com, Aug. 28, 2025.

28. Top 500, "June 2025," June 2025.

29. Ray Wang, "NVIDIA's new Rubin CPX targets future of large-scale inference," Futurum, Sept. 18, 2025.

30. In 2025, Deloitte Consulting LLP performed an analysis of the data center market, including a rough bill of materials for the various components, and market sizes. This analysis is due to be published in December 2025.

31. Omdia, "New Omdia forecast: AI data center chip market to hit $286bn, growth likely peaking as custom ASICs gain ground," Aug. 28, 2025.

32. Anthony Di Pizio, "Jensen Huang predicts annual data center spending will hit $1 trillion by 2028. Here's the ultimate semiconductor ETF to buy right now." The Motley Fool, May 1, 2025.

33. Dave Lawler, "Exclusive: 'Massive ten-year' AI boom is just starting, AMD CEO says," *Axios*, Sept. 17, 2025.

34. Paul Mah, "AI training is going to multiple data centers," CDO Trends, Sept. 11, 2024.

35. Chris Thomas, Akash Tayal, Duncan Stewart, Diana Kearns-Manolatos, and Iram Parveen, "Is your organization's infrastructure ready for the new hybrid cloud?" *Deloitte Insights*, June 30, 2025.

36. Mike Robuck, "Google strikes deals for flexible AI data centre power use," Mobile World Live, Aug. 5, 2025.

37. Tyler H. Norris, Tim Profeta, Dalia Patino-Echeverri, and Adam Cowie-Haskell, "Rethinking load growth: Assessing the potential for integration of large flexible loads in US power systems," Nicholas Institute for Energy, Environment & Sustainability, Duke University, February 2025.

38. Anton Shilov, "TSMC could charge up to $45,000 for 1.6nm wafers — rumors allege a 50% increase in pricing over prior-gen wafers," Tom's Hardware, June 4, 2025.

39. Francisco Jeronimo, "The rise of gen AI smartphones," IDC, July 5, 2024.

40. Dan Shipper, "Vibe check: OpenAI drops two new open-weight models," Every Media, Aug. 5, 2025.

41. Cyfuture Cloud, "NVIDIA DGX H100 price 2025: Cost, specs, and market insights," Cyfuture Cloud Knowledgebase, accessed October 2025.

42. Tae Kim, "NVIDIA's multi-million dollar AI servers are getting more expensive," Barron's, Aug. 28, 2025.

43. Skye Jacobs, "NVIDIA Blackwell server cabinets could cost somewhere around $2 to $3 million each," TechSpot, July 28, 2024.

44. Qualcomm, "Flight RB5 5G platform," accessed Sept. 19, 2025.

45. Ali Kani, "NVIDIA DRIVE Thor strikes AI performance balance, uniting AV and cockpit on a single computer," NVIDIA, Sept. 20, 2022.

46. There are a variety of suppliers for chips for driving assistance, but as one example, NVIDIA's auto segment is at a US$2 billion run rate as of August 2025: Pras Subramanian, "NVIDIA's auto business surges 69% from self-driving tech," Yahoo Finance, Aug. 25, 2025.

47. Karthik Ramachandran, et al, "AI for industrial robotics, humanoid robots, and drones," Deloitte Insights.

48. Jameel Rogers, "AI chips for data center and cloud to exceed US$400 billion by 2030," IDTechEx, May 8, 2025.

ACKNOWLEDGMENTS

The authors would like to thank **Brandon Kulik**, **Amy Scimeca**, **Karan Aggarwal**, **Kate Hardin**, **Diana Kearns-Manolatos**, **Mike Luk**, **Baris Sarer**, **Jason Chmiel**, **Dan Hamling**, **Jan Nicholas**, **Jordan Bish**, **Nitin Mittal**, **Rohit Tandon**, **Nicholas Merizzi**, and **Dan Littmann** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

Unlocking exponential value with AI agent orchestration

Autonomous AI agents may be transformational, but orchestration can be key for intelligent automation. Open source and proprietary communication protocols will compete to lead the way.

ARTICLE • 9-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

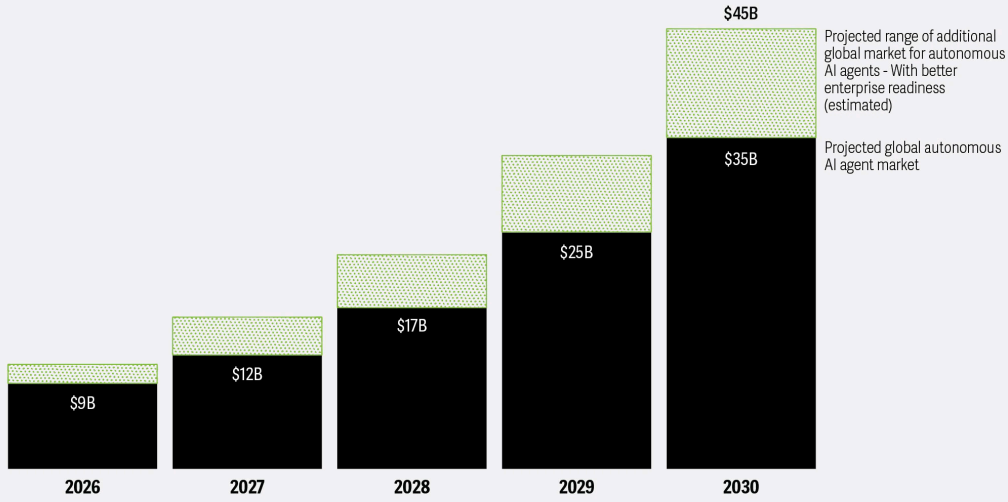
As companies integrate multiagent systems—where different AI reasoning engines interact seamlessly across domains—agent orchestration (the effective coordination of role-specific agents) will be essential to help unlock their full potential. Thoughtful orchestration unleashes intelligent workflows by enabling multiagent systems to interpret requests, design workflows, delegate and coordinate tasks, and continuously validate and enhance outcomes.[1]Conversely, poor agent orchestration can significantly limit this business value.

On average, market estimates suggest that the autonomous AI agent market could reach US$8.5 billion by 2026 and US$35 billion by 2030 (figure 1).[2] **Deloitte predicts** that if enterprises orchestrate agents better and thoughtfully address the associated challenges and risks, this market projection could increase by 15% to 30%—or as high as US$45 billion by 2030. According to an estimate, more than 40% of today's agentic AI projects could be cancelled by 2027, due to unanticipated cost, complexity of scaling, or unexpected risks.[3] These projects could drive significant revenue growth if enterprises remediate the potential pitfalls preemptively.

Figure 1

The AI agent market may expand with better enterprise readiness to orchestrate agents

Projected global autonomous AI agent market



Note: All numbers have been rounded to the nearest whole number.
Source: Deloitte analysis.

Deloitte. Insights | deloitteinsights.com

To leverage multiagent systems fully, businesses will likely work on their readiness to orchestrate agents with a specific degree of autonomy and address the early potential pitfalls. At the same time, multiagent systems will likely work for those businesses that focus on agent interoperability and management and implement the required changes in workflows and talent, effectively.

Making businesses work for multiagent systems

As businesses work through decisions related to their agent orchestration preparation, these three guideposts will likely be pivotal.

From single-purpose agents to multiagent systems: Are enterprises ready?

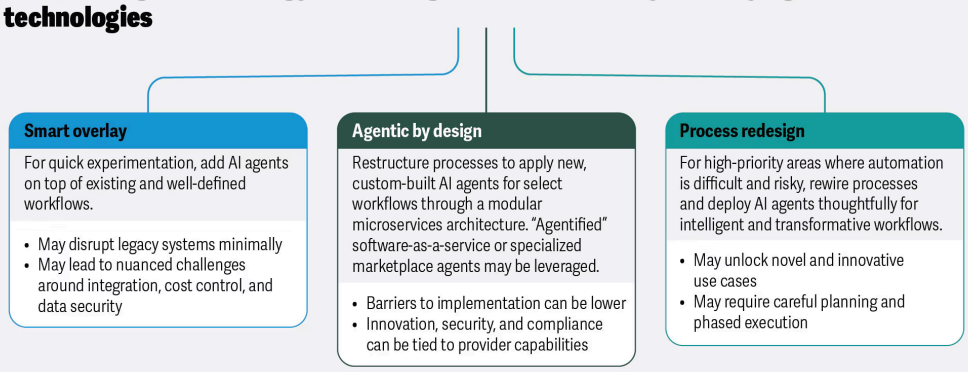
Enterprises today could leverage single-purpose AI agents to carry out multiple steps autonomously.[4] Increasingly, they're realizing that the benefits of agentic AI also extend to multiagent systems, unlocking broader and exponential enterprise value.[5] However, tech implementations could be far from maturity for many organizations.

In Deloitte's 2025 Tech Value Survey of nearly 550 US cross-industry leaders, 80% of respondents believe their organization has mature capabilities with basic automation efforts, whereas only 28% believe the same with basic automation and AI agent–related efforts. Furthermore, among those pursuing each strategy, 45% expect that their basic automation efforts could yield the desired return on investment within three years, whereas only 12% expect the same for basic automation and agents, within a similar time frame.[6]

How can they get there faster? Step one is to consider the three potential multiagent approaches (figure 2).[7]

Figure 2

An example of agentic strategy, depending on task complexity, underlying workflows, and technologies



Source: Deloitte analysis.

Deloitte. Insights | deloitteinsights.com

The human layer in agent orchestration

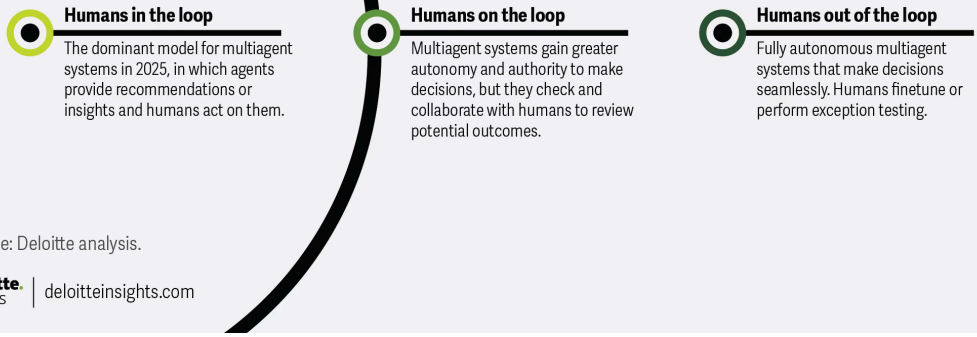
In 2025, businesses have been implementing relatively simple yet promising agent orchestrations in specific domains, like financial investment research and health care for critical illnesses.[8] In such applications, agents often work together under the purview of human supervision or a dedicated "supervisor agent" to provide insights for human professionals to act on. More complex and autonomous agent orchestration spanning across multiple business domains has been limited, for the most part, to select industry leaders.[9] As such efforts intensify, businesses will increasingly need to balance agentic autonomy and human oversight—carefully weighing innovation against risk, accountability, and trust.

Research suggests that today's emerging multiagent systems can perform better with humans in the loop, as they benefit from human experience and remain aligned with the nuanced organizational expectations.[10] We predict that, in the next 12 to 18 months, more businesses will accelerate experimenting and scaling of complex agent orchestrations, keeping humans in the loop. They will likely adopt frameworks and solutions to integrate human judgment into agentic workflows for higher confidence, quality, and accountability.[11]

Additionally, a progressive "autonomy spectrum"—humans in the loop, on the loop, and out of the loop—will emerge based on task complexity, business domain, workflow design, and outcome criticality (figure 3). While the humans out of the loop approach will still need continuous monitoring—human-in-the-loop and human-on-the-loop approaches will rely more on platforms and agent telemetry dashboards offering outcome tracing, orchestration visualization, and other details to guide human interventions. We predict, in 2026, the most advanced businesses will begin to lay the foundation of shifting toward human-on-the-loop orchestration.

Figure 3

The progressive evolution of AI agent autonomy over time



Source: Deloitte analysis.

Deloitte. Insights | deloitteinsights.com

Taming the fragmented AI agent proliferation

In 2026, AI agent sprawl is likely to increase across different programming languages, frameworks, infrastructure, and communication protocols. To add complexity, some agents might need multimodal capabilities (the ability to interpret different information types and formats like text, audio, and images) to reach peak intelligence. Additionally, web protocol developments for agents, like Massachusetts Institute of Technology's project NANDA, can define how agents coordinate on digital interfaces, external to businesses.[12] In the longer term, it can enable strategic agent orchestration across internal and external networks of businesses, unlocking new capabilities.

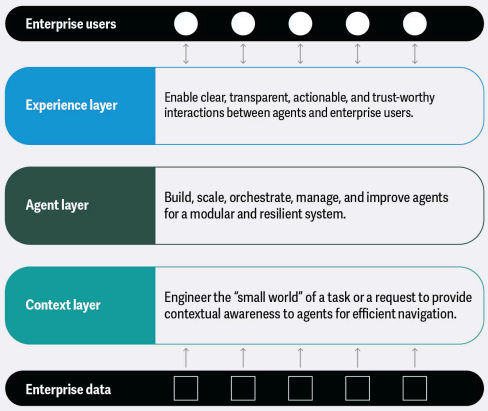
These variables will make multiagent interoperability critical yet challenging. Additionally, businesses will increasingly look for ways to direct, observe, and manage disparate AI agents through a unified platform. Lack of digital workforce operational standards may make building, configuring, and deploying AI agents decentralized and uncoordinated. This, in turn, will likely increase potential risks and costs of performance degradation and ethical, cyber, and regulatory compliance issues.

Businesses can draw inspiration from previous technologies that shaped today's information technology and business architecture, like cloud and microservices. Standardized protocols (like HTTPS, JSON, etc.), clear application programming interface blueprints, and domain-specific microservices enabled interoperability, stability, and ownership. Service registries, distributed tracking, and centralized logs improved discovery of capabilities, error resolution, and service management. Governance, service catalogs, and "zero-trust" security ensured robust systems and prevented confusion about versions. All these measures could offer lessons for building resilient and scalable multiagent systems. However, businesses should also adopt a fresh approach and focus on creating unique layers in their enterprise architecture.

Enterprise architecture for resilient and scalable multiagent systems

Figure 4

Enterprise architecture to build resilient and scalable multiagent systems



Source: Deloitte analysis.

Deloitte. Insights | deloitteinsights.com

1. **Context layer:** This robust knowledge engineering foundation is important for scalable AI agent architecture. It translates raw and diverse data into structured and well-governed knowledge representations (for example, knowledge graphs, ontologies, domain taxonomies, etc.) to provide agents with a "small world" model of the problem space. Optimized context retrieval techniques can empower agents with precise and timely access to relevant information, while context shaping can refine inputs to reduce noise and conflicts, enhancing agent accuracy and efficiency.

2. **Agent layer:** This component leverages the underlying context layer to enable agent operations, focusing on safety, autonomy, and interoperability. Central to this layer is a modular and composable architecture that can integrate and adapt to new technologies. Strategies emphasizing tool relevance and abstraction help prevent agent overload. Additionally, thoughtful memory strategies optimize access to the right blend of factual, experiential, and procedural memories to enhance context awareness. This layer also selects appropriate AI models (ranging from compact, specialized models to expansive, powerful ones) to optimize agent performance across orchestration tasks. Robust security measures and comprehensive observability via advanced telemetry help ensure secure, transparent, and reliable agent activities.

3. **Experience layer**: This primary interface between enterprise users and agents helps to control and course-correct agent actions. It provides users with relevant information like agent status and contextual data. It also enables prompt suggestions and comprehensible results in easy-to-review formats. Intuitive controls for human oversight, advanced feedback capabilities, and explainability features like displaying agent reasoning help make the outcomes more transparent and trustworthy. Additionally, when errors or ambiguous situations arise, it provides clear explanations and options to recover.

Making multiagent systems work for businesses

As businesses master the technical foundations, these three guideposts can help enable better alignment with business imperatives.

Flexible, scalable, and secure communication protocols

Multiagent orchestration requires a standard form of communication among agents and between agents and other tools or platforms. It's essential for predictable messaging on agent capabilities, insights, and actions. Over the last year, several inter-agent communication protocols have emerged, each promising coordination among agents built on different frameworks or models. These include Google's A2A, Cisco-led AGNTCY, Anthropic's MCP, and others.[13] Tech providers are rallying their partners, alliances, and customers to achieve dominance in this category. Additionally, some of these protocols are being extended for trustworthy agent interoperability in specific domains like financial transactions.[14]

Excessive competition across protocols could risk the development of "walled gardens," where companies are locked into one communication protocol and agent ecosystem.[15] It's likely, however, that, by next year, these protocols will begin converging, resulting in two or three leading standards that other tech providers will need to align with to remain competitive.

Which select protocols rise to the top will likely depend on multiple parameters and how businesses prioritize them according to their multiagent use, industry, and orchestration maturity. For example, lightweight protocols with standard application programming interfaces and developer tools for testing and simulation can ease experimentation. Support for peer-to-peer and hub-and-spoke agent interactions with shared context and memory and built-in negotiation, delegation, and conflict resolution can enable diverse orchestrations. Agent registries for trusted discovery and workload balance, asynchronous messaging, high throughput, low latency, and support for chained and nested workflows can help scale up agent orchestrations. Additionally, authentication, secure messaging, and access control can help mitigate security risks, while inter-agent messages and explanations can ensure auditability and error traceability.

Management platforms and observability tools

As multiagent systems scale, businesses will increasingly need to manage agents and understand the decisions being taken by them. They can leverage the unified and scalable platforms available, with supervising capabilities or "supervisor agents"—to interpret requests, route tasks, grant and manage access, and execute parallel or multi-step processes.[16] It's likely that, in the next year, tech companies will launch new capabilities here, leaving businesses to decide how they want such orchestration platforms set up. For example, central in-house platforms can limit vendor dependency and increase data and agent control. However, off-the-shelf platforms can help accelerate testing and manage the cost of innovation.

Whatever businesses choose, agent orchestration platforms will be important to track operational metrics, enhance performance, and manage cost. Currently, some platforms are developing ways to integrate monitoring of agent telemetry such as latency, error rates, token usage, and other tool insights.[17] Guardrail assessments and capabilities to detect unusual behaviors can help mitigate risks. Over time, such platforms will likely bring innovative features, such as layered business insights and additional control mechanisms. For example, an emerging category called guardian agent can both own tasks and govern other agents to sense and manage risky behaviors.[18]

Agent orchestration platforms will also need to incorporate regulatory compliance, an area where international efforts are advancing. The European Union AI Act sets requirements around risk assessment, transparency measures, technical safeguards, and human oversight.[19] In addition, the EU's standards bodies are working to develop harmonized legal standards as per the EU AI Act.[20]

Business process and workforce changes

Gartner® predicts that, by 2028, "33% of enterprise software applications will include agentic AI, up from less than 1% in 2024, with at least 15% of day-to-day work decisions being made autonomously through AI agents."[21] To get there, more businesses will likely begin reimagining their workflows in 2026,

defining concrete and unique modules. This will help determine the kinds of agent orchestration needed, depending on criticality, dependencies, task predictability, and targeted resilience. For example, some modules may benefit from agents working sequentially—where one agent's output becomes another's input—while other modules might leverage agents operating in parallel or collaboratively.

Another major consideration is how humans will collaborate with multiagent systems. A global survey of 200 human resources leaders found that 86% of chief human resources officers see integrating digital labor (that is, technologies performing intelligent work) as central to their role.[22] Early models show humans acting as "agent bosses," or working alongside agents.[23] In 2026, businesses will likely delve deeper into these collaboration models across more roles, functions, and tasks to identify where agent orchestration can enhance efficiency and where human strengths and collaboration can bring more meaningful value.[24]

By next year, enterprises will also likely start reimagining how existing roles can unlock higher-value outcomes with multiagent systems.[25] For example, human contributions can include more creative prompting and guiding multiagent systems while solving problems and taking strategic decisions efficiently. At the same time, businesses will also likely focus on defining the new human skills and responsibilities for agent training, orchestration, oversight, and governance.[26] Tailored training programs and developing leaders to manage both human and digital workers will be important—to embed higher quality, accountability, and resilience in multiagent decisions while leveraging uniquely human skills.[27]

The bottom line: 2026 could be an inflection point for agent orchestration

Agent orchestration will likely shape the next era of intelligent enterprises. Next year, we expect businesses to start scaling multiagent systems, bringing additional complexity to their IT and business environments. Agent communication protocols will likely consolidate around those offering ease of experimentation, flexibility, scalability, and security. Enterprise workflows will likely start becoming more modular, powered by agents—built internally or acquired through software as a service and other third-party providers. New and modified roles for human workers will begin emerging, facilitating effective collaboration with multiagent systems.

However, businesses and technology providers should act decisively to shape that journey.

Considerations for businesses adopting multiagent systems

- ***Define ownership and accountability.*** Businesses should identify who in the C-suite will own their company's AI agent vision, strategy, and execution with aligned incentives and accountability. This role could most naturally align with those leading strategic technology initiatives and driving innovation, but an integrated function can demonstrate more holistic impact and risk management.
- ***Design for evolution, not just deployment.*** Agents and orchestration capabilities are advancing fast. Modular "plug-and-play" orchestration frameworks can help businesses boost flexibility, cost-efficiency, and innovation, while minimizing disruption to system architectures.
- ***Stress-test orchestrations rigorously.*** Before scaling, businesses should simulate agent orchestration with real complexities of businesses—incomplete data, conflicting goals, or adversarial scenarios. Controlled environments can reveal hidden failure points and strengthen safeguards before enterprisewide deployments.
- ***Take governance and measurement seriously.*** AI agent governance will be critical to help ensure secure, compliant, and reliable orchestration on a scale. Setting clear rules for AI agent roles, defining their accountability, designing fallback routes to address errors, and oversight can help prevent misuse, ensure auditability, and build trust. Beyond technical readiness, enterprises should

identify and track metrics that connect agent orchestration to value creation—such as quicker decisions, better customer experience, or faster innovation.

Considerations for tech providers

- ***Build with interoperability***. Besides adhering to inter-agent communication standards, tech providers should design solutions that are modular, and where agents understand each other's intent and context of actions, to enable seamless coordination.
- ***Rethink trust.*** Insight delivery won't be enough; the ability to understand or validate AI agent output is essential for trust and adoption. Novel security measures like digital identity for agents will also be pertinent to build and run trustworthy multiagent systems.
- ***Make governance inherent***. Learning what businesses will need over time, to align with human values and organizational policies, could be key to providing relevant governance frameworks. Future solutions should have innovative agent monitoring and advanced governance, and ethical guardrails to enable compliance and efficacy.
- ***Expand the ecosystem.*** Tech providers should continue forming and strengthening industrywide alliances to achieve necessary standards in communication protocols, trust, and governance. Innovative and cross-platform orchestration tools are gaining traction, signaling opportunities for new and established tech players to strengthen their market position through acquisitions, partnerships, and collaboration.[28]

BY

Sayantani Mazumder
India

China Widener
United States

Gillian Crossan
Global

Girija Krishnamurthy
Global

Baris Sarer
United States

Diana Kearns-Manolatos
United States

ENDNOTES

1. Deloitte, "The cognitive leap: How to reimagine work with AI agents," December 2024.
2. The baseline projection is derived from a Deloitte analysis of global autonomous AI agent market projections as per seven publicly available and third-party research reports. The estimated increase of 15% to 30% in the projected market is modeled on future scenarios where fewer agentic AI projects are cancelled owing to improved enterprise readiness.
3. Gartner, "Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027," press release, June 25, 2025.
4. Bojan Ciric and Prakul Sharma, "Generative AI meets the virtual world: A model for human-AI collaboration," *Deloitte Insights*, Feb. 10, 2025.
5. Abdi Goodarzi and Nitin Mittal, "A new digitally-enabled workforce era: How AI agents can help deliver functional efficiency and value across the enterprise," *Forbes*, Aug. 18, 2025.
6. Tim Smith, Gregory Dost, Garima Dhasmana, Parth Patwari, Diana Kearns-Manolatos, and Iram Parveen, "Digital budgets are rising, but investment strategies may need a recalibration," *Deloitte Insights*, Oct. 16, 2025. The survey asked respondents about four types of AI automation and their incremental actions across each: mature or very mature respondents for basic automation (n = 443) and basic automation and AI agents (n = 153); and those with up to three-year expectations for basic AI automation (n = 245) and basic automation and AI agents (n = 68).
7. Prakul Sharma, Val Srinivas, and Abhinav Chauhan, "How banks can supercharge intelligent automation with agentic AI," *Deloitte Insights*, Aug. 14, 2025; Kausik Chaudhuri, "Applying agentic AI to legacy systems? Prepare for these 4 challenges," CIO, July 16, 2025; SaaS meets AI agents: Transforming budgets, customer experience, and workforce dynamics; Bojan Ciric and Prakul Sharma, "Scaling AI agents may be risky without an enterprise marketplace," *Deloitte Insights*, Sept. 15, 2025.
8. Julian Horsey, "AI investment research agent "Ask David" built by JP Morgan," Geeky Gadgets, May 30, 2025; Irene Iglesias Álvarez, "The agentic AI assist Stanford University cancer care staff needed," CIO, May 30, 2025.
9. Isabelle Bousquette, "Why Walmart is overhauling its approach to AI agents," *The Wall Street Journal*, July 24, 2025.
10. Henry Peng Zou et. al, "A call for collaborative intelligence: Why human-agent systems should precede AI autonomy," *arxiv*, June 11, 2025.
11. Jesus Olivera, "Ensuring accuracy in AI with human-in-the-loop," *Medium*, Sept. 27, 2024.
12. John Werner, "They're making TCP/IP for AI, and it's called NANDA," *Forbes*, May 01, 2025
13. Emilia David, "Google's Agent2Agent interoperability protocol aims to standardize agentic communication," VentureBeat, April 9, 2025.
14. Emilia David, "Google's new agent Payments Protocol (AP2) allows AI agents to complete purchases — is your enterprise ready?" VentureBeat, Sept. 16, 2025.
15. Leslie Joseph and Rowan Curran, "Interoperability is key to unlocking agentic AI's future," Forrester, March 25, 2025.
16. Alfred Shen and Anya Derbakova, "Design multi-agent orchestration with reasoning using Amazon Bedrock and open source frameworks," Amazon Web Services, Dec. 19, 2024; IBM, "Multiagent orchestration, accessed Oct. 7, 2025.
17. Amazon Web Services, "Observe your agent applications on Amazon Bedrock AgentCore Observability," accessed Oct. 13, 2025.

18. Gartner, "Gartner predicts that guardian agents will capture 10-15% of the agentic AI market by 2030," press release, June 11, 2025.

19. The Future Society, "How AI agents are governed under the EU AI Act," June 4, 2025.

20. CEN-CENELEC, "Artificial intelligence," accessed Oct. 7, 2025.

21. Daniel Sun, "Capitalize on the AI agent opportunity," Gartner, Feb. 27, 2025.

GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

22. Salesforce, "HR leaders to redeploy a quarter of their workforce as agentic AI adoption expected to grow 327% by 2027," May 5, 2025.

23. Ibid; Atikah Amalia, "The marketer's new job title: AI boss," Content Grip, April 29, 2025.

24. Kyle Forrest, Brad Kreit, Abha Kulkarni, Roxana Corduneanu, and Sue Cantrell, "AI, demographic shifts, and agility: Preparing for the next workforce evolution," *Deloitte Insights*, Aug. 25, 2025.

25. Michael Caplan et al., "The technology operating model of the future: Rise of the agentic enterprise," *The Wall Street Journal*, Aug. 23, 2025.

26. Ritu Jyoti, "The rise of the agentic economy: How autonomous AI is reshaping the future of work," CIO, Sept. 8, 2025.

27. Isabelle Bousquette, "Digital workers have arrived in banking," *The Wall Street Journal*, June 30, 2025.

28. Marina Temkin, "Why AI agent startup /dev/agents commanded a massive $56M seed round at a $500M valuation," TechCrunch, Nov. 28, 2024; Hui Wong, "Questflow secures $6.5M seed round to build AI agent economy for every workflow," Marketers Media, July 24, 2025.

ACKNOWLEDGMENTS

The authors would like to thank **Prakul Sharma**, **Rajib Deb**, **Mark Szarka**, **David Jarvis**, **Abhinav Chauhan**, **Michael Steinhart**, **Ankit Dhameja**, and **Iram Parveen** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

AI for industrial robotics, humanoid robots, and drones

Can more powerful AI models and chips catalyze what has been a relatively stagnant industry?

ARTICLE • 9-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

A factory floor bustling with humanoid robots that can see and act akin to human intelligence is a compelling vision for 2030 or 2040, and may even be possible. But the reality in 2026 is different. Deloitte predicts that cumulative installed capacity of industrial robots will surpass 5 million units in 2025 and could reach 5.5 million by 2026, globally.[1]

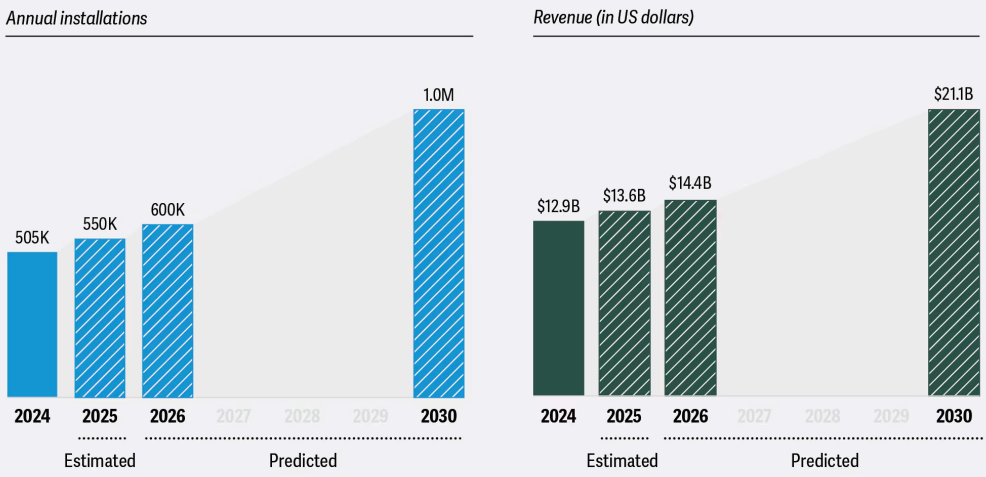
With greater integration of AI capabilities in robotic systems and the emergence of specialized foundational models, robots can permeate multiple industries and applications from smart factories to public utility services and even autonomous drones. But unless the broader technology, AI, and robotics ecosystem address bottlenecks related to data quality, integration, and cyber security, the market for industrial robots is likely to stay at its current level of relatively modest annual growth.

Advanced and special types of AI models as catalysts for industrial robots

Annual sales of new industrial robots have remained flat at roughly 500,000 units since 2021, in line with what Deloitte predicted in the *2020 TMT Predictions* about industrial robots' slow pace of growth.[2] Longer term projections suggest massive growth in the far future, with one estimate pegging the humanoid robotic industry at US$5 trillion by 2050.[3] Nonetheless, even as early as 2030, we could see an inflection point with annual new robot shipments doubling from current levels to reach 1 million a year, with projected revenues of US$21 billion in 2030, almost twice 2024 levels (see figure 1).[4]

Figure 1

Annual shipments of industrial robots, many AI-powered, could reach one million units by 2030, generating over US$20 billion in annual revenue



Source: Deloitte analysis based on publicly available information from sources including the International Federation of Robotics, Interact Analysis, and IPF Online.

Deloitte. Insights | deloitteinsights.com

What "robots" are covered in this article?

"Robots" is a broad term, ranging from dishwashers (yes, really), to more intelligent and autonomous home vacuum cleaners costing a few hundred dollars, to industrial robots on assembly lines worth millions of dollars each. And the definition sometimes includes flying robots (drones), driving robots (full self-driving cars), and humanoid robots that can do pretty much anything a human being can do, and more.

In this prediction chapter, the focus is primarily on industrial robots, humanoid robots meant for industrial use, and drones. There appears to be a rise in physical AI, robotics, and drones, and there is already a lot of articles and analyst coverage on autonomous vehicles. Therefore, this chapter will focus on industrial robotics and drones.

Two growth catalysts may create a turning point for industrial robots' increased adoption between 2026 and 2030. First, developed countries face persistent labor shortages due to ageing populations.[5] As these regions increasingly bolster domestic manufacturing and build resilient supply chains, demand for robots capable of handling increasingly sophisticated tasks will likely only go up. Second, and perhaps more importantly, exponential advancements in computing power and the emergence of specialized foundational AI models—different from typical large language models—are accelerating the development of AI robots and embodied AI systems.[6] Special-purpose models may be paving the way for highly sophisticated AI engines that can allow robots to move beyond simple command-and-control to comprehending natural language, perceiving physical surroundings, and learning and navigating complex tasks in a generalized way just like humans do.[7]

Despite the enthusiasm and emergence of advanced technologies, certain hurdles to robotics advancement remain. For instance, integration of robotic systems into existing industrial workflows is complex, particularly concerning data quality, interoperability, and legacy system compatibility. Many companies struggle to harness clean, unified datasets (e.g., real-world data, physical surroundings, spatial data), which are essential to train the robots.[8] Moreover, the prospect of security and privacy breaches or malicious cyberattacks on connected robotic networks remains a critical concern.[9] Additionally, the safety of human workers is an essential aspect that industrial robots and humanoid robots need to address.[10]

Deloitte believes that a tighter integration of gen AI and agentic AI with robotics and automation tools would help bring AI-enabled robotic devices out of the realm of science fiction and into modernized workplaces.[11] As a case in point, a smart factory in Wichita, KS, used to simulate cutting-edge, real-world use cases, houses diverse tech capabilities including gen AI, agentic AI, unlimited reality, as well as robotics such as drones, autonomous mobile robots, quadrupeds, and humanoid robots.[12]

Industrial robots already appear to be unlocking value for multiple industries such as manufacturing, health care, warehouse, and even national defense (figure 2).[13] But what's likely shaping new opportunities for industrial robots appears to be the innovation that some technology companies have been demonstrating, especially with an advent of multimodal AI models, as well as advanced chips and hardware.

Figure 2

Powered by AI, industrial robots are generating value for multiple industries

Industry environment	Benefits and use cases of AI-powered robots
Manufacturing	Robots, including cognitive humanoids with bionic hands, can use synthetic data to self-learn and work in a coordinated fashion in high-end factories. Equipped with 3D object recognition, such robots can assist with machine loading, injection molding, and maintenance.
Health care	Robots can assist nurses to run errands (like picking samples and delivering meds), help surgeons perform delicate procedures that require high levels of precision, support personnel to handle hazardous materials (such as virus samples), and perform high-risk jobs like disinfecting rooms.
Warehouse and logistics	AI-powered robots can use deep-learning vision to recognize and handle a range of items of varying shapes and sizes. Autonomous mobile robots use AI algorithms and advanced sensor data generated from cameras and 3D vision to map and navigate the warehouse environment in real time, and station themselves accurately without relying on any fixed infrastructure or markers.
Defense and military	Robot dogs can spot and dispose of bombs and classify and identify objects of threat using advanced sensors and AI-based analytics. AI robots can manage surveillance and reconnaissance missions, carry supplies, and assist with casualty evacuations.

Source: Insights gathered from conversations with industry subject matter experts, as well as multiple publicly-available sources including: American Machinist, Admedica, World Economic Forum, PHS Innovate, and ASDNews.

Vision-language-action models are likely to make humanoid robots smarter and more autonomous

Some AI startups and major tech companies are developing vision-language-action (VLA) models that can make it possible for robots to advance from performing pre-programmed tasks to understanding context and making decisions autonomously. VLA enables robots to gain more autonomy, allowing them to develop higher order planning and spatial reasoning, and providing them with dexterity to navigate challenging terrains.[14] With large scale reinforcement learning in simulation and multimodal learning, robots can get pre-trained on vast datasets.

VLA integrates visual perception (observing the environment and the laws of physics), natural language understanding (verbal commands and comprehension), and real-world actions to perform (responding to visual and textual instructions).[15] Typically, as of mid-2025, VLAs were anywhere from 500-million to 7-billion parameter models, enabling humanoid robots to learn, perceive, and act.[16] There are select examples where VLA models are being used to augment robotics development in the United States, with the potential for wider commercial adoption between 2026 and 2030:

- NVIDIA's open foundational model for humanoid robots combines reasoning and actions to help advance robotics development.[17] Robotics companies like Boston Dynamics are building humanoid robots by using libraries from NVIDIA's model and other supporting technologies from NVIDIA.[18]

- Figure AI's Helix is a VLA model that trains robots using visual and natural language prompts, enabling humanoid robots to learn intimately about real-world scenes and objects and develop fine motion control.[19]
- Hugging Face developed open-source data and models specifically for robots, even as it continues to build and test its own open-source humanoid robot,[20] allowing developers to customize their own robots.[21]

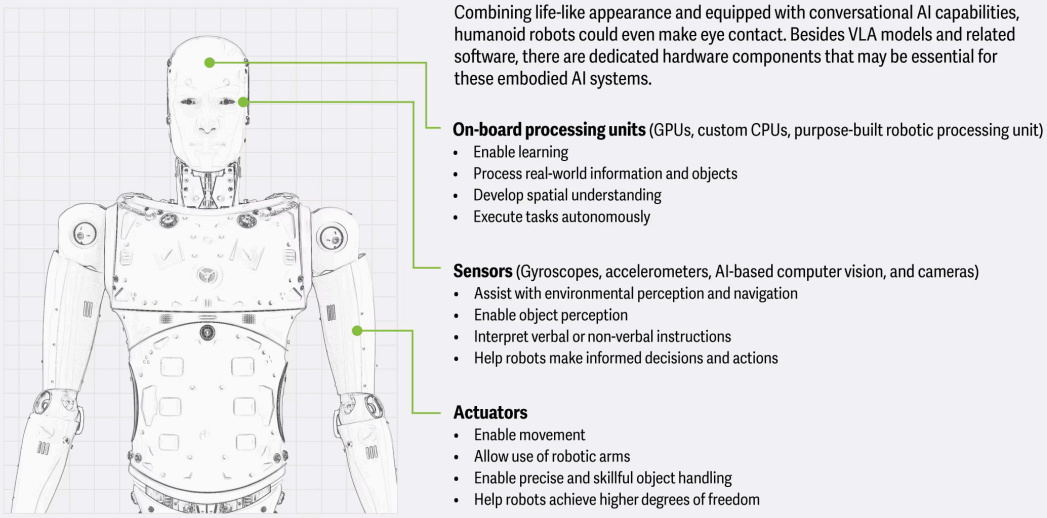
Outside the United States, humanoid robots are being developed in Asia and Europe as well, with emphasis on custom foundational models and training on physical world data. For instance, South Korea-based startup RLWRLD is developing foundational AI models that would allow traditional manual-intensive processes to be performed autonomously by robots through automated learning and mimicking human expertise.[22] In Japan, FANUC Corporation is focused on developing a range of AI-powered robots across various sizes, designed for industrial environments.[23] In Europe, Neural Foundry (London-based) and NEURA Robotics (Germany) are building AI robots for industrial environments by integrating cognitive capabilities and developing custom models.

In China, startups such as AgiBot and MagicLab are designing humanoid robots capable of handling complex tasks in manufacturing environments.[24] And the likes of Unitree Robotics and UBTECH Robotics are advancing toward mass production and making humanoid robots accessible and affordable to help drive wider adoption.[25] Various chip components and hardware go into building a humanoid robot (figure 3), indicating a strong revenue potential for semiconductors (across chip hardware and related software and services) from this market.[26]

While AI-powered humanoid robots meant for industrial use at scale may still be in early stages, Deloitte estimates annual unit shipments to be in the range of 5,000 to 7,000 in 2025, which may increase to 15,000 in 2026.[27] At an average price of US$14,000 to US$18,000 per unit,[28] the AI humanoid robot market for industrial use could be worth around US$210 million to 270 million in 2026.[29] As the robotics industry overcomes technology, price, and operational barriers between 2026 and 2030, the market for humanoid robots could reach US$600 million to US$700 million (roughly three times the market size of the 2026 baseline scenario) or even attain US$1 billion (four times the market size of the 2026 optimistic scenario) by 2032.[30]

Figure 3

The chips and hardware that can make humanoid robots and embodied AI possible



Note: Vision language action (VLA); graphics processing unit (GPU); central processing unit (CPU).

Source: Deloitte analysis.

Deloitte Insights | deloitteinsights.com

Advanced AI is likely to make drones more autonomous and versatile

Most drones, also known as unmanned aerial vehicles (UAVs), are currently manually operated, but their autonomous capabilities appear to be advancing rapidly. Many drones now use AI for real-time navigation, communicate with each other, avoid obstacles and collision, and may soon be able to execute missions without human intervention. As a case in point, scientists in Hungary studied patterns and movements of various animals including pigeons and wild horses. They used those insights to help build an algorithm that would guide a swarm of drones capable of making onboard, autonomous decisions. These drones can not only navigate, avoid collisions and hover safely in the skies, but perform missions in diverse environments including land surveying, meteorology, and wildfire management.[31]

Drones: What's inside?

New age, sophisticated, AI-enabled drones are often equipped with various types of tech and chips; single or dual microcontrollers serving as flight controllers; onboard power systems that can include lithium-based batteries and distribution boards to supply power to various components; radio-frequency modules to enable communication between drones and ground control units; GPS modules for navigation and positioning; sensors (such as accelerometers, gyroscopes, magnetometers, optical flow sensors, and lidar and ultrasonic); and onboard flight control software and platforms to manage aerial operations.[32]

As noted in Deloitte's 2024 prediction on agricultural technology, a combination of spectral sensors, chips, and cameras mounted on UAVs or drones gather large volumes of data (soil moisture, plant health, etc.) that AI models can analyze to offer insights for targeted spraying operations.[33] Besides agriculture, drones can be used to inspect wind turbines and electric power lines, minimizing the need for manual inspections.[34] China, South Australia, and the United Kingdom are experimenting with UAVs that carry out fully autonomous long-range, remote inspection of high-voltage power lines. They not only help human workers and engineers by taking up such dangerous and critical tasks but can also auto-capture and transmit dozens of images that would help engineers to detect and analyze corrosion through AI and advanced analytics.[35]

Several countries are aiming to deploy autonomous drones for aerial surveillance to assist with disaster relief (for example, autonomous drones mapped damaged areas following Southwest Florida's Hurricane Ian during September 2022, assisting emergency responders), as well as to detect and counter potential border threats.[36] In many of these applications, the drones are only remotely operated by humans for part of the mission: The AI acts like an autopilot on a commercial jet and handles the relatively simpler task of getting the drones close to their destinations, before handing off to the human operator. But the recent efforts by several countries in drone swarms for military applications [37] indicate how they could possibly influence nonmilitary (industrial and civilian) applications as well. For instance, a swarm of autonomous drones could inspect high-tension power lines in remote and difficult-to-access terrains and even monitor offshore wind turbines in harsh weather conditions.

The bottom line: Commercialization, safety, and workforce readiness

Industrial robots are already an important end market for semiconductor companies, despite the industry's relatively modest growth in recent years. For example, an industrial robot worth roughly US$ 200,000 could contain approximately US$25,000 to US$50,000 worth of chips and related electronics components.[38] Further, making industrial robots better will likely rely on increasingly advanced chips, ranging from processors to networking to sensors, and the semiconductor content per robot will likely

increase. Additionally, the semiconductor industry is a significant consumer and an end-user of industrial robots as of 2025, using them in various aspects: fab manufacturing processes, wafer handling, testing, and sorting, advanced packaging, and clean rooms.[39] In the journey toward "lights out" manufacturing, the chip industry will potentially use even more industrial robots as part of its operations.

As market opportunities for industrial AI-powered robots including humanoid robots and drones appear promising, many semiconductor and technology companies are actively investing in this area for the long term. Robotics startups are in pilot stages in real-world contexts like warehouses, logistics, and aerial autonomy. Venture capital investments in robotics are growing, which is expected to be the only non-AI market category that may experience an increase in funding during 2025.[40] Cloud and IT infrastructure is also falling in place, even as synthetic data generation and physics simulators may be accelerating development and lowering reliance on high-cost real-world trials.

Here are five action steps that AI, robotics, and tech industry leaders can consider taking to help address some of the potential challenges related to industrial robotics commercial adoption, as well as to help address matters related to data integration, privacy and cyber, safety, and workforce readiness.

1. **Demonstrate commercial viability through open innovation:** Tech and AI companies should demonstrate ROI via broader commercialization by promoting open, full-stack robotics ecosystems that allow for the wide-ranging deployment and coordination of robots; and create a collaborative general ecosystem to move toward general-purpose embodied AI.[41]
2. **Enhance data quality and address data integration:** Ecosystem players should prioritize data standardization and collaboration for common platforms and middleware for a more seamless integration of diverse types of robots into industrial environments.
3. **Fix cyber vulnerabilities:** Companies should embrace common interoperability protocols, adopt privacy and security-by-design approaches, and proactively engage cyber specialists to craft clear and flexible security frameworks.
4. **Address safety as an essential and integral feature:** Right from development and early-testing and prototyping phase, robots should be programmed for safety, whether it's about working alongside humans safely without causing physical injury, or in ensuring they don't collide with each other accidentally. Simulation-based training, computer-aided safety planning tools, and proactive collision-free motion planning are novel technologies and approaches that can help make robots safer.
5. **Augment current workforce proactively:** Reskilling and upskilling the workforce on emerging AI tech can be critical for every single company. As robots are increasingly working alongside humans in this next wave of industrial AI automation, companies should assess and level up their workforce's AI skills on a more regular basis to help stay at the forefront of industrial robot adoption and integration into their broader enterprise fabric.

The way forward appears quite clear: AI, robotics, and technology industries should take the starting steps as there's both the necessary advanced AI tech and the commercial appetite and interest. A complete 360-degree systems thinking and an ecosystem-based approach may be essential to demonstrate progress across the five areas presented above—related to open innovation, data, cyber, safety, and talent—and accelerate commercial adoption of industrial robotics in 2026 and beyond.

BY

Karthik Ramachandran
India

Duncan Stewart
Canada

Jeroen Kusters
United States

Tim Gaus
United States

Gillian Crossan
Global

Girija Krishnamurthy
Global

ENDNOTES

1. Deloitte analysis and estimates based on data from publicly available information sourced from the International Federation of Robotics, Interact Analysis, and Automation.com.

2. Duncan Stewart et al, "Robots on the move: Professional service robots set for double-digit growth," *TMT Predictions 2020*, November 2019. To read further, see "Professional services robots on the move," *The Wall Street Journal-CIO Journal*, April 8, 2020.

3. Morgan Stanley research, "**Humanoids: A $5 trillion market**," May 14, 2025.

4. Methodology and assumptions: From 500,000 annual installations each year, in 2025 and 2026, we anticipate annual industrial robot installations could grow by 100,000 units every year between 2027 and 2030, reaching 1 million installed units in 2030. These calculations are based on insights gathered from IFR press release dated September 24, 2024 ("**Record of 4 million robots in factories worldwide**"). From our conversations with industry experts, we believe growth and availability of computing power, especially new types of AI models (LLMs, but also VLAs and world models), plus the active role that some major tech and robotics companies are playing to invest and bring forth robotics chips and solutions to market, will help drive robotics adoption during 2026 to 2030 and beyond. Additionally, average unit price per industrial robot has declined by approximately 3.2% between 2018 and 2024. We expect average price to continue to decline in that range through 2030, given the broader availability of chips, sensors, and other components, including open model-based robots. Between 2025 to 2030, we have assumed average annual price per industrial robot could decline approximately 3.1 to 3.2 percent based on information gathered from IPF Online's article dated June 27, 2025 ("**Global industrial robot shipments down in 2024, recovery likely in 2025**").

5. OECD, **OECD Employment Outlook 2025: Can we get through the demographic crunch?**, July 9, 2025.

6. Deloitte analysis of the various foundational models released by technology companies and niche LLM players during 2024 and 2025.

7. Standard bots, "**The most advanced robots in 2025**," August 7, 2025.

8. Cem Dilmegani, "**Data quality in AI: Challenges, importance, & best practices**," AIMultiple research, July 9, 2025.

9. Ainsley Lawrence, "**AI's impact on robots in manufacturing**," September 11, 2024.

10. Brian Heater, "**Figure AI details plan to improve humanoid robot safety in the workplace**," *TechCrunch*, January 28, 2025.

11. Tammy Whitehouse, "**AI robots in the workplace: Preparing for humanoid colleagues,**" Deloitte-*WSJ CIO Journal*, July 26, 2025.

12. **The Smart Factory by Deloitte website**, "Home page," accessed Oct 29, 2025.

13. Deloitte analysis based on insights gathered from interviews and conversations with industry subject matter experts, and supplemented with information gathered from multiple publicly available sources including: ***American Machinist***, **Admedica**, **World Economic Forum**, **PHS Innovate**, and ***ASDNews***.

14. Reyk Knuhtsen, et al, "**Robotics levels of autonomy,**" *SemiAnalysis*, July 30, 2025.

15. Sudhir Pratap Yadav, "**Vision-Language-Action (VLA) models: LLMs for robots**," Black Coffee Robotics, April 17, 2025; Raman Thakur, "**How Vision-Language-Action models powering humanoid robots**," Labellerr, March 5, 2025.

16. Deloitte analysis based on information gathered about multiple VLA models that are commercially available in the market.

17. Andrew Liszewsk, "NVIDIA says 'the age of generalist robotics is here'," *The Verge*, March 19, 2025.

18. Automation World, "Boston Dynamics working with NVIDIA on next-gen humanoid robots," May 21, 2025.

19. Brian Heater, "Figure's humanoid robot takes voice orders to help around the house," *TechCrunch*, February 20, 2025; Wei Sun, "Figure AI Unveils its 2nd-Gen Robot, Extending Focus from Factory to Home After OpenAI Split," Counterpoint Research, August 14, 2025.

20. Rebecca Szkutak, "Hugging Face unveils two new humanoid robots," *TechCrunch*, May 29, 2025.

21. Michael Nunez, "Hugging Face just launched a $299 robot that could disrupt the entire robotics industry," *VentureBeat*, July 9, 2025. The company launched a sub US$ 300 robot, which can integrate with the Hugging Face Hub, enabling its developer community to access pre-built AI models, hardware designs, and software and assembly instructions.

22. Kate Park, "RLWRLD raises $14.8M to build a foundational model for robotics," *TechCrunch*, April 14, 2025.

23. The Robot Report, "RBR50 Spotlight: FANUC produces one-millionth industrial robot," August 12, 2024.

24. domainB, "China's AI-powered humanoid robots set sights on transforming global manufacturing," May 13, 2025.

25. Based on publicly available secondary sources that reference Unitree and UBTECH.

26. Based on multiple publicly available data and research reports that highlight the various chip components and hardware that are used to build humanoid robots.

27. Deloitte analysis based on data and information gathered from select major AI humanoid robot makers in the US and China.

28. Deloitte analysis based on data and information gathered from select major AI humanoid robot makers in the US and China.

29. Note to calculations: Using the 2026 estimated price range of US$14,000 to US$18,000 per unit, and 15,000-unit shipments, we multiplied the two variables to arrive at US$210 to US$270 million as overall revenue opportunity.

30. Using the variables and methodology noted in end note No. 26, we took the baseline scenario range of US$210-270 million for 2026 and multiplied it by 3X and 4X to arrive at the other two probable 2032 market revenue potential presented in this paragraph. Our underlying assumptions for these relatively optimistic scenarios are mainly based on how fast the broader AI, robotics and tech industry might be able to address and workaround data, integration, safety, and cyber related challenges, and as price points become relatively attractive over time.

31. Justin Spike, "Data on animal movements help Hungarian researchers create a swarm of autonomous drones," AP News, December 19, 2024.

32. Deloitte analysis based on information gathered from publicly available sources about AI-enabled drones.

33. Karthik Ramachandran, Gillian Crossan, Duncan Stewart, and Ariane Bucaille, "On solid ground: AgTech is driving sustainable farming and is expected to harvest US$18 billion in 2024 revenues," *TMT Predictions 2024*, November 29, 2023.

34. Damon Johnson, "From Sci-Fi to reality: The latest in drone technology for 2024," Raising Drones, July 12, 2025.

35. Yahoo! Finance, "Britain to allow drones to inspect power lines, wind turbines," October 15, 2024; Joe Macy, "Autonomous UAS inspection system for power lines introduced," Unmanned

Systems Technology, March 14, 2025.

36. Damon Johnson, "From Sci-Fi to reality: The latest in drone technology for 2024," Raising Drones, July 12, 2025.

37. Aja Melville, "Drone Wars: Developments in Drone Swarm Technology," Forecast International, January 21, 2025.

38. Deloitte analysis based on publicly available price information of select major industrial robots in the market.

39. Gregory Haley, "Increasing roles for robotics in fabs," *Semiconductor Engineering*, Aug. 19, 2024.

40. Rebecca Szkutak, "We are entering a golden age of robotics startups — and not just because of AI," *TechCrunch*, September 12, 2025.

41. Deloitte China, "Open Full-stack Intelligent Service Robot Ecosystem white paper," April 24, 2025.

ACKNOWLEDGMENTS

The authors would like to thank **Dan Hamling**, **Rohini Prasad**, **Viswanath Anakkara**, **Joe Mariani**, and **Adam Routh** for their contributions to this article.

Cover art by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

SaaS meets AI agents: Transforming budgets, customer experience, and workforce dynamics

As AI agents pervade the SaaS market, how businesses experience and leverage software will likely change—shifting business models, capabilities, and expectations

ARTICLE • 10-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

As agentic AI capabilities mature and enterprise software-as-a-service (SaaS) vendors build out their platforms to create, integrate, and orchestrate AI agents, how organizations purchase and use software could shift dramatically. In 2026, SaaS applications will likely become more intelligent, personalized, adaptive, and autonomous, evolving towards a federation of real-time workflow services that can learn from their experiences. This evolution should disrupt traditional pricing models. Subscriptions and seat-based licensing could give way to hybrid approaches that blend usage- and outcome-based pricing. All these advancements will likely introduce new complexity in both software implementation and monetization—potentially redefining the entire SaaS business model.

What is an AI agent?

In artificial intelligence, an intelligent agent is an entity that perceives its environment, takes actions autonomously to achieve goals, and may improve its performance through machine learning or by acquiring knowledge.[1]

AI agents could drive a gradual transformation of SaaS markets starting in 2026

To put things in perspective, let's take a step back and look at how overall AI adoption appears to be evolving in the market. Deloitte's 2025 Tech Value survey found that 57% of respondents were putting between 21% and 50% of their annual digital transformation budgets into AI automation, and 20% of respondents were investing 50% or more (US$700 million on average for a company with $13 billion in revenue).[2] Nearly three-quarters of surveyed leaders said their organizations funded AI and generative AI technology capabilities over the last 12 months (the No. 1 area) and 39% funded agentic AI.

Based on this, Deloitte predicts that up to half of organizations will put more than 50% of their digital transformation budgets toward AI automation in 2026, and agentic AI will see an even higher percentage

of companies investing, perhaps reaching 75%. Although the Tech Value survey focused on US respondents only, we believe that companies around the world will follow a similar path, possibly delayed by a year or two. SaaS is often foundational to digital transformation efforts, and treating these broader spending shifts as a proxy, we expect commensurate increases in spending toward autonomous AI agents as part of SaaS in the next year.

Where could all this investment and technological advancement eventually lead? There are some optimistic visions of the future getting attention. Some have stated that parts of, or even entire, enterprise applications could eventually be replaced by agents.[3] Deloitte predicts that this future may ultimately come to pass for some enterprise applications, but it won't be in 2026. It will likely take at least five years or more to come to fruition, even with the rapid pace of technological development and investment around agentic AI. There are challenges to this vision, as traditional SaaS providers have large footprints across complex workflows that will likely be hard to supplant.[4]

In 2026, we will likely see a lot of experimentation, a general augmentation of capabilities, and a slow restructuring of the SaaS market, with AI-first companies competing. This moderate pace is likely because the "agentification" of SaaS is often not only about technological change, but business and operating model change as well—for both vendors and users.

With agentic AI, SaaS is about to get more complex

Many CIOs and CTOs continue to face pressure to reduce costs and streamline the number of vendors that they use.[5] In an agentic AI era, the question often arises, when and how should organizations start to shift their investments toward solutions with AI agents in the hopes of greater efficiency?

There are a couple of different paths some of the largest SaaS providers are taking in their approach to providing these capabilities to their customers. Many are adding AI agents to existing products and producing brand-new AI agent–powered products (Salesforce Agentforce, SAP Joule Agents, ServiceNow Now Assist AI Agents, and Workday Illuminate Agents are recent examples).[6] Many are also creating agent-building frameworks built on top of current services and introducing new data management and orchestration capabilities to help make the creation and management of AI agents easier (Google Cloud Agent Development Kit, Oracle AI Agent Studio, SAP Business AI, Workday Build and Adobe Experience Platform Agent Orchestrator are recent examples).[7]

In an agentic AI era, the question often arises, when and how should organizations start to shift their investments toward solutions with AI agents in the hopes of greater efficiency?

In addition, some new AI-native companies appear to be developing agentic solutions that could potentially disrupt these incumbents. In the short term, "easier" business processes like customer service are more likely to be disrupted, but disruption could spread to more complex markets like ERP (enterprise resource planning) and CRM (customer relationship management). Significant amounts of investment are powering many of these startups.[8] Many of these emerging companies are likely to get acquired in the next few years as incumbents look to expand their portfolio of agents and seek differentiation. In fact, Gartner® says, "By 2030, 35% of point-product SaaS tools will be replaced by AI agents or absorbed within larger agent ecosystems of major SaaS providers."[9]

Today, organizations have access to AI agents through their existing SaaS providers, which can make it easier to test and learn how to build agentic solutions through built-in functionality. While organizations may take this agentic-by-default approach initially, as they gain more experience, they will likely shift toward a more deliberate tack. Building around their data, Deloitte predicts they will pick capabilities

from a broad and complex agentic ecosystem, develop their own agents, and weave everything into an integrated and autonomous multi-agent system.

Navigating the transition to agentic AI

To more successfully get to this future, several challenges should be addressed:

Pricing becomes more complex

An area that will likely significantly impact both SaaS users and vendors alike will be how using AI agents will be priced and paid for. When software was mostly on-prem, you typically had a perpetual software license and paid for upgrades and upkeep. The SaaS revolution, driven by the cloud, shifted things to subscriptions. Today, there are a couple of common pricing approaches for SaaS. Generally, organizations are charged based on the number of users or seats that they have. These seats could include a tiered pricing option, where different tiers provide different sets of capabilities based on the type of user. Such pricing can be relatively straightforward and predictable. Usage or consumption-based pricing appears to also be increasingly common, and less predictable. This model is often based on the number of API calls or tokens (units of text or data an AI model processes) used.

As AI agents enter more widespread use, these traditional pricing models won't be adequate to reflect the true value exchange between provider and consumer.[10] AI agents could conceivably give one user the power of many users and reduce the need for the number of seats needed in an organization, impacting the revenue of SaaS providers. Additionally, AI agents operate autonomously, and their actions aren't necessarily predictable; they may take novel or inefficient paths while completing their tasks.

There will likely be a lot of effort needed to shift to these newer models, and we expect to see pricing variety and experimentation in 2026 and beyond. It could take years for standard practices to emerge, if they ever do. There are a couple of pricing models that are expected to gain in popularity: usage-based and outcome- or value-based. Gartner says that "by 2030, at least 40% of enterprise SaaS spend will shift toward usage-, agent-, or outcome-based pricing."[11]

Usage-based pricing

In usage-based models, a customer could be charged every time an agent takes an action or completes a task. Pricing could also be based on computing time, API calls, the number of tokens used for generative tasks, or how long an agent is in action (or a combination of all of those). There could also just be a flat fee per time period for the use of a single agent, like a salary for a digital worker. In a recent survey of SaaS companies, Maxio found that 83% of AI-native SaaS companies currently offer usage-based pricing.[12] Usage-based pricing is often attractive because it is quantifiable and therefore auditable.

Outcome- or value-based pricing

Pricing model changes will impact multiple functions within organizations and may transform how SaaS vendors operate.

Outcome- or value-based pricing is based on the real business results that SaaS applications with AI agents produce—something that can be much harder to measure. This could be as simple as the number of customer support tickets that get resolved or how many employees were eventually hired because of an HR agent, or it could be as complex as an increase in overall revenue AI agents contributed to. There's likely still a long way to go before there's widespread use of this model, though some are pursuing it.[13] Agentic systems still need to prove that they can produce consistent and reliable value.

These pricing model changes will impact multiple functions within organizations and may transform how SaaS vendors operate. First, there should be agreements around basic definitions for things like "an agent," "a task," "a process," "an interaction," and "an outcome." What "value" is and how it is attributed should be clearly defined, communicated, and agreed upon contractually. This will likely take significant effort and coordination from engineers, sales people, legal teams, and others. Proving that an AI agent created value or a business outcome could be challenging, especially if multi-agent systems composed of agents from different vendors are used. Revenue for vendors and costs for customers could become less predictable and highly variable. System instrumentation and metering may have to become more advanced and data observability, billing, and financial compliance may have to become more real time and autonomous.

The sales models for many vendors will likely need to change. Sellers will have to educate customers on these new models and convince them that AI agents will create value and the shift won't cost them more than their subscription-based services. Sellers will also likely have to be measured and compensated differently and may have to drive deeper relationships with customers.

Customer experience and user interfaces could become greater differentiators

AI agents are, by nature, supposed to be autonomous, so why do they need to have a user interface? Like APIs, agents are "headless." They don't have a direct connection to a user interface. However, someplace for interaction and visibility is necessary. So, what will that look like? Will there be a single, primary AI agent interface or multiple ones? Will a SaaS provider or third party "control" a gateway to agents?

Over the next few years, Deloitte predicts that the user experience and interface for SaaS AI agents will become more:

- *Personalized and proactive:* The interface will adapt to the individual user providing needed tools and tasks based on specific responsibilities and prior actions. It will provide tailored insights and suggest specific actions for users to take.
- *Conversational:* Interaction will move from menus and clicks to natural language and voice commands. AI agents will translate natural language into a structured series of API calls, eliminating the need for pre-defined workflows. It will be less about *telling* software what to do and more about *asking* software to achieve a particular outcome.
- *Diagnostic:* Because of the autonomous nature of AI agents, if something wrong or unexpected happens, users will have to be able to reconstruct the agent's decision-making process and understand why it happened. Transparency, explainability, reversibility, and auditability will be crucial for trust.

Another open question is where will the interaction layer be? Deloitte predicts a lot will be done in stand-alone SaaS apps. Many SaaS providers want to keep users in their application as much as they can to maintain worker efficiency and keep users using their products. They will increasingly provide access to not only their own suite of agents, but agents from other providers as well. Interaction could also take place through a separate management platform. These might be provided by a SaaS vendor or they could come from a third-party company (like current SaaS management platforms). These "control centers" could integrate agents' activities from multiple vendors and internally developed agents—tracking usage, expenditures, access, performance, status, security, and compliance.[14] There will also likely be agent marketplaces, where internal and external agents get published and businesses can discover and integrate new capabilities dynamically.[15] This interaction, or attention, layer has the potential to provide significant value, and there is likely to be considerable competition around it.

The bottom line

In 2026, the usage of AI agents through SaaS applications is set to rapidly grow, with many major SaaS providers working to implement more robust agentic AI solutions with their customers. We expect

increased investment in all AI-powered automation, extending into SaaS applications. Organizations will be seeking process efficiency, cost savings, greater flexibility and personalized capabilities for workers. There will be a lot of experimentation and pricing variety. Overall, Deloitte predicts there will be a gradual move toward a future powered by integrated, autonomous multi-agent systems.

Considerations for SaaS customers to help prepare:

- *Invest in data management:* For AI agents, access, integration, observability, and data governance can become even more important. Data doesn't necessarily need to be centralized in a single repository, but it should be consistent and accessible across an organization.
- *Embrace the growing complexity:* There will be more models, more agents, more vendors, new ecosystems, and new data relationships. Organizations will have to get agents from different vendors to work together, potentially causing pricing and operational complexity.
- *Expect multifaceted pricing models:* Pricing models for AI agents could create ambiguity as hybrid models that include a mix of licenses and usage-, value-, or outcome-based pricing become standard. Bolster your real-time finance capabilities.
- *Help workers become AI orchestrators:* More time could be spent managing AI agents like co-workers—setting goals, supervising their work, and validating and correcting their actions. When rearchitecting workflows, clearly define what humans will do, what agents will do, and what they will do together. It's a cultural shift, not just a software upgrade.

Considerations for SaaS vendors to help prepare:

- *Prepare for greater competition:* As it becomes easier and easier for anyone to write code through generative AI tools, the cost of producing code approaches zero. This is going to create greater competition with AI-native companies and even customers themselves—requiring greater product differentiation.
- *Focus on interoperability:* Agents will have to operate across multiple systems, coordinate tasks, and share data and goals—all while maintaining security and compliance. Organizations should prepare for a more open and interoperable environment, in which customers can easily switch providers if expectations aren't met.
- *Shift sales models:* With varied pricing models, things like revenue forecasting become more challenging. How sales teams are measured and compensated will have to evolve. Help customers predict costs and provide hybrid pricing models that are simple and flexible. Use conversations around pricing models to expose unmet needs and deepen relationships.

BY

David Jarvis
United States

Sayantani Mazumder
India

Girija Krishnamurthy
United States

Gopal Srinivasan
United States

China Widener
United States

Gillian Crossan
Global

ENDNOTES

1. Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (New York, NY: Pearson, 2021).
2. Tim Smith et al., "AI is capturing the digital dollar. What's left for the rest of the tech estate?," Deloitte, October 16, 2025; AI automation includes basic automation, process automation with agents, process reimagination, and organizational reimagination.
3. Eric Newmark, "The agentic evolution of enterprise applications," IDC, April 4, 2025.
4. Dan Gallagher, "Software's death by AI has been greatly exaggerated," *Wall Street Journal*, August 27, 2025.
5. Zylo, "111 unmissable SaaS statistics for 2025," accessed October 2025; Matt Ashare, "AI drives up compute costs as cloud inflation slows," *CIO Dive*, February 18, 2025.
6. Salesforce, "Agentforce," accessed October 2025; SAP, "Joule Agents," accessed October 2025; ServiceNow, "AI Agents," accessed October 2025; Workday, "Workday unveils next generation of Illuminate Agents to transform HR and finance operations," press release, May 19, 2025.
7. Oracle, "Oracle introduces AI Agent Studio," press release, March 20, 2025; SAP, "Business AI," accessed October 2025; Adobe, "Adobe launches Adobe Experience Platform Agent Orchestrator for businesses to activate AI agents in customer experiences and marketing workflows," news release, March 18, 2025; Workday, "Workday unveils Workday Build, giving developers the tools to build the future of work," press release, September 16, 2025; Erwin Huizenga and Bo Yang, "Agent Development Kit: Making it easy to build multi-agent applications," Google for Developers, April 9, 2025.
8. Joanna Glasner, "AI autonomous agents are top 2025 trend for seed investment," *Crunchbase News*, June 17, 2025; Jacob Robbins and Kia Kokalitcheva, "Y Combinator is going all-in on AI agents, making up nearly 50% of latest batch," PitchBook, June 11, 2025.
9. Gartner, *AI agents are disrupting SaaS pricing: What must CIOs do?*, July 16, 2025 (ID G00834627). GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.
10. Adrian Radu, "Billing infrastructure in the age of co-pilots and AI agents," Lightspeed, March 6, 2025.
11. Gartner, *AI agents are disrupting SaaS pricing: What must CIOs do?*, July 16, 2025 (ID G00834627). GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.
12. Maxio, *2025 pricing trends: Usage-based models and the path to SaaS growth*, 2025.
13. Zendesk, "Zendesk first in CX industry to offer outcome-based pricing for AI agents," August 28, 2024.
14. Salesforce, "Agentforce Observability," accessed October 2025; Google Cloud, Gemini Enterprise, accessed October 2025.
15. Bojan Ciric and Prakul Sharma, "Scaling AI agents may be risky without an enterprise marketplace," *Deloitte Insights*, September 15, 2025.

ACKNOWLEDGMENTS

The authors would like to thank **Diana Kearns-Manolatos**, **Faruk Muratovic**, **Rohan Gupta**, **Khusro Khalid**, **Laura Shact**, **Girish Srinivasan**, **Pavan Srivastava**, **and Prakul Sharma** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

New technologies and familiar challenges could make semiconductor supply chains more fragile

With escalating trade restrictions on critical next-gen AI chip technologies, leaders should adapt quickly to make supply chains more resilient

ARTICLE • 9-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

Geopolitical tensions and escalating trade restrictions are reshaping semiconductor supply chains, with far-reaching impacts for artificial intelligence chip innovation, the global economy, national security, and scientific progress. Many of these high-tech processes and materials rely on a handful of suppliers, whose dominance in key regions has prompted governments to impose trade barriers to protect strategic interests and reduce dependency. Making the world's most advanced chips for next-generation AI systems and high-performance computing data centers has, for a long time, meant navigating fragile supply chains, but the stakes are much higher now.

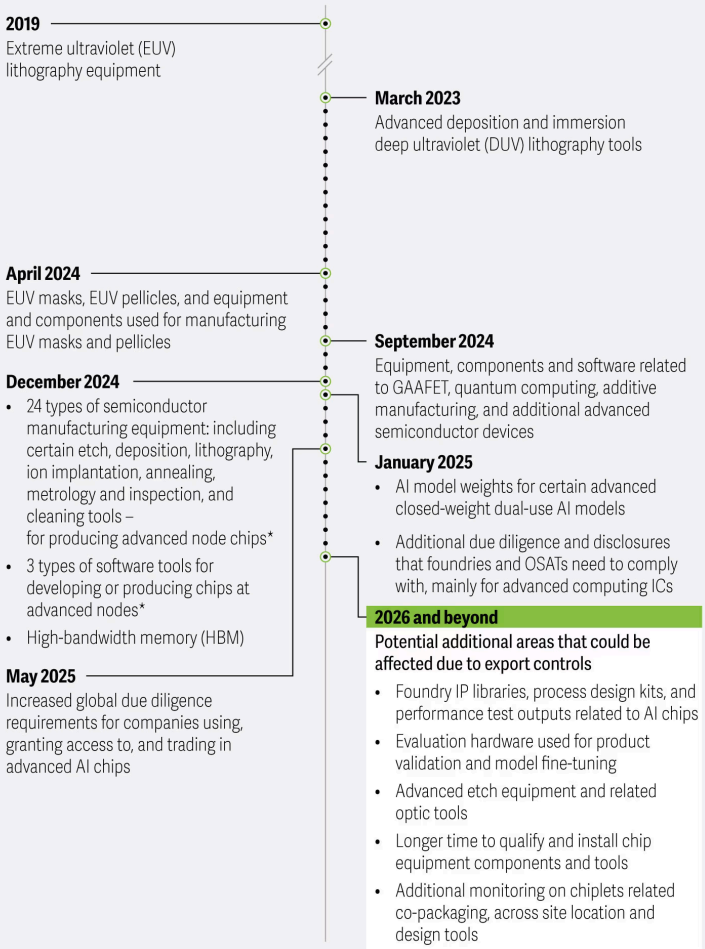
Deloitte expects that, by 2026, semiconductor technologies, including front-end and back-end chip manufacturing such as etching and gate-all-around (GAA) transistors, electronic design automation (EDA), and software tools that enable advanced AI models, will become additional supply chain chokepoints. And Deloitte predicts that, in 2026, at least US$30 billion will be spent on various critical technologies, including extreme ultraviolet (EUV) lithography equipment and high-bandwidth memory co-packaging tools, which will be affected by trade barriers.[1] However, this investment will be dwarfed by the approximately US$300 billion AI chips market that these technologies will enable, underscoring the critical role in the global semiconductor supply chain.[2]

AI (re)writes and (re)shapes global semiconductor supply chains

Deloitte's analysis of semiconductor content in AI data centers noted that the global semiconductor supply chain is deeply interdependent, and countries are working to protect their access to AI chips and hardware components that are critical for generative AI, high-performance computing, and autonomous systems.[3] Therefore, it's not surprising that export controls and other trade restrictions have started to affect a broader footprint of semiconductor equipment, materials, software, design tools, various kinds of chips, and packaging and assembly tools in 2025 and 2026 compared to two or three years ago (figure 1).

Figure 1

Trade controls in the United States and Europe have broadened to cover multiple types of semiconductor technologies in 2025, 2026, and beyond



Note: 2025 information as of October 8, 2025.

Source: Deloitte analysis. Data for 2019 to 2025 based on information gathered from publicly available sources including documents and announcements published on the sites of Federal Register and Bureau of Industry and Security (BIS). 2026 information based on conversations and forward-looking insights gathered from industry subject matter specialists. *Bureau of Industry and Security, "Commerce strengthens export controls to restrict China's capability to produce advanced semiconductors for military applications," U.S. Department of Commerce, December 2, 2024.

Deloitte. Insights | deloitteinsights.com

An AI system's performance depends on a narrow stack of several globally distributed technologies, including advanced AI logic design, leading-edge front-end node fabrication, and advanced packaging. Delivering these capabilities involves collaboration among multiple stakeholders, such as integrated device manufacturers, foundries, equipment makers, design vendors, outsourced semiconductor assembly and test (OSAT) vendors, system integrators, outsourced channel distribution partners, and government bodies from different countries.[4]

Export controls redefine the future of advanced AI logic design

In 2024 and 2025, US restrictions tightened and then eased on multiple critical semiconductor technologies, especially EDA tools.[5] EDA processes constitute the design logic, chip layout and placement, simulation, AI-enhanced design, verification, and integration workflows, all of which are vital for developing advanced AI accelerators.

As an example, there was an existing restriction for chips developed based on gate-all-around field-effect transistor (GAAFET).[6] GAAFET is an emerging transistor architecture for sub-5 nm and sub-3 nm logic design, offering performance and power efficiency benefits for compute-intensive gen AI workloads. In

December 2024, the United States further broadened export controls to include software and tools that support the development and design of advanced computing nodes.[7] As these new export controls emerge, they are likely to have implications for the broader EDA ecosystem and foundry partners in 2026.

Prediction and perspectives for 2026 and beyond

As restrictions on GAAFET-based chips increase, foundries in non-US allied countries using GAAFET process design kits for leading nodes will require EDA tool support for validation. But if a region lacks access to these tools, it may have to rely on older, less efficient nodes, or be pushed toward developing domestic EDA capabilities, both of which will likely stretch product cycles and dent competitiveness. Moreover, added controls on advanced computing chips and new controls on AI model weights have increased compliance requirements for companies collaborating with customers and business partners, especially in China.[8] Increasingly, AI models and the scale and quality of AI model weights are influencing the capabilities of AI-powered EDA tools that are used to design chips.[9]

By 2026, Deloitte predicts that EDA and logic design players will likely be impacted by these controls: They could face more intense checks and granular disclosure requirements regarding entity, location, and end use of foundry intellectual property libraries, process design kits, and performance test outputs tied to AI accelerators. Evaluation hardware, typically used for product validation and model fine-tuning (including reference model weights for testing purposes and outputs), may come under closer scrutiny.[10] Companies involved in AI hardware co-design may need to establish trusted country pathways or may have to retool workflows: For example, they could keep model weights within the United States or an ally's secure IT infrastructure while allowing foundry partners to run tests remotely.[11]

Chokepoints in developing leading-edge front-end node fabrication for AI systems

The United States and the Netherlands continue to restrict access to EUV equipment, which is widely regarded as essential for producing the most advanced process nodes.[12] While the United States does not have domestic EUV production capabilities, it influences which countries can buy these machines by coordinating export restrictions with allies (such as the Netherlands), mainly to secure technological and national security. At the same time, China has pushed forward to develop lithography equipment by customizing deep ultraviolet technology using multiple patterning techniques through its domestic chip equipment companies.[13] While these methods appear effective, they operate at much slower speeds and higher costs.[14] To safeguard national security interests, the United States introduced additional export restrictions on tools used for precision etching that are essential to carve intricate AI architectures.[15]

Prediction and perspectives for 2026 and beyond

Advanced etch technology is critical for fabricating leading-edge AI chips at sub-5 nm nodes. The chip industry employs double, quadruple, and spacer-based patterning to manufacture delicate features on the most modern AI chips.[16] As a result, the US-originated process equipment for etching, as well as etching equipment and tools designed or manufactured abroad using the United States' etch tech IP, could emerge as new chokepoints in 2026. In addition, components such as optics (lenses and mirrors) and reticles (photomasks), which are integral to wafer fabrication equipment and hold the blueprint of the pattern to be printed on a wafer, may also attract restrictions.

Furthermore, specialty gases (such as silane and fluorinated derivatives)[17] and critical minerals (including gallium, germanium, and antimony)[18] that are part of the advanced node manufacturing process introduce additional friction points in the global chip supply chains.

With a broad range of front-end process equipment, components, and input materials facing export controls, Deloitte predicts that sub-5 nm and sub-3 nm production ramps would continue to accelerate in

the United States, Taiwan, and South Korea through 2026 and beyond. Meanwhile, China is expected to continue focusing on mature deep ultraviolet technology with multiple-patterning workarounds.

Consequently, multinational chip equipment companies should adjust their front-end wafer fabrication–related capital expenditure planning at the regional level. Fabrication equipment vendors, components and parts suppliers, and foundries may face longer qualification, upgrade, and installation cycles compared to those experienced in 2024 and 2025. And as chip design companies adapt to the new requirements—developing de-featured or stepped-down AI XPUs (reduced performance versions of high-end AI chips) and region-centric process libraries to meet the growing gen AI chip demand in China and other non-US–allied countries—the need for enhanced support from front-end fabrication equipment providers will likely also rise.

Trade controls disrupt advanced packaging and testing

Advanced packaging technologies have quickly become strategic targets for export controls. Measuring and inspection equipment is facing export restrictions from the Netherlands [19] due to its critical role in high-density chip stacking,[20] an essential building block for current and future gen AI chips.[21] Specific types of chip equipment (etch, deposition, lithography, ion implantation, annealing, metrology and inspection, and cleaning tools) that are essential for testing and validating advanced AI chips are under export control.[22] This is because they're considered sensitive and potential dual-use technologies, and they may continue to attract additional trade controls in the future.

Prediction and perspectives for 2026 and beyond

As highlighted in the 2024 TMT Predictions, chiplets and heterogeneous architectures are fast emerging as preferred packaging models for gen AI chips designed for high-performance computing AI workloads.[23] However, the complexity involved in sourcing and packaging multiple dies and components from diverse vendors from different regions will likely make chiplets a major geopolitical chokepoint in 2026. Notably, chiplet-based solutions are estimated to be worth approximately US$100 billion to US$110 billion in annual revenues in 2026.[24]

High-bandwidth memory (HBM) has also become crucial for gen AI training and inference workloads. As of mid-2025, HBM co-packaging was being monitored more closely, including the identification of locations where HBM and logic are co-packaged.[25] As a result, semiconductor players involved in assembly, testing, and packaging will likely be required to provide additional disclosures. These may include naming the OSAT providers or back-end manufacturing vendors involved in packaging, specifying the location where the system is co-packaged, indicating the destination country where the interim or finished product is shipped to, and detailing relevant performance thresholds.

What is likely to become more prominent in 2026 and beyond is the growing dependence on the effectiveness of the back-end process to ensure new products reach the market on time. As routing and documentation requirements grow increasingly stringent for co-packaging sites—particularly those involving HBM, logic, and high-speed input/output—every aspect of the supply chain, from front-end wafer fab schedules and design sign-offs for EDA vendors to product launches by end-customer original design manufacturers and original equipment manufacturers, will become more dependent on the pace at which advanced packaging-related process clearances and procedures are completed. Any delays on the packaging vendor or the OSAT's side could affect yield ramps and tuning, in turn, triggering re-shoring or friend-shoring by relocating facilities to allied countries.

Collectively, these factors could impact the rollout of AI data centers planned for 2026 (and beyond) across multiple regions. Hyperscalers, cloud providers, and companies across industries combined are expected to spend roughly US$500 billion in 2026 and US$1 trillion in 2028 on AI data centers,[26] with chip solutions accounting for roughly 50% to 60% of that spending. Given the anticipated growth,

supply chain disruptions could affect tens or even hundreds of billions of dollars' worth of semiconductors over this three-year period.

The bottom line

China bolsters its domestic semiconductor ecosystem

Stringent export controls and restrictions on a range of semiconductor technologies have inhibited China's access to state-of-the-art AI chips. This has prompted China to accelerate domestic semiconductor innovation, especially as it sees the moves could hamper its progress toward sub-7 nm and sub-5 nm, even as non-China chip fabs move from 3 nm and 2 nm in 2025 to 1.8 nm in 2026 and 2027.[27]

As China develops workarounds to deal with export controls, it may explore multiple facets of the global semiconductor supply chain, not just front-end manufacturing but also chip design and advanced packaging.[28] While sophisticated chips using older manufacturing nodes can be used for advanced packaging, the United States is likely to implement additional controls and checks to limit the performance of such packaged systems meant for leading-edge AI chips.

Race to build sovereign tech stacks accelerates, ushering new regional equations

Technology sovereignty is aspirational as countries aim to independently develop, control, and regulate digital technologies.[29] Since AI is widely viewed as the next major driver of economic development and national competitiveness, its ecosystem is receiving attention as governments seek greater direct control over its digital infrastructure. Countries and regions do not want to be left further behind or involuntarily forfeit their authority. This urgency is heightened because advanced AI capabilities are currently concentrated among a few countries and companies. Moreover, as both the United States and Europe are reshoring high-end chip manufacturing, they are likely to invest in alternative advanced assembly and test hubs through 2026 and beyond, domestically as well as in countries such as India, Vietnam, and Malaysia.[30]

Need for the semiconductor industry to bolster supply chain resilience

Chip companies across the ecosystem may need to proactively prioritize resilience through internal stress-testing exercises, primarily to self-assess their end-to-end supply chains and bolster cybersecurity preparedness.[31]

Robust supply-chain diversification across regions and investment in alternate sourcing strategies and channel partnerships are crucial. The strategic importance of securing independent supply chains for critical materials and components requires accelerated localization and regulatory adaptability. Moreover, geopolitical issues could fragment global AI ecosystems, presenting risks such as exporting chips through gray markets and intensifying pressures on companies to bolster product and supply chain monitoring and tracking capabilities.

Though the market for AI inference-optimized chips is expected to grow to billions of dollars in 2026, most of the advanced computing will be performed on leading-edge AI chips that would mainly reside in hyperscale data centers or at on-prem servers that use the same chips and racks as data centers do.[32] Therefore, new and additional export controls and requirements could possibly be directed at AI inference chips and related infrastructure, for which the broader semiconductor industry should develop alternate supply chain options across sourcing to distribution.

And with the shift from training to inference, software's importance as a more integral part of semiconductors will also grow, for instance, using software programming techniques to reconfigure one large monolithic AI GPU (meant for training) into multiple smaller GPU slices or virtual GPU instances (usable for inference).[33]

Additionally, US- and Europe-based device original equipment manufacturers may need to shift production and assembly away from China and toward the emerging hubs in Southeast Asia and India. This shift could increase costs in the short term, potentially driving consumer tech device inflation. Semiconductor companies should remain agile and operate at scale, anticipate and adapt to evolving trade patterns beyond 2026, and explore alternate strategic country-level alliances to safeguard critical logistics routes and infrastructures.

As trade tensions reshape global alliances and channel partnerships, the chip industry's resilience faces an unprecedented test heading into 2026. The interconnected and highly strategic nature of global chip supply chains highlights the urgent need for proactive engagement and collaboration among multiple industry stakeholders to make the semiconductor supply chain more resilient.

BY

Karthik Ramachandran
India

Duncan Stewart
Canada

Jeroen Kusters
United States

Deb Bhattacharjee
United States

Girija Krishnamurthy
Global

Jan Thomas Nicholas
Malaysia

ENDNOTES

1. A note to methodology. Estimates include projected aggregate spending for 2026 on extreme ultraviolet equipment, AI-based etch equipment, select advanced packaging equipment including high-bandwidth memory co-packaging tools, and AI chip design software and tools.

2. In 2025, Deloitte Consulting LLP performed an analysis of the data center market, including a rough bill of materials for the various components and market sizes. This analysis is due to be published in December 2025.

3. Ibid.

4. Ibid. Importantly, an AI server rack is not just a monolithic unit but a far more complex, integrated system that comprises tens of thousands of components ranging from advanced chips, memory dies, analog integrated circuits, controllers, power devices, and passives like substrates and capacitors.

5. Karen Freifeld and Surbhi Misra, "**As trade war truce with China holds, US lifts curbs for chip design software and ethane**," *Reuters*, July 3, 2025; Joe Cash, "**China says successful US trade talks make return to tariff war unnecessary**," *Reuters*, July 18, 2025.

6. Bureau of Industry and Security and US Department of Commerce, "**Federal Register, vol. 89, no. 173**," Sept. 6, 2024.

7. New software and technology controls included restrictions on electronic computer-aided design and technology computer-aided design software and technology, especially when these are used for designing advanced node-integrated circuits. To read further, see: Bureau of Industry and Security and US Department of Commerce, "**Commerce strengthens export controls to restrict China's capability to produce advanced semiconductors for military applications**," Dec. 2, 2024.

8. Bureau of Industry and Security and US Department of Commerce, "**Framework for artificial intelligence diffusion**," Federal Register, Jan. 15, 2025.

9. Wenji Fang, Jing Wang, Yao Lu, Shang Liu, Yuchao Wu, Yuzhe Ma, and Zhiyao Xie, "**A survey of circuit foundation model: Foundation AI models for VLSI circuit design and EDA**," *arXiv*, March 28, 2025.

10. For further information on AI model weights related technology controls, see: US Department of Commerce and Bureau of Industry and Security, "**Federal Register, vol. 90, no. 9**," Jan. 15, 2025.

11. Insights based on conversations and interviews with Deloitte experts in the areas of the semiconductor industry, supply chains, and export control impact.

12. Chris Miller, "**How US export controls have (and haven't) curbed Chinese AI**," AI Frontiers, July 8, 2025.

13. Stefano Lovati, "**China invests €37 billion to develop domestic EUV lithography systems**," *Power Electronics News*, Feb. 11, 2025.

14. Pablo Valerio, "**China semiconductor ambition and adversity**," *EE Times*, May 19, 2025. Additionally, US regulations included restricting and capping the production of advanced AI chips far below the domestic demand in China.

15. See Bureau of Industry and Security and US Department of Commerce, "**Federal Register, vol. 89, no. 173**," p. 7. As noted in this document, atomic layer etching helps produce vertical edges required in high-quality, leading-edge advanced devices and structures, including gate-all-around field-effect transistor and similar 3D structures. Anisotropic dry etching is critical for gate-all-around field-effect transistor and similar 3D structure fabrication. It is also an important tool for fin-shaped field effect transistor (FinFET) fabrication.

16. Ibid.

17. US Department of Commerce and Bureau of Industry and Security, "Foreign-produced direct product rule additions, and refinements to controls for advanced computing and semiconductor manufacturing items," Dec. 5, 2024.

18. Sara Bulter, "How China's rare earth metals export ban will impact supply chains in 2025," Optilogic, Feb. 17, 2025.

19. Deloitte analysis based on conversations and insights gathered from industry experts and cross-validated with multiple secondary sources, including: Abbie Windsdale, "Netherlands takes bold step to tighten semiconductor export control," Tech Announcer, Jan. 16, 2025.

20. For example, hybrid bonding is fundamental to developing advanced 2.5D and 3D chip designs and heterogeneous architectures (or chiplets), as it enables ultra-fast data transfers (up to 17 TB/s) that are critical for AI and high-performance computing. To read further, see: Sam Naffziger, "Future of AI hardware enabled by advanced packaging," IEEE Electronics Packaging Society, May 28, 2024.

21. Duncan Stewart, Karthik Ramachandran, Prashant Raman, and Ariane Bucaille, "Silicon building blocks: Chiplets could move Moore's Law forward," *Deloitte Insights*, Nov. 19, 2024.

22. Bureau of Industry and Security, "Commerce strengthens export controls to restrict China's capability to produce advanced semiconductors for military applications," press release, Dec. 2, 2024.

23. Stewart, Ramachandran, Raman, and Bucaille, "Silicon building blocks."

24. Xiaoxi He and Yu-Han Chang, "Chiplet technology 2025-2035: Technology, opportunities, applications," IDTechEx, accessed Oct. 1, 2025.

25. US Department of Commerce and Bureau of Industry and Security, "Foreign-produced direct product rule additions, and refinements to controls for advanced computing and semiconductor manufacturing items."

26. Duncan Stewart, et al, "Why AI's next phase will likely demand more computational power, not less," Deloitte Insights.

27. For context, state-of-the-art chip fabs in the United States and Taiwan were already pushing the boundaries toward sub 7 and sub 5 nm as of 2020 to 2021, indicating China is probably at least four to five years behind (see Deloitte 2024 semiconductor outlook). Therefore, initiatives such as Beijing's Big Fund III actively support the expansion of local semiconductor capabilities, notably electronic design automation (EDA) and lithography tech development. To read further, see: Anton Shilov, "China to pivot $50 billion chip fund to fighting U.S. squeeze as trade war escalates — country to back local companies and projects to overcome export controls," Tom's Hardware, June 27, 2025.

28. The Chinese Academy of Sciences worked with domestic chip design players on an open-source project to develop an AI system that used large language models to accelerate chip design and build fully functional central processing units. To read further, see: Mark Tyson, "China claims to have developed the world's first AI-designed processor — LLM turned performance requests into CPU architecture," Tom's Hardware, June 12, 2025. Additionally, Huawei's breakthroughs in developing EDA tools capable of supporting 14 nm processes and above mark significant milestones. To read further, see: Omar Sohail, "Huawei has reportedly developed 14nm EDA tools, which the company will employ to mass manufacture its Kirin 9020, but the company is still limited to the 7nm architecture," WCCF TECH, June 11, 2025.

29. David Jarvis, et al, "A new era of self-reliance: Navigating technology sovereignty," Deloitte Insights.

30. Analysis based on multiple publicly available secondary sources that discuss the chip industry's plans to commence new AT hubs in countries including India, Malaysia, and Vietnam.

31. Aside from trade-related issues, as we already mentioned in our 2024 Global Semiconductor Outlook report, cyber threats are surging, requiring chip fabs and AI systems to intensify security

measures against malware targeting critical infrastructure.

32. Duncan Stewart, et al, "**Why AI's next phase will likely demand more computational power, not less**," Deloitte Insights. Deloitte analysis based on conversations and insights gathered from industry experts.

33. Gwangoo Yeo, Jiin Kim, Yujeong Choi, and Minsoo Rhu, "**PREBA: A hardware/software co-design for multi-instance GPU based AI inference servers**," *arXiv*, Nov. 28, 2024.

ACKNOWLEDGMENTS

The authors would like to thank **Nina Zhang**, **Amy Scimeca**, **Karan Aggarwal**, **Jesse Singh**, **Michael Greco**, and **Pablo LeCour** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

Tiny episodes, massive appeal: Short-form serials are gaining viewers and empowering independent studios

From independent creators to major platforms, micro-series are helping redefine how viewers connect and consume content worldwide

ARTICLE • 8-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

Foragers, snackers, and grazers: Confronted by an abundance of content, social media audiences are constantly searching for something good to consume. They sift through endless streams that might match their metadata but might not satisfy their emotional or intellectual appetites. Could serialized short-form storytelling—like micro-series and micro-dramas—offer greater sustenance and continuity in a highly fragmented attention economy?

A micro-series—sometimes called a micro-drama or short-form serial—is a scripted video series told in bite-sized episodes lasting just a few minutes each, designed for mobile-first consumption and rapid engagement. Mobile apps like DramaBox, ReelShort, ShortMax, and DramaWave, among others, are generating billions in revenue and hundreds of millions of users in Asia and the United States.[1] This explosive growth is redefining what audiences expect from digital entertainment—and signals new opportunities and challenges for creators, platforms, and brands alike.[2]

In 2025, in-app revenue for micro-series content is forecast to reach US$3.8 billion.[3] **In 2026, Deloitte predicts** that the revenue growth of in-app micro-series will more than double, reaching US$7.8 billion. Deloitte also predicts that the United States will account for half of global revenue in 2025, but its share will decline to 40% as other markets convert more views and downloads into cash. As more audiences are exposed to micro-series, we believe they will find the combination of short-form and serial entertainment compelling, buoyed by increased virality on social media. Additionally, we anticipate that more micro-serials will break out on social platforms, commanding more attention time and climbing the charts of US social media engagement. Finally, we expect some savvy video-streaming providers will experiment more with short-form serialized content offerings directly on their services.

Although short-form serials seem like a made-for-social innovation, they could challenge the dominance of leading social platforms. The capricious nature of the algorithmic feed on social platforms could make it difficult to "follow" a series and keep up with new episodes. This could push more audiences and creators onto competing micro-series apps. At the same time, there is some evidence that younger generations are feeling overwhelmed by social media, unable to keep up with, or let go of, the infinite feed.[4] Could a short-form throwback to linear TV be the solution?

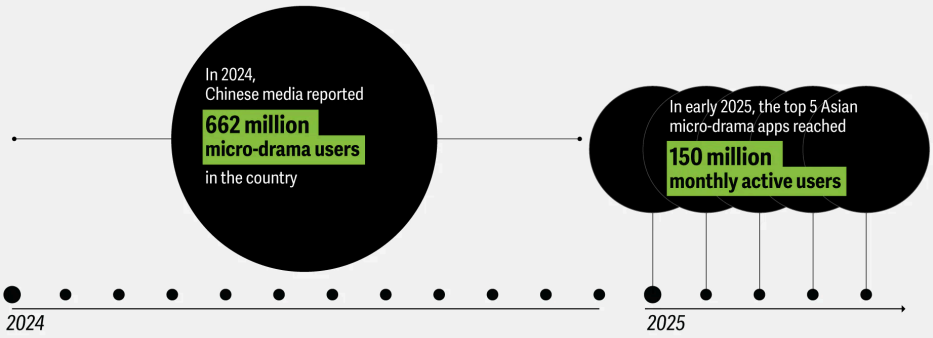
Micro-dramas are capturing audiences

The growing popularity of serialized short-form content appears to be gaining traction on leading social video platforms while also supporting the growth of new competitors: successful micro-drama services that are competing for the finite amount of time people have for digital entertainment.

Micro-drama mobile apps are growing in popularity, offering potentially hundreds of 60- to 90-second serialized episodes loaded with plot twists and cliffhangers to keep audiences engaged and wanting more. These new micro-serials are produced quickly and cheaply, constantly refined by audience interactions, and often leverage leading social video platforms to drive discovery and buzz.[5]

Figure 1

Micro-drama popularity in Asia has helped drive global success



Source: Omdia, "Emerging Micro-drama Trend in Asia," March 25, 2025.

Deloitte. Insights | deloitteinsights.com

China's iQiyi offers over 15,000 free and paid micro-dramas and has seen considerable growth in its watch time over the past year, adding e-commerce capabilities around the micro-drama ecosystem.[6] As of 2024, Chinese media reported approximately 662 million micro-drama users nationwide.[7] Leading Chinese video streamers are partnering with short-video platforms to coproduce premium mini-dramas, perhaps anticipating an integrated future of long- and short-form content.[8]

Given the growing global momentum of micro-dramas, India is seizing the opportunity with a surge of innovative platforms and established media companies entering the market.[9] Over-the-top platforms like Zee Entertainment, and Kuku FM have launched dedicated micro-drama verticals, with platforms reporting doubled daily watch times following the introduction of short-form pilot content.[10] In India's highly price-conscious market, where average revenue per user is modest, platforms are experimenting with micro-payment options for individual episodes, while others are rolling out flexible subscription plans, including hybrid models that blend subscription fees with advertising income.[11] Viewers can watch the first few episodes for free, but then they need to pay to watch the story unfold.

Leading micro-drama apps now regularly appear among the top 25 US app store downloads.

The appetite for short-form serials and micro-drama apps is also spreading beyond Asia.[12] One report found that global revenue from micro-drama apps surged from US$178 million in Q1 2024 to nearly US$700 million in Q1 2025.[13] The United States has become the top-grossing market for short-drama apps like DramaBox, ReelShort, and GoodShort.[14] Leading micro-drama apps now regularly appear among the top 25 US app store downloads.[15] Crossovers onto social video platforms have given them a

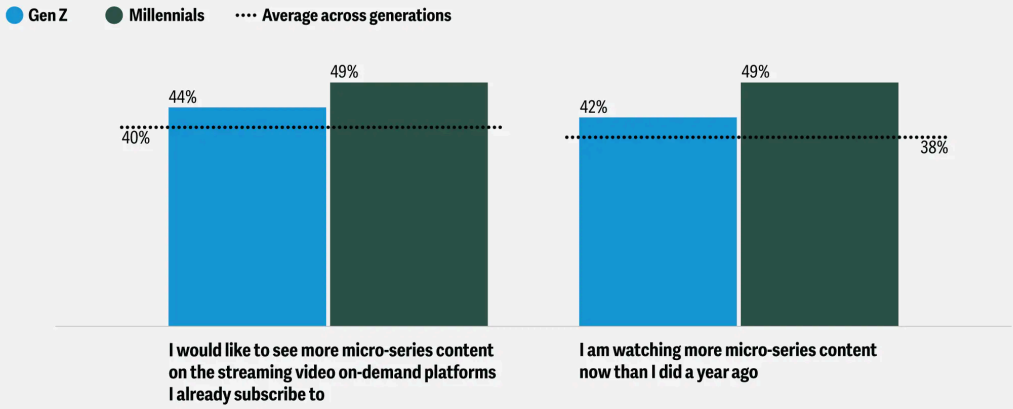
boost in the United States,[16] and social platforms themselves are seeing more engagement with micro-series content.[17] Even leading streamers are dabbling in short-form video and vertical content.[18]

Deloitte's own **Digital Media Trends** survey of US consumers found that in March 2025, about 30% of Generation Z and millennials were familiar with micro-series or micro-dramas. Among them, nearly half are watching more micro-series content now than they did a year ago, and nearly half would like to see more micro-series content on the subscription video-on-demand platforms they already subscribe to, suggesting a competitive path for streamers (figure 2).

Figure 2

Younger generations are engaging more with serialized short-form video content: micro-series and micro-dramas

Percentage of consumers who are familiar with micro-dramas or micro-series who selected "strongly agree" or "somewhat agree"



Note: n (DMT 19.5 respondents who are familiar with micro-dramas or micro-series) = 1,620; n (Gen Z) = 483; n (millennials) = 497; n (Gen X) = 399; n (boomers)=230; n (matures) = 11.

Source: 2025 Digital Media Trends, 19.5 edition.

Deloitte. Insights | deloitteinsights.com

More audiences are being drawn to independently created short-form, narrative-driven content. In response, a growing number of creators are building their own independent studios, leveraging data, artificial intelligence, and social platforms to amplify their reach.

The rise of social, cost-effective, and data-driven studios

Media and entertainment are being reshaped by the behaviors and economics of short-form content, the capabilities and reach of social video platforms, and the dual forces of prestige and unwieldy costs associated with premium content. As audiences devote more of their entertainment time and ascribe greater value to short-form content, independent creators are evolving into modern studios, elevating the quality of independent video at a fraction of the cost.

More than just an economic advantage or capitalizing on shifting audience behaviors, new creator studios can be fast and responsive, quickly adapting to audience feedback. They leverage engagement data to see what works and what doesn't, reducing the risks of content decisions.[19] They interact with viewers to reinforce community bonds and grow fandoms. They employ AI wherever it can shorten time-to-market, reduce production overhead, and grow their reach across geographies.[20] And they are free to experiment with editorial tactics that maximize engagement and retention.

Rising engagement with micro-series could mint more creator studios powered by audience interactions and amplified by technologies, making it easier for them to move fast and reach global audiences. It could also drive more competition for top creator talent among micro-drama apps, streaming video services,

and social video platforms. Of these, social platforms could be the least advantaged unless they make it easier for audiences to discover and keep up with serials.

Tools and tactics supporting modern short-form content creators

Creators and media executives should consider how new independent studios—built from scratch—are leveraging tools and platforms to reach and engage global audiences at minimal cost.

AI-enabled production pipelines: Studios can use generative AI tools to compress production cycles and lower the barrier to achieving high production value, like auto-generating B-roll or simple animations, and automated subtitle generation, voice cloning, or dubbing AI to overcome language, dialect, and accent gaps.[21] Tools are also emerging that can convert long-form stories into short-form serials.

Editorial tactics: Stories often leverage tricks like increasing the density of "hooks" (plot twists and reveals) and ending every episode on a cliffhanger to trigger the viewer's "need-to-know" impulse. Successful micro-dramas frequently borrow from fan fiction and online novel tropes, such as rich versus poor romances and time-travel revenge, which have proven audience appeal.[22]

Community growth: Creators should engage viewers through comments, even adapting later episodes in response to fan feedback. Participating in behind-the-scenes streams and social media discussions can help transform a series into a thriving fandom. Releasing episodes on a steady schedule of daily drops at consistent times helps build appointment-viewing habits. Micro-dramas can become durable mini-soap franchises with recurring characters or themes that can extend into multiple seasons or spin-offs. Creators who master cross-promotion, intellectual property merchandising, and multiplatform distribution—like novelizations or soundtrack releases from AI-generated music—can find an edge in sustaining their "mini-Marvel" universes on a budget.

New key performance indicators: A data-driven feedback loop can help short-drama studios test multiple storylines and double down on those with high retention. Key metrics include tracking completion rates, average episodes watched per user, series subscription uptake, and even story-specific return on investment, like whether a series drives merchandise sales or increases platform watch time.

Monetization: From episodic micro-payments and monthly subscriptions to soundtrack sales, merchandising, ads, and product placement, short-form serials are exploring multiple monetization pathways to fund their growth.[23]

An antidote to brain rot and doomscrolling?

The 2024 word of the year, as determined by Oxford University Press, was "brain rot," a term used to "capture concerns about the impact of consuming excessive amounts of low-quality online content, especially on social media."[24] A related term, "doomscrolling," reflects the tendency of some users to display addictive behaviors on social media.[25]

Some evidence suggests that more people are moving away from social media toward smaller, more intimate, and protected sources of information, entertainment, and community.[26] Perhaps this is a response to a sense that social media is no longer "social" but just "media"—too fragmented and commoditized.[27] It could be due to documented concerns about the mental health implications of overuse.[28] It could soon arise from an inability to trust online content and information, as some synthetic content begins to erode truth and evidence. Or it may simply be that foraging for intermittent rewards is fundamentally tiring and unfulfilling.[29]

Micro-dramas may not signal a great change in the new mass media, but the growing interest in more serialized short-form independent content could challenge social video platforms, traditional studios, and streamers, shifting the balance of power and potentially elevating a new tier of high-quality, cost-effective independent studios.

Some evidence suggests that more people are moving away from social media toward smaller, more intimate, and protected sources of information, entertainment, and community.

The bottom line: Creators and independent studios are becoming more capable of meeting and responding to audiences

Independent creators and studios are gaining influence, forging closer ties with brands, and discovering new channels to reach their audiences. Challenges with engagement and monetization on social platforms —or the chance to come together on their own platforms—could push more creators toward other channels, such as video streamers and micro-drama apps, and even lead to new creator-led entertainment services.

Creators are amassing audiences across platforms and building closer relationships with brands.[30] This has led to some tension between popular creators and the platforms they publish on, particularly around profit-sharing and content moderation.[31]

Social media platforms have evolved from the social graph to the interest graph, offering endless streams of content based on user interactions rather than on who users explicitly choose to follow. This can make it harder for creators to connect with their audiences when the algorithm decides that something else is more likely to foster engagement with the platform. Creators can spend large amounts of time and money developing audiences and brand relationships, only to see their content deprioritized or even shut down by seemingly capricious algorithms.

For the most part, algorithmic feeds and interest graphs are not geared to support serialized narratives. If micro series popularity continues to rise, social platforms may adapt trending algorithms to support ongoing narrative content, helping creators effectively reach and retain dedicated viewers, and signaling a potential shift in media consumption dynamics. They may implement “series-aware” algorithms, or “continue watching” rails.

The alternative could see more creators migrating to dedicated creator-studio applications, like micro-drama apps that are dominating mobile downloads. They may also be poached by streaming video providers looking for more short-form content to fill their slates and appeal to younger audiences. Indeed, streaming video services could be well-positioned, having built their services on serialized and appointment-based content. If more creator studios embrace serialized short-form productions, streamers could be the beneficiaries. Or perhaps the time is right for new, creator-led platforms to emerge, built on the technologies and learnings of streamers and social media.

Chris Arkenberg
United States

Tim Bottke
Germany

Ankit Dhameja
India

Gillian Crossan
Global

ENDNOTES

1. Rui Ma, "State of short drama apps 2025," Mobile App Insights, July 2025.
2. Stephanie Yang, "Two-minute TV shows have taken over China. Can they take over the world?" *Los Angeles Times*, March 16, 2025.
3. Ma, "State of short drama apps 2025."
4. Gaby Hinsliff, "It's the age of regret: Gen Z grew up glued to their screens, and missed the joy of being human," *The Guardian*, March 7, 2025.
5. *Xinhuanet*, "Love, twist and one-minute cliffhangers: China's micro dramas go global," July 29, 2025.
6. CMB Global Markets' equity research, May 23, 2025 (private report accessed via AlphaSense).
7. Mandy Zuo, "China's addictive micro-dramas show how commercial demand is fuelling a netcasting boom," *South China Morning Post*, March 27, 2025.
8. Jeff Huang, "How China's $7 billion micro drama industry is taking on the US entertainment industry," *CNBC*, July 22, 2025.
9. Kunal Purandare, "The VC-backed rise of micro dramas in India," *Forbes India*, Aug. 8, 2025.
10. Systematix Institutional Research and Morning Brew, July 10, 2025 (private report sourced via AlphaSense).
11. *The Economic Times*, "Stage set for micro-dramas; WhatsApp's monetisation bid," June 17, 2025.
12. Robert Steiner, "Microdrama plot twist: A threat to the apps' stratospheric US growth," *Variety*, April 30, 2025.
13. Ma, "State of short drama apps 2025."
14. Ibid.
15. Appfigures, "Top ranked iOS app store apps," accessed Oct. 23, 2025.
16. Steiner, "Microdrama plot twist: A threat to the apps' stratospheric US growth."
17. Paige Gawley, "People on TikTok are obsessed with a fake group chat," Vice, April 9, 2025.
18. Lauren Forristal, "Netflix is getting into short videos with a new vertical feed for mobile," *TechCrunch*, May 7, 2025.
19. Global economic outlook and investment strategy, 2H 2025; ICBC International Research (private research brief via AlphaSense).
20. Carson Taylor, "Microdramas: China's new craze goes global," *Naavik*, Sept. 8, 2024.
21. Focus on structurally high-growth segments, Huatai Securities, Aug. 23, 2025 (private documents sourced via AlphaSense).
22. Kristian Monroe, "Told one minute at a time, micro dramas are soap operas designed to fit in your hand," *NPR*, March 19, 2025.
23. Taylor, "Microdramas: China's new craze goes global."
24. *Oxford University Press*, "'Brain rot' named Oxford word of the year 2024," Dec. 2, 2024.
25. Sian Boyle, "Is doom scrolling really rotting our brains? The evidence is getting harder to ignore," *The Guardian*, Dec. 9, 2024.

26. Annalee Newitz, "Social media is dead – here's what comes next," *NewScientist*, July 23, 2025.

27. Rodney Mason, "Social isn't social anymore—now what?" *Forbes*, April 28, 2025.

28. Jessica A. Kent, "Need a break from social media? Here's why you should—and how to do it," Harvard Summer School, Aug. 28, 2023.

29. Sanzana Karim Lora, Sadia Afrin Purba, Bushra Hossain, Tanjina Oriana, Ashek Seum, and Sadia Sharmin, "Infinite scrolling, finite satisfaction: Exploring user behavior and satisfaction on social media in Bangladesh," *Arxiv*, April 15, 2025.

30. Deloitte Digital, "2025 state of social research: How efficiency can meet impact with the right investments," May 15, 2025.

31. Gillian Follett, "How creators are shaping Cannes Lions—from business discussions to the campaigns winning awards," *AdAge*, June 12, 2025.

ACKNOWLEDGMENTS

The authors would like to thank **Jeff Loucks**, **Brooke Auxier**, and **Daniela Gonzales** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

Video podcasts dominate: Opportunity for brands, competition for traditional video

Podcasting is becoming a video-first, multilingual medium with booming reach that may help brands reach global audiences, while occupying a larger share of viewers' screen time

ARTICLE • 7-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

Podcasts aren't just for listening anymore: Now, you can watch them too. Video podcasts—or vodcasts—are redefining how audiences consume long-form media by blending audio storytelling with visual appeal, making content more immersive and shareable. As they blur the lines between podcasts, social media, and streaming video, creators are using cross-platform distribution to boost engagement, build communities, expand ad revenues, and unlock new sponsorships. And now, vodcasts may be starting to claim the screen time once monopolized by traditional TV and streaming platforms.

At the same time, audiences in emerging markets like India, Nigeria, and Brazil are embracing podcasts for their mobile-first, low-bandwidth appeal. The rise of localized, multilingual content is also fueling this growth, making podcasts a truly global and culturally diverse medium—even as issues around monetization, language accessibility, and infrastructure gaps pose real obstacles. **Deloitte predicts** that the annual global ad revenues for podcasts and vodcasts will reach roughly US$5 billion in 2026, marking a nearly 20% year-over-year increase in revenues.[1]

Combine the rising popularity of vodcasts with the global expansion of podcasts, and what you get is a market poised for significant growth in terms of audience, reach, and advertising revenues. Still, the industry's path forward will likely depend on how effectively creators and platforms can navigate challenges around discoverability, monetization, and scalability.

Vodcasts innovate to engage audiences in new ways

These visually focused podcasts will continue to gain traction with consumers—and advertisers—into 2026 and beyond. The drivers of this growth are likely threefold: their seamless integration into existing and popular media platforms, the use of social clips to drive buzz and virality, and their ability to more deeply connect vodcast creators with their loyal audiences.

We predict that the percentage of popular podcasts with video will rise and consumers will increasingly gravitate toward platforms that embrace video.

Some streaming music and audio services, including Spotify, Wondery, Podbean, and YouTube,[2] have integrated podcast video feeds directly into their user interfaces in recent years, making them accessible and available to consumers. In tandem, some major providers have equipped podcasters with the tools and know-how needed to create and monetize their video assets, making vodcast offerings more plentiful.[3] Although some of the other podcast services that are supporting vodcasts are long-time and well-known players in the space, YouTube is a relatively new entrant as of 2022,[4] but is already having an impact: They boasted one billion monthly vodcast viewers in early 2025, launched a ranking of top podcasts list for the US market,[5] and set a Guinness World Record for a vodcast episode of "New Heights" in August of 2025 with a live audience of 1.3 million concurrent viewers.[6] Making these vodcasts available on popular platforms and services—where many people are already spending their time and subscription dollars—lowers the barriers to entry for many consumers and increases uptake and engagement with both content and ads. Different podcast platforms have different approaches to video versus audio-only formats: Some are audio only, or nearly so, while others feature all podcasts with video versions. Spotify, which only started offering video versions in 2022, now has over 60% of its most popular shows offering a video component as of mid-September 2025.[7] We predict that the percentage of popular podcasts with video will rise and consumers will increasingly gravitate toward platforms that embrace video. For their part, as of the Fall 2025 Digital Media Trends report, 27% of US consumers say they watch vodcasts weekly, a trend that is led by Gen Zs and millennials.[8]

The video component of vodcasts brings audiences into the conversation in a way audio alone simply can't. Viewers see hosts' facial expressions, body language, and the visual context of the environment, which creates a sense of closeness. When viewers see the podcast hosts they know and love, the parasocial relationship strengthens—along with trust and perceived authenticity—which drives community and engagement.[9] All this engagement adds up: Users who *watch* vodcasts consume 1.5 times more content than users who only *listen* to podcast content.[10] Video also adds a compelling visual storytelling layer that appeals to younger, digitally native audiences and enables creators to reach new viewers across platforms. Advertising and sponsorship opportunities also increase with a visual format, as they allow for logo and product placements and the creation of short clips ready-made for social sharing.[11]

As such, social media platforms are also key to the success of vodcasts.[12] Short vodcast clips can be repurposed and shared across social media platforms to reach different audiences and highlight the most engaging and buzzworthy parts of an episode. The use of social platforms and viral social clips extends the reach of the vodcast, allows for discoverability, expands the scope for ads, and gives the podcast creator a chance to directly interact with their audiences.

The vodcast boom may continue to put pressure on other video entertainment offerings and platforms, as these video assets compete for coveted—and finite—screen and TV time. Already in 2024, almost half of podcast viewers say they watch on a connected TV.[13] And competition from this format may be what's pushing some traditional streaming video providers to think about entering the podcast space.[14] As video podcasts popularize and take over the living room, it's clear the format is changing consumers' behaviors in ways worth paying attention to: Whereas audio podcasts can be consumed while doing other activities like commuting or exercising, vodcasts require more focused attention. Forty-four percent of US vodcast watchers say they never multitask while watching compared with 29% of podcast listeners who say they never multitask while listening.[15]

This more focused attention on the content could lead to greater engagement and increased subscriber growth, which can attract more attention and investments from advertisers and sponsors who want in. As it stands, roughly a quarter of US podcast watchers and listeners (and more than a third of Gen Z and millennial podcast watchers and listeners) say they often purchase products or services that they hear advertised on podcasts, according to the latest Fall 2025 Digital Media Trends data.[16] More advertisers and sponsors mean increased revenue, which will drive reach, growth, and innovation. In short, the rise of vodcasts is likely to define the next chapter of the podcast industry's evolution.

Local, multilingual podcast content: Coming to a market near you

What started as a largely US-centric audio format is rapidly evolving into a dynamic global medium.[17] While the percentage of weekly podcast listeners varies widely across regions, the global average is around 22%—with markets like Indonesia (42.6%) and Mexico (41.8%) leading the way in listenership.[18] Several factors are fueling the surge in podcast consumption across emerging markets: expanding mobile connectivity, increasing global investments by streaming audio platforms, and growing availability of local and multilingual podcast content and vodcasts.

Established audio streamers are expanding globally and fueling podcast growth by investing in local language content.

The expansion of mobile internet access globally has democratized connectivity and content consumption. In countries like India, Nigeria, and Brazil, affordable smartphones and data plans have brought millions online.[19] For example, mobile data costs in Nigeria have dropped by roughly 97% over the past decade: While 1GB of mobile data cost US$11.15 in 2014, by 2023, it decreased to US$0.39.[20] Access to lower-cost devices and plans is making on-demand audio and video content—like podcasts and vodcasts—accessible to more people in more places.

Established audio streamers are expanding globally and fueling podcast growth by investing in local language content. Spotify, for instance, is funding creators and forming exclusive partnerships across Latin America, Africa, and Asia to develop regionally relevant shows.[21] Other platforms are licensing popular local podcasts, producing original content, and building tools to support regional talent.[22] Despite much of the industry being English centric, there is a growing understanding that multilingual and culturally specific programming could be key to sustaining global podcast growth. Meanwhile, new platforms in countries like Lebanon, India, and Nigeria [23] are emerging as hubs for local content and are increasingly partnering with global players to expand their reach.

New formats like vodcasts are also driving engagement, as they appeal to younger, digitally native audiences in emerging markets that typically have younger populations than more developed economies.[24]

The global rise of podcasts has implications for monetization strategies, with platforms boosting discovery and consumption especially for non-English content.[25] More and more, multinational brands might lean into ad placements in regional shows and within culturally relevant storytelling to reach diverse, engaged audiences—though cost per mille in emerging markets remain low, making revenue generation challenging for creators. Despite the obstacles, market globalization is fueling a surge in localized ads, branded content, and creator partnerships that cross borders.[26]

The bottom line: Untapped global audiences and new growth opportunities

Vodcasts and podcasts are taking over the living room and pushing into emerging markets globally, presenting opportunities and challenges for several players in the media and entertainment space.

Streaming audio and music platforms might focus on building—or improving—tech capabilities that allow for the seamless streaming of vodcasts directly within their app, though this involves investments in infrastructure, technology systems, and personnel. For those with existing capabilities, exploring dynamic, shoppable elements in vodcast advertising—like the ability to click right on the video ad to shop or

purchase—is the next step toward securing lucrative partnerships and sponsorships with advertisers, and growing industry monetization.

Success will likely depend on the ability to localize content, build trust with diverse audiences, and navigate an increasingly complex competitive landscape.

These same platforms should explore expansions into emerging markets, which might include investing in local and regional content and personalities (as well as gen AI capabilities to auto-translate and lip-synch audio and video content [27]) to appeal to new markets and grow audiences. Though there are upsides, globalizing the market involves a nuanced understanding of regional preferences, regulatory environments, and monetization models. Success will likely depend on the ability to localize content, build trust with diverse audiences, and navigate an increasingly complex competitive landscape. Streaming audio providers should also explore offering offline access to content, file compression, and lower bitrate streaming modes, especially in regions where bandwidth is still expensive and networks are unstable.

There may also be unique opportunities for subscription video-on-demand providers to capitalize on the vodcast boom—most notably by launching companion vodcasts that keep audiences engaged between seasons while expanding their content slate.[28] Streamers might also consider podcasts as low-cost incubators for new stories and emerging talent, with the podcast-to-screen funnel tapping into existing fanbases and reducing development risk.[29] Likewise, partnering with established creators who already command loyal followings and understand how to spark social buzz offers a fast track to cultural relevance. These tactics aren't about chasing the vodcast hype. They transform audio-first storytelling into a strategic engine for retention, deeper fan engagement, and sustainable long-term growth.

BY

Brooke Auxier
United States

Akash Rawat
India

Gillian Crossan
Global

Tim Bottke
Germany

Duncan Stewart
Canada

Wenny Katzenstein
United States

ENDNOTES

1. Based on Deloitte analysis; Brooke Auxier, Bree Matheson, Duncan Stewart & Kevin Westcott, "Shuffle, subscribe, stream: Consumer audio market is expected to amass listeners in 2024, but revenues could remain modest," *Deloitte Insights*, Nov. 29, 2023.
2. Spotify Newsroom, "Spotify unveils uninterrupted video podcasts, audience-driven payments, and the new Spotify for Creators platform, Nov. 13, 2024; Wondery, "Now playing: Video podcasts on the Wondery app for Wondery+ subscribers," accessed Oct. 23, 2025; Angela Yang, "Podcasts are taking over TV screens as video formats grow increasingly popular," *NBC News*, Dec. 23, 2024.
3. Spotify Newsroom, "From audio to video, Spotify's $100 million payout fuels creator success stories," Apr. 28, 2025.
4. Ariel Shapiro, "YouTube launches a dedicated page for podcasts," *The Verge*, Aug. 23, 2022.
5. Todd Spangler, "YouTube says it now has more than 1 billion monthly viewers of podcast content," *Variety*, Feb. 26, 2025; Zach Vallese, "YouTube launches weekly top podcast list to rival Spotify and Apple," *CNBC*, May 15, 2025.
6. Alex Schiffer, "Taylor Swift draws 1.3 million live viewers in 'New Heights' appearance," *Front Office Sports*, Aug. 13, 2025; Vicki Newman, "Taylor Swift earns podcast record with appearance on boyfriend Travis Kelce's New Heights, Guinness World Records," Aug. 26, 2025.
7. Based on Deloitte analysis of publicly available data.
8. Data from Deloitte's Fall 2025 Digital Media Trends 19 survey.
9. Edison Research, "YouTube is the preferred podcast listening service," Oct. 23, 2024.
10. Ellie Hammonds, "Vodcasts: Is it the future of podcasting?" The Media Leader, Aug. 28, 2025.
11. Molly Fuard, "Visual podcasting is now a thing and here's what advertisers should know," *Adweek*, accessed Oct. 23, 2025.
12. Lloyd George, "Why social media is a game-changer for growing your podcast, Acast, accessed Oct. 23, 2025.
13. Alexander Lee, "Podcast consumption shifts towards connected TVs," Digiday, May 7, 2025.
14. Eve Upton-Clark, "Netflix is eyeing video podcasts as it expands beyond TV and film," Fast Company, April 21, 2025.
15. Data from Deloitte's Fall 2025 Digital Media Trends 19 survey.
16. Ibid.
17. Sara Fischer, "Axios media trends," *Axios*, April 22, 2025.
18. Simon Kemp, Digital 2025: The essential guide to the global state of digital," Meltwater, Feb. 5, 2025.
19. Global System for Mobile Communications Association, "The mobile economy 2025," accessed Oct. 23, 2025.
20. Paula Gilbert, "Nigeria's 1GB data price has dropped 75% over five years," Connecting Africa, June 5, 2020; Peter Oluka, "$0.39 [604 NGN] per 1GB: Nigeria among countries with cheapest data rates," Tech Economy, Jan. 10, 2025; Bruno Venditti, "The cost of 1GB of mobile data worldwide," Visual Capitalist, Oct. 21, 2024.
21. Spotify Newsroom, "Get to know the 13 podcast grantees of Spotify's new Africa podcast fund," Oct. 24, 2022; *Blueprint Magazine*, "Spotify and the pod network enters a new era of Filipino podcasting with the launch of their state-of-the-art studio," April 28, 2025.

22. Spotify Newsroom, "The Spotify partner program expands to nine new markets, giving more creators new ways to monetize their content," March 27, 2025.

23. IndustryPods, "Podcast distribution on international platforms," December 2024; The Storiez, "How Anghami is dominating the music streaming market globally," Sept. 14, 2024; Peerzada Abrar, "Kuku FM raises $25 mn from investors; aims to expand content, improve tech," *Business Standard*, Sept. 20, 2023; Samuel Viavonu, "The podcast boom in Nigeria: an era of noise or knowledge?" Afrocritik, Feb. 19, 2025.

24. Acast, "The video podcast opportunity," June 10, 2025; Devan Kaloo and Robert Gilhooly, "Demystifying emerging markets," Aberdeen Investments, Sept. 8, 2023.

25. BeMultilingual, "What are the most popular languages on YouTube?" July 26, 2025; David R. Gonzalez, "The state of podcasting in Latin America," *PodNews*, Feb. 15, 2024.

26. Aaron Chow, "Nike Japan launches 'NIKELAB RADIO'," HypeBeast, July 28, 2021.

27. Burt Helm, "How AI for lip dubbing could change the film industry," Fast Company, November 2023.

28. The New York Times Style Magazine: Australia, "Forensic fandom and the age of the companion podcast," Feb. 27, 2025.

29. Damion Taylor, "How podcasts are becoming Hollywood's new development pipeline," *Forbes*, Jan. 30, 2025.

ACKNOWLEDGMENTS

The authors would like to thank **Matt Varraveto**, **Abel Sun**, and **Kenny Gold** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

A new era of self-reliance: Navigating technology sovereignty

Countries and regional blocs are racing to build out their own sovereign tech and AI infrastructures. What are the implications, and how can global businesses prepare?

ARTICLE • 9-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

As the global geopolitical environment becomes increasingly complex and uncertain, businesses and policymakers are urging their countries and regions to take greater control of their digital infrastructure, especially components related to artificial intelligence. Gartner® estimates that "by 2028, 65% of governments worldwide will introduce some technological sovereignty requirements to improve independence and protect against extraterritorial regulatory interference."[1]

Technology sovereignty is based on the ability of countries and regional blocs to independently develop, control, regulate, and fund digital technologies such as cloud, quantum computing, AI, semiconductors, and digital communication infrastructure.[2] It can include specific geographic, legal, and regulatory requirements around flows of data and where physical facilities are, who owns them, who governs them, who operates them, and who provides the hardware, software, and services that power them.

The desire for sovereignty is not new, but the shift toward technology sovereignty will likely quicken in 2026. Over the next decade, significant investment will flow into cloud computing, semiconductors, data centers, AI models, connectivity, and satellite communications. In an interconnected world, total sovereignty is unlikely to be achieved by any country or region, but many aim to become at least more sovereign.

Since AI is widely regarded as the next major driver of economic development and national competitiveness, its ecosystem is currently getting a lot of attention. This urgency is keenly felt because advanced AI capabilities like computing power (also called "compute") are currently controlled by very few countries and companies.

Research from the Oxford Internet Institute found that "only 34 countries host any public AI compute; only 24 of those have access to training-level compute; and most rely on cloud or chip infrastructure controlled by a small number of foreign actors."[3] The same study found that 90% of all AI compute is managed by US and Chinese companies.[4]

In 2026, Deloitte predicts that more countries will gain greater access to AI compute, and over US$100 billion will be committed to building sovereign AI compute. By 2030, the share of AI compute, managed by companies outside the United States and China, will likely double from its current 10% of global capacity. Signaling this shift, AI and accelerated computing platform company NVIDIA predicts it will sell US$20 billion worth of AI chips for sovereign data center markets in 2025—an increase of 100% year over year.[5]

Greater Europe is leading the drive

In September 2024, the European Union released the "Draghi report," which outlined recommendations for improving overall European economic competitiveness.[6] Part of the report focused on how to potentially advance its domestic tech sector and how the sector could improve innovation, technology adoption, and worker productivity. The report preceded the launch of the EuroStack Initiative—a call from over 200 European companies and organizations for "radical action" around increasing technology sovereignty.[7] This included advocating for buying European, pooling and leveraging existing assets more effectively, focusing less on research and development and more on productization, ensuring adequate capital, and protecting data for European cloud users. Overall efforts of the European Commission are being led by a designated Commissioner for Technology Sovereignty. This continues a long history of the European Union seeking sovereignty in tech, believing that sovereign solutions are best suited for supporting the EU philosophy, values, and principles—embodied in frameworks such as the General Data Protection Regulation, the Digital Services Act, and the AI Act.

Initial fervor and expectations may have moderated somewhat since early 2025, as reflected in the recent EU International Digital Strategy, which is focused more on cooperation with other countries around AI, semiconductors, quantum computing, and cybersecurity.[8] The debate on the best strategic approach to take is ongoing, but there's likely to be over €100 billion in public and private investment for European cloud computing, AI data centers and companies, semiconductors, and satellite communications efforts over the next five years.

Cloud computing

Local European cloud providers comprise a very small percentage (less than 20%) of the overall market.[9] They would require significant investment and time to develop into true competition for global hyperscalers. What's more likely to happen is that global players will increasingly provide European-specific adaptations of their capabilities. Amazon Web Services (AWS) announced that it will invest almost €8 billion in a European Sovereign Cloud located in Germany. Goals for the project include allowing customers to keep their data in the European Union, providing independence, and ensuring it is led, operated, secured, and governed by EU citizens.[10] Microsoft has also announced a set of commitments to Europe—specifically around AI, cybersecurity, privacy, resiliency, and economic competitiveness—and a Microsoft Sovereign Cloud platform and solutions.[11]

AI models and data centers

There are several initiatives from both the government and commercial sectors looking to improve overall AI capabilities. The European Commission's AI Continent Action Plan seeks to develop a series of AI factories and gigafactories across Europe, building on existing supercomputing infrastructure, and driving net-new investment through the InvestAI program.[12] This program will make €20 billion available for up to five new AI gigafactories that will enable the creation of advanced, cutting-edge AI models known as "sovereign frontier models." The Action Plan also looks to improve data availability for AI models, the use of AI applications, and skills and workforce development. On the commercial side, NVIDIA and Perplexity are teaming up to help train and make open-source, localized AI models widely available.[13] NVIDIA is a backer, along with investment firm MGX, Mistral AI, and others, to create Europe's largest AI data center by 2028 at a cost of €8.5 billion.[14] There is also Stargate UK, a phased effort to build out AI infrastructure across the country and accelerate domestic AI adoption.[15]

Semiconductors

Much like the United States, Europe wants to onshore more semiconductor manufacturing, strengthen the resilience of its supply chains, advance a stronger local ecosystem, and boost European companies. To that end, the EU Chips Act (2023) established a fund, pilot lines for experimentation, a collaborative design platform, and competency centers, and provides resources for quantum chips—€43 billion in total investment through 2030.[16] There is already significant commercial investment happening, including a FinFET (field-effect transistor) pure-play foundry, a "Smart Power Fab," and a silicon carbide chip manufacturing plant, among others.[17]

Satellite communications

Another key initiative for Europe is building its own satellite communications constellations to reduce dependence on providers outside the bloc—ensuring secure and reliable services for military, government, and commercial applications. The two main efforts consist of the Infrastructure for Resilience, Interconnectivity and Security by Satellite (IRIS²) constellation and Eutelsat OneWeb. IRIS² will eventually consist of almost 300 satellites in multiple orbits at a cost of about €11 billion.[18] Eutelsat is looking to accelerate its efforts to build out and enhance its OneWeb low earth orbit satellite internet constellation, which currently has more than 630 satellites in orbit.[19] It recently received fresh investment from both the UK and French governments to make this happen.[20] In an increasingly crowded and competitive market, it will take time (IRIS² completion is planned for 2031) and significantly more investment for both of these constellations to reach the point where they can effectively challenge current services and fully support European needs.[21]

What about the rest of the world?

Although Europe is driving a significant amount of technology sovereignty activity, other countries and geographic regions are pursuing their own unique and innovative approaches, with most of the efforts focused on AI. This isn't meant to be comprehensive but rather to show the breadth and depth of global activity.

- **South Korea:** South Korea aims to develop sovereign AI capabilities based on its language and tailored to its culture.[22] One example is Kakao partnering with OpenAI on new personalized digital services.[23] To bolster domestic infrastructure, SK Group and AWS announced that they will jointly build South Korea's largest AI data center by 2029, at an estimated cost of US$5 billion.[24]
- **Japan:** The country is looking to advance its AI capabilities and reinvent its domestic semiconductor industry through the Rapidus initiative—a new company focused on 2-nanometer technology—and a proposed US$65 billion government investment package through 2030.[25]
- **Africa:** Africa's first AI factory, powered by NVIDIA's AI and accelerated computing platform capabilities, will be located in Cassava Technologies' data center facilities in South Africa—with plans to expand to other locations across the continent, including Egypt, Kenya, Morocco, and Nigeria.[26]
- **India:** There is a strong drive for self-reliance across all layers of the tech stack in India, and government programs like the India Semiconductor Mission and IndiaAI are working to address those needs.[27] India will face some unique challenges in developing its AI models, including compute availability, multiple languages to support, and a lack of high-quality training data.[28] India's strong domestic digital capabilities developed for the India Stack could be expanded and exported to other countries to create a new, competitive digital ecosystem.[29]
- **Canada:** The Government of Canada's sovereign AI compute strategy focuses on improving private investment, public infrastructure, and funding access to compute resources.[30] It has also announced a partnership with domestic AI firm Cohere to explore how they can both improve Canada's overall technological capabilities.[31] In addition, several Canadian telecommunications companies are planning to build sovereign AI data centers, including TELUS, SaskTel, and Bell.[32]

- **Middle East:** Many countries in the region are increasing their investment in sovereign cloud and AI data centers—including major projects like the Stargate UAE initiative, a 1 gigawatt AI cluster.[33] The Public Investment Fund of Saudi Arabia established HUMAIN in 2025, a new company looking to develop end-to-end AI infrastructure with the help of global partners like AWS, NVIDIA, and others.[34] US$23 billion in investment has been announced in relation to these partnerships.[35]

Dealing with the consequences

What happens if, as expected, most governments pursue robust technology sovereignty policies and programs in the near future? There are a variety of potential benefits to having greater control over end-to-end technological capabilities. These include economic ones such as greater tax revenue and private capital investment, better employment opportunities for citizens, and a greater chance for homegrown tech companies to flourish. By being more self-reliant, there is a belief that overall resiliency can be improved, privacy and security can be enhanced, and exposure to potential political disruption from foreign countries can be reduced. Additionally, when it comes to AI, if foundational models are created within a country, they can better reflect its local language, customs, and data sets.

We could also see challenges arise, such as:

- **Shifting investment flows.** Foreign direct investment, mergers and acquisitions, and joint ventures could potentially face increasing numbers of conditions and requirements. Venture capital investment could also shift focus. Will venture capital firms put national strategic interests above more global opportunities?[36]
- **Increased fragmentation.** Taking a more insular, zero-sum approach may lead to lower levels of collaboration, fractured international relationships, and fewer academic partnerships. We could also see reduced cross-border flows of data and proprietary communications infrastructure, as well as an increase in the number of standards and regulations.
- **Workforce impacts.** With countries putting greater emphasis on domestic technological capabilities, overall global mobility for highly skilled workers could shift. This could be especially acute in critical areas like AI, cybersecurity, and chip design. There will also likely be greater investment to bolster broader national workforce capabilities.[37]
- **Environmental impacts.** A large increase in the construction of fabs, labs, data centers, and associated supporting infrastructure will put strains on resources. Some countries' power grids are already maxed out, and with new data centers demanding thousands of additional megawatts, they could compete with residential energy needs.[38] There is also the challenge of utilizing non-polluting, low- or zero-carbon electricity sources.[39]
- **New partnerships.** Not everyone can do it alone. In the future, we may see more bilateral agreements, regional frameworks, and nontraditional technology alliances seeking to capitalize on each other's strengths.[40]
- **Overcapacity.** How many foundational AI models can the market support? There are massive global capital expenditures happening just for AI infrastructure—estimated to be almost US$3 trillion through 2028. Will all of it produce a return on investment?[41] Long-term demand may not meet extraordinary expectations, and new technological innovations may lessen the need for current approaches.[42]

The Bottom Line: Prepare for a more self-reliant future

In 2026, expect the drive for technology sovereignty to continue with more debate, government action, and investment activity. While the motivations and eventual outcomes of this drive are open to discussion, action is underway—and more *will* be taken—because many believe that the future prosperity of their countries and regional blocs is at stake.

- **Audit global dependencies.** Identify and assess all critical dependencies—data flows, public cloud, vendors, supply chains, financial, and regulatory. Build new, and strengthen existing, partnerships that could provide the most global flexibility. Be able to transparently explain your global operations.
- **Anticipate regulatory complexity.** Prepare for a fast-evolving regulatory environment. Expect new rules on data localization, cybersecurity, mergers and acquisitions, and capital flows. Pinpoint where your business is most exposed and build scenario plans now. Bolster your compliance programs.
- **Revisit your cloud strategy.** Think about the balance between your public and private cloud capabilities. Adopt a multicloud or sovereign cloud model to enhance resilience and compliance. Prioritize portability, interoperability, and control across environments. Ensure your vendors support automatic compliance for data storage, processing, and transfer. Prepare contingency plans to stay agile in the face of geopolitical shifts.
- **Strengthen talent resilience.** Know where your critical talent comes from—and what happens if access is disrupted. Develop alternative talent-sourcing strategies and leverage government workforce programs and university partnerships to grow specialized skills.

David Jarvis
United States

Duncan Stewart
Canada

Nick Seeber
United Kingdom

Gillian Crossan
Global

Tim Bottke
Germany

Girija Krishnamurthy
Global

ENDNOTES

1. Gartner, "Gartner reveals top technologies shaping government AI adoption," press release, Sept. 9, 2025; Gartner is a registered trademark and service mark of Gartner Inc. and its affiliates in the United States and internationally and is used herein with permission. All rights reserved.

2. Sean Fleming, "What is digital sovereignty and how are countries approaching it?" World Economic Forum, Jan. 10, 2025.

3. Zoe Hawkins, Vili Lehdonvirta, and Boxi Wu, "AI compute sovereignty: Infrastructure control across territories, cloud providers, and accelerators," SSRN, June 24, 2025.

4. Adam Satariano and Paul Mozur, "AI computing power is splitting the world into haves and have-nots," *The New York Times*, June 21, 2025.

5. Yahoo Finance, "NVIDIA Corporation (NVDA) Q2 FY2026 earnings call transcript," Aug. 27, 2025.

6. Mario Draghi, "The Draghi report on EU competitiveness," European Commission, Sept. 9, 2024.

7. EuroStack, "Building Europe's digital future," accessed Oct. 30, 2025; EuroStack, "Open letter: European industry calls for strong commitment to sovereign digital infrastructure," March 14, 2025; Natasha Lomas, "European tech industry coalition calls for 'radical action' on digital sovereignty—starting with buying local," *TechCrunch*, March 16, 2025.

8. European Commission, "The international digital strategy for the European Union," July 8, 2025.

9. Diana Goovaerts, "Europe's cloud market poised for 24% growth," *Fierce Network*, July 28, 2025.

10. Amazon, "AWS plans to invest €7.8 billion into the AWS European Sovereign Cloud," May 15, 2024; Amazon, "Built, operated, controlled, and secured in Europe: AWS unveils new sovereign controls and governance structure for the AWS European Sovereign Cloud," June 3, 2025.

11. Brad Smith, "Microsoft announces new European digital commitments," Microsoft, April 30, 2025; Judson Althoff, "Announcing comprehensive sovereign solutions empowering European organizations," Microsoft, June 16, 2025.

12. European Commission, "Commission sets course for Europe's AI leadership with an ambitious AI Continent Action Plan," press release, April 9, 2025.

13. Belle Lin, "Nvidia and Perplexity team up in European AI push," *The Wall Street Journal*, June 11, 2025.

14. Amiya Johar, "Nvidia, MGX lead €8.5B project to build French AI data center," *Mobile World Live*, May 20, 2025.

15. OpenAI, "Introducing Stargate UK," Sept. 16, 2025; Tom Bristow, "US tech firms pour £30B into UK as Trump lands," Politico, Sept. 16, 2025.

16. European Commission, "European Chips Act: The Chips for Europe Initiative," Nov. 4, 2024; European Commission, "European Chips Act," accessed Oct. 30, 2025.

17. Jingyue Hsiao, "TSMC breaks ground on EUR10 billion semiconductor fab in Dresden," *Digitimes Asia*, Aug. 21, 2024; Infineon, German government issues final funding approval for new Infineon fab in Dresden," press release, May 8, 2025; Adrià Calatayud and Mauro Orru, "Apple supplier STMicroelectronics to build $5.4 billion chip plant in Italy," *The Wall Street Journal*, May 31, 2024.

18. Jeff Foust, "Europe signs contracts for IRIS² constellation," *SpaceNews*, Dec. 16, 2024.

19. Eutelsat, "High-speed, low-latency connectivity," accessed Oct. 30, 2025.

20. Jason Rainbow, "French government to lead Eutelsat's $1.56 billion capital boost," *SpaceNews*, June 19, 2025; Rachel Jewett, UK to join Eutelsat's capital raise with $105M investment," *Via Satellite*, July 10, 2025.

21. Margherita Stancati, Matthew Dalton, and Vera Bergengruen, "Europe scrambles to break its dependence on Musk's satellites," *The Wall Street Journal*, April 13, 2025.

22. Byun Hee-won and Kim Mi-geon, "South Korea to pour $735 bn into developing sovereign AI built on Korean language and data," *The Chosun Daily*, June 17, 2025.

23. Zinnia Lee, "Korea's Kakao teams up with OpenAI to develop AI products," *Forbes*, Feb. 4, 2025.

24. Zinnia Lee, "Billionaire Chey's SK Group partners with Amazon to build a $5 billion AI data center in Korea," *Forbes*, June 23, 2025.

25. Dylan Butts, "Japan is ramping up efforts to revive its once dominant chip industry," *CNBC*, Nov. 13, 2024; Rapidus, "Rapidus Corporation," accessed Oct. 30, 2025.

26. Cassava Technologies, "Cassava to upgrade its data centres with NVIDIA supercomputers to drive Africa's AI future," accessed Oct. 30, 2025; Nell Lewis, "Africa's first 'AI factory' could be a breakthrough for the continent," *CNN*, April 3, 2025.

27. INDIAai | Pillars; Government of India, "India semiconductor mission," accessed Oct. 30, 2025.

28. Shadma Shaikh, "Inside India's scramble for AI independence," *MIT Technology Review*, July 4, 2025.

29. India Stack, "India Stack," accessed Oct. 30, 2025.

30. Government of Canada, "Canadian sovereign AI compute strategy," Oct. 1, 2025.

31. Government of Canada, "Canada partners with Cohere to accelerate world-leading artificial intelligence," press release, Aug. 19, 2025.

32. Telus, "TELUS to launch Canada's leading sovereign AI factory, powered by NVIDIA to drive the nation's AI future," March 19, 2025; Bell, "Increasing sovereign AI capacity: Introducing Bell AI Fabric," May 28, 2025; SaskTel, Deloitte Canada and SaskTel announce strategic alliance to bring Artificial Intelligence (AI) capabilities and solutions to market, advancing Canada's AI vision," press release, Sept. 23, 2025.

33. OpenAI, "Introducing Stargate UAE," May 22, 2025.

34. Amazon, "AWS and HUMAIN announce a more than $5B investment to accelerate AI adoption in Saudi Arabia and globally," May 13, 2025; Nvidia, "HUMAIN and NVIDIA announce strategic partnership to build AI factories of the future in Saudi Arabia," press release, May 13, 2025; PIF, "HRH Crown Prince launches HUMAIN as global AI powerhouse," press release, May 12, 2025.

35. Natasha Turak, "Saudi AI firm Humain is pouring billions into data centers. Will it pay off?" *CNBC*, Aug. 27, 2025.

36. Chris Metinko, "Defense tech venture funding gains traction," *Crunchbase News*, Feb. 12, 2025.

37. 3MTT, "Shaping the future of Nigeria's digital workforce," accessed Oct. 30, 2025; European Commission, "Commission to invest €1.3 billion in artificial intelligence, cybersecurity and digital skills," press release, March 28, 2025.

38. Goldman Sachs, "AI to drive 165% increase in data center power demand by 2030," Feb. 4, 2025; Felicity Barringer, "Thirsty for power and water, AI-crunching data centers sprout across the West," Stanford University, April 8, 2025.

39. Karthik Ramachandran, Duncan Stewart, Kate Hardin, Gillian Crossan, and Ariane Bucaille, "**As generative AI asks for more power, data centers seek more reliable, cleaner energy solutions**," *Deloitte Insights*, Nov. 19, 2024.

40. ECDPM, "**Von der Leyen in India: A tech sovereignty partnership in the making**," Feb. 28, 2025; Nii Simmonds and David Timis, "**How Europe and Africa can unlock tech opportunities through stronger collaboration**," World Economic Forum, Aug. 18, 2025.

41. Rolfe Winkler, Nate Rattner, and Sebastian Herrera, "**Big tech's $400 Billion AI spending spree just got Wall Street's blessing**," *The Wall Street Journal*, July 31, 2025; *Financial Times*, "**What'll happen if we spend nearly $3tn on data centres no one needs?**" July 30, 2025.

42. Caiwei Chen, "**China built hundreds of AI data centers to catch the AI boom. Now many stand unused**," *MIT Technology Review*, March 26, 2025.

ACKNOWLEDGMENTS

The authors would like to thank **Richard Nunan**, **Michael Greco**, **Julia Tavlas**, **Paul Lee**, **Ben Stanton**, **Manel Carpio**, and **Mosche Orth** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

Generative AI video is perfect for social media, but could disrupt social media companies

Approaching Hollywood quality, the latest generative AI video models appear to be supercharging independent video but could provoke a stronger regulatory response against social video platforms

ARTICLE • 6-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

Sasquatch selfie adventures. Dating tips from fairy tale princesses. Fast-breaking news that may or may not be real. Creative and sometimes concerning uses of generative video are populating social media feeds and competing for eyes in the attention economy. Reality, it seems, may be facing even greater competition than it has already.

Generative video could empower independent creators to produce more for less while reinforcing social media platforms' ability to deliver compelling short-form entertainment and gain a greater share of digital advertising. The same capabilities could also overwhelm audiences, erode authenticity, and provoke regulators to try to contain the potential negative side effects of generative video.

When anyone can produce realistic video and publish it to potentially millions as "news," branded content, fan fiction, and much more, or use it to scam, coerce, or deliberately misinform people, the potential for misuse could strengthen the drumbeat of regulators seeking to contain new media.

Deloitte predicts that, in 2026, generative video could provoke a regulatory response in the United States, potentially driving more age verification in more states, refreshing federal challenges to Section 230 protections established in 1996 by the Communications Decency Act,[1] and requiring labeling for AI content published on social platforms. Such regulatory efforts have already begun in some US states, like New York, Tennessee, and Utah.[2] The US Supreme Court has declined to hear objections to an age-verification law for social media use in Mississippi.[3] The European Union's Digital Services Act also includes provisions for "effective age assurance methods."[4] In 2026, a US election year, social platforms may be compelled to use their AI and data capabilities to better manage generative content. Some platforms are already advancing these solutions.[5]

Generative AI will likely enable a glut of video content while also powering better moderation of this content, all at scale. Regulators may look to see how effective these efforts are at managing perceived online harm. Platforms will likely do the same, while also monitoring for reduced engagement, lower monetization, and challenges in meeting compliance.

Some independent content creators are being empowered by generative tools

Generative video models can create short clips of high-quality video and audio that are nearly indistinguishable from "real" content.[6] The relative ease of use and cost-effectiveness is empowering some creatives to run with their creativity, try things with much lower risk of failure, and even rapidly test creative ideas in the hyper-competitive marketplace of social video.

Though they may not be able to deliver 30-minute TV shows or full two-hour movies, generative video tools are very capable of producing compelling, made-for-social video content—like high-profile ads bringing internet memes to life.[7] In fact, the perceived limitations of generative video that may slow Hollywood adoption seem to be empowering some creators and social video platforms, where short-form, fast cuts, and selfies are common to virality, and where audiences may be less discerning about the free, open-source entertainment they receive.

For independent creators—and maybe soon for all media production—generative tools may be less about replacing the entire production stack to render fully synthetic content, and more about eliminating costly micro-tasks, compressing the time to create, and empowering smaller outfits to do more.

Generative AI and video tools are powering cheaper and faster content creation, eliminating more of the micro-tasks in production, distribution, and measurement.[8] This can amplify outputs to help keep up with a fast cadence of publishing, often necessary to engage followers and stand out in the algorithmic feeds of social platforms.

Many tools focus on time- and money-saving shortcuts, like quickly generating videos from scripts and "one-click" clip generation.[9] This can help enable creators to rapidly test variants to determine which approaches work better with specific audiences and trending algorithms. Other tools are enabling creators to generate AI avatars of themselves that can reduce fatigue while still engaging audiences and even enabling greater personalization at scale.[10] The same features are expanding into generative ads.[11]

Generative AI tools can also support non-generative content with faster editing, like removing "ums," silences, and bad takes, fixing shaky cameras, and automating the removal of dead space.[12] Multilingual dubbing tools can open access to foreign language audiences, expanding engagement and ad revenue potential.[13]

With these capabilities, creator studios can more quickly ideate, generate content, target audiences, measure results, and repeat. This could not only disrupt the economics of content but also lead to exponentially more content. A greater supply of content could create more competitive pressure among creators, which could inspire even more creative content.

Generative video could threaten Hollywood and social media platforms alike

As Deloitte showed in its "2025 Media and Entertainment Outlook," some major studios and publishers have been exploring generative video but have been hesitant to integrate it into productions. This caution may come, in part, from a fear of undermining their premium content offerings with synthetic media, but also due to challenges from talent. The SAG/AFTRA strike of 2023 included demands for limiting the use

of generative AI in productions.[14] Yet, Hollywood studios are often overburdened by high production costs and may wish for generative AI to eventually reduce this burden.[15]

At the same time, traditional studios and streamers face greater competition for advertising dollars.[16] While some Hollywood studios work to stem ad losses from a declining linear TV business and migrate their advertising businesses to connected TVs and streaming video services, many social platforms have been taking more digital advertising dollars. Advertiser spending on these platforms is showing significantly greater growth than other digital media, like streaming video services.[17]

Generative AI's ability to both quickly generate content and predict which segments and individuals will engage with it appears to be transforming digital advertising.[18] With simple prompts, social platforms can automatically generate thousands of ads with small variations, and then instantly test which variants perform the best.[19]This is enabling ad buyers to spend less on creating ads and more on testing variants that return the highest success rates, reinforcing the competitive advantage of social platforms.[20]

Yet, social platforms will also likely see greater risks from their own generative video efforts. They may confront a further boom in the amount of video content they should deliver and manage, some of it likely treading into copyright violations or worse categories. They may risk audiences being overwhelmed by "AI slop" and a rapid devaluation of the principal currency that has fed much of social media's growth: authenticity.

A year ago, we asked people in the United States how they felt about generative media.[21] Sixty-four percent agreed that generative AI on social media is dangerous; 76% agree that online content creators should be transparent about when and where they use generative AI in their content; and 53% agreed that online content creators who use generative AI are not authentic.

Now, a year later, the capabilities of generative video and the amount of it on platforms have both grown considerably. Generative video appears to be quickly approaching parity with reality and could soon tread into dangerous territory, potentially to both attract regulators and empower bad actors. There are concerns of potential fraud as models enable bad actors to use AI to impersonate people.[22] Along with sasquatch selfie videos come the likely influence campaigns, scams, political disinformation, and conspiratorial rantings. This could even impact legal proceedings if video evidence becomes untrustworthy. Yet, without regulatory oversight, audience exodus, or punitive damages, platforms are typically unincentivized to rein it in.

The bottom line: Social platforms can protect the truth —and their best interests

Synthetic media, AI slop, and disrupted business models could pale in comparison to the societal challenges when anyone can make and distribute realistic videos, and video evidence is no longer a reliable form of truth. Watching the leading edge of generative videos—especially the ones trying to illustrate the risks of fake news feeds, celebrity sightings, false flags, and political gaffes—it's hard to downplay what could become a wave of disruption that seems to be fast approaching.

To get ahead of these risks, social platforms should work to develop and integrate watermarking, AI labeling, and ways to track and reveal the provenance of all content, including ads, that are uploaded to or generated by their services. Seemingly inevitable political manipulation and consumer deception should move regulators to work with platforms and establish stronger guardrails for generative content, such as requiring labeling and watermarking.

In the United States, Section 230 of the Communications Decency Act, which has protected open platforms from liability for content they host, could be challenged if they don't get ahead of these accelerating risks.[23] European regulators have already shown a strong willingness to regulate US social platforms and data collectors.[24] Developing stronger compliance automation, like compliance agents that

monitor the outputs of other generative tools, could enable platforms to rapidly respond to violations at scale.

For all its connectivity, transparency, and celebration of humanity, social media has also advanced the fragmentation of information, the deregulation of media, and the capabilities of bad actors. Without strong efforts from the social platforms enabling these capabilities, generative video could greatly amplify this condition and further unmoor society from any shared sense of ground truth.

BY

Chris Arkenberg
United States

Gillian Crossan
Global

Tim Bottke
Germany

ENDNOTES

1. U.S. Congress, Senate, "S. 314 – A bill to protect the public from the misuse of the telecommunications network and telecommunications devices and facilities," accessed Oct. 22, 2025.

2. Mayer Brown LLP, "Children's online privacy: recent actions by the states and the FTC," Feb. 25, 2025.

3. Ella Lee, "Supreme Court — Mississippi social-media law and minors' access," *The Hill*, Aug. 14, 2025.

4. European Commission, "Commission press corner detail: IP/25/1820," press release, July 14, 2025.

5. James Beser, "Extending our built-in protections to more teens on YouTube," YouTube News & Events Blog, July 29, 2025.

6. *The New York Times*, "AI video deepfakes – quiz and playground," June 29, 2025.

7. Bill Chappell, "AI video ad, Kalshi advertising NBA finals," NPR, June 23, 2025.

8. Thomas H. Davenport and Nitin Mittal, "How generative AI is changing creative work," *Harvard Business Review*, Nov. 14, 2022.

9. Torin Anderson and Shuo Niu, "Making AI-enhanced videos: Analyzing generative AI use cases in YouTube content creation," In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–7. 2025.

10. Collectively Inc., "How content creators are embracing generative AI and AI avatars: insights from our latest survey," Jan. 14, 2025.

11. Jess Weatherbed, "TikTok ads may soon contain AI-generated avatars of your favorite creators," *The Verge*, June 17, 2024

12. Anderson and Niu, "Making AI-enhanced videos: Analyzing generative AI use cases in YouTube content creation."

13. Yael Malamatinas, "7 of the best AI dubbing tools to translate videos into different languages," Vimeo, blog, April 28, 2025.

14. Screen Actors Guild – American Federation of Television & Radio Artists, "SAG-AFTRA statement on the use of artificial intelligence and digital doubles in media and entertainment," March 17, 2023.

15. Katie Kilkenny, "Higher costs are hitting film and TV producers even as studios keep trimming budgets," *The Hollywood Reporter*, April 17, 2025.

16. Chris Arkenberg, Jeff Loucks, Kevin Westcott, Danny Ledger, and Doug Van Dyke, "2025 media and entertainment outlook," *Deloitte Insights*, April 23, 2025.

17. Interactive Advertising Bureau, "Digital ad revenue surges 15% YoY in 2024, climbing to $259 B," April 17, 2025.

18. Ryan Browne, "AI is disrupting the advertising business in a big way — industry leaders explain how," *CNBC*, June 15, 2025.

19. Charles James, "Generative AI for retail ad campaign variants and A/B testing automation," ResearchGate, Nov. 9, 2024.

20. Interactive Advertising Bureau, "Nearly 90% of advertisers will use Gen AI to build video ads, according to IAB's 2025 video ad spend & strategy full report," July 15, 2025.

21. China Widener, Jana Arbanas, Doug Van Dyke, Chris Arkenberg, Bree Matheson, and Brooke Auxier, "2025 digital media trends: Social platforms are becoming a dominant force in media and entertainment," *Deloitte Insights*, March 25, 2025.

22. Clare Duffy, "OpenAI's Sam Altman warns of an AI 'fraud crisis'," *CNN*, July 22, 2025.

23. Paris Martineau, "Exclusive: Section 230 may finally get changed — lawmakers prep new bill," *The Information*, accessed Oct. 22, 2025.

24. Dawn Carla Nunziato, "The Digital Services Act and the Brussels Effect on platform content moderation," *Chicago Journal of International Law* 24, no. 1 (2024): pp. 1–37.

ACKNOWLEDGMENTS

The authors would like to thank **Rohan Gupta** and **Jana Arbanas** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

Public media partnerships with streaming giants could be a model for making traditional TV sustainable

Public service broadcasters are publishing to social platforms, co-producing with streamers, and forming partnerships with the largest video distributors. They can offer lessons to for-profit US media companies.

ARTICLE • 10-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

In the United States, traditional television continues to be a profitable but declining business. Amid the rise of streaming wars, social video, and interactive entertainment, content catalogs are migrating to streaming video services, rights are being renegotiated, and linear TV businesses are being restructured and sold.[1] Yet many traditional media conglomerates have been acting—and spending—to rebuild the golden age of TV profitability around their own new and expensive streaming video services. So far, that golden age hasn't returned.

Outside the United States, more adaptive models for success are emerging from public service broadcasters (PSBs). With a history of proven storytelling exported to larger audiences, PSBs in Europe are bolstering production and distribution by making co-production deals with streamers. To reach younger audiences, they are publishing and promoting content on social platforms. To expand their reach, they are experimenting with staggered releases between their own services and global partners.[2] These creative strategies can offer valuable lessons for smaller US networks grappling with similar pressures to evolve and stay competitive.

In 2025, there has been an acceleration with three notable deals between broadcasters and streamers in just a few months.[3] Deloitte predicts another handful of broadcaster-and-streamer deals for 2026. Further, we see more co-productions and other initiatives, once again led by the PSBs. This could bring tens of thousands of additional hours of broadcaster content to streaming video services and social platforms, with potential gains in ad revenue shares and global viewing hours.

For now, PSBs are moving faster at these sorts of deals, motivated more by extending their reach than by profit, and streaming deals appear to be an effective pathway. Interestingly, some commercial broadcasters are also striking similar streaming deals, but only time will tell if they are outliers or a sign of things to come for other commercial TV broadcasters. Regardless, the broadcasting world is watching to see if these deals yield "happily ever after" endings.

Public service broadcasters are embracing disruption and finding innovative ways forward

Globally, many PSBs are large cultural institutions that have played an outsized role in representing and reaching the public, and in shaping entertainment, news, education, and culture.[4] Yet, they are also under threat from the same demographic and behavioral changes that have disrupted traditional, linear television.[5]

There are important differences, however, between commercial broadcasters and PSBs. PSBs typically have a mission to produce TV, movies, news, and documentaries that serve the public interest, regardless of their funding model. If domestic viewers aren't tuning into their linear TV channels or streaming services, PSBs may fail to serve the public, intensifying the debate over whether they should keep their funding and preferred access to the audiences.

Dependent on citizens and governments for funding, yet chartered to fulfill a public mission, PSBs can often be underfunded but remain more committed to outreach and culture than to profits. This condition has enabled them to innovate more quickly and flexibly—and even more daringly—than their private counterparts whose risk tolerance is often anchored to profits and shareholders.

The following examples offer a model for other PSBs working to fulfill their mission in the modern media landscape, and for private traditional media companies that may be both blinded by the success of leading streamers and hesitant to wade into more innovative and disruptive opportunities. There may be some perils with the opportunities, too, if relationships with for-profit providers undermine the value and mission of public media.

Reaching younger viewers on YouTube: ARD and ZDF feel the funk

In 2016, German PSBs ARD and ZDF could already see where younger audiences were tuning in—and where they were tuning out. They launched "funk," a bold digital-content initiative to connect with young viewers on their preferred turf—social media. Instead of posting full episodes or short clips of existing TV shows, funk's studios create videos specifically for social platforms like YouTube, Instagram, and TikTok.[6] Funk publishes dozens of original formats designed for digital natives, including snappy explainers, edgy comedy, and short documentaries. For example, a funk science explainer channel, maiLab, became popular by making chemistry and COVID-19 research accessible.

In 2026, nearly 41% of Germans under 30 now watch funk's offerings at least weekly.[7] Within a two-year period, funk content garnered roughly 2.2 billion views on YouTube and 173 million hours of viewing. As young viewers have moved to new platforms featuring short videos, funk has moved with them. TikTok and Instagram now contribute heavily to funk's reach. In fact, in the past two years, funk content logged even more views on TikTok (around 2.3 billion) than on YouTube, reflecting the rise of bite-sized clips.

Connecting with young Germans where they are is important to ARD and ZDF's public mission. By providing professional, publicly funded content in the same spaces dominated by influencers and algorithmic feeds, they offer an alternative to purely commercial social media.

The Canadian Broadcast Corporation's YouTube U-turn

Until recently, the focus of Canada's 90-year-old public broadcaster, the Canadian Broadcasting Corporation (CBC), had been on its own streaming app, CBC Gem. But in 2023, the CBC's digital strategy team led an experiment: They uploaded full episodes and even entire seasons of older CBC shows onto YouTube, treating the platform as a new distribution channel.[8] They adopted a "test-and-learn"

approach, ready to pull content if it siphoned too many viewers from CBC's own services. Far from eroding CBC Gem, YouTube became an *additive* platform—a marketing funnel drawing new, younger viewers to CBC content. Many viewers discovered shows on YouTube then sought out more episodes on CBC Gem, creating a virtuous "flywheel" effect.

The CBC now manages a portfolio of more than 50 YouTube channels, spanning news, comedy, children's programming and more.[9] Short clips like comedy sketches and viral news segments routinely rack up millions of views. The CBC News YouTube channel now boasts over 4.4 million subscribers and 2.6 billion total views.[10] CBC also posts full 20-plus minute episodes of dramas, documentaries and kids' shows that account for nearly half of all viewing time.[11] While quick clips drive clicks, it's full-length shows that keep viewers engaged on the channel.

By the end of 2024, the CBC's experiment on YouTube was gaining viewers. Total watch-time across its channels jumped by 65%, exceeding the 25% growth target the team had set.[12] YouTube has expanded CBC's reach to demographics that traditional TV is challenged to reach effectively. The approach has enabled them to make their content work harder, give new life to back-catalog programming, and create new revenues.

French broadcasters leverage the biggest distributors

In July 2025, PSB France Télévisions struck an "historic distribution agreement" with Amazon's Prime Video.[13] Under the deal, Prime Video subscribers in France can access the live feeds of France Télévisions' channels, and 20,000 titles from their on-demand catalog at no extra cost.[14] The home screen of Prime Video now features a dedicated france.tv section showcasing the broadcaster's content within Amazon's interface.[15] In effect, Amazon's streaming service has become a new virtual cable operator carrying France's public channels.

For France Télévisions, the benefit is greater visibility among younger, cord-cutting audiences. In a fragmented viewing landscape, being present on a popular streamer's menu is a way to stay relevant.

Some European private broadcasters seem to agree. One notable example is France's largest private broadcaster, TF1, which has signed a similar deal with Netflix. Starting in 2026, the partnership—the first of its kind for Netflix anywhere—will let French Netflix subscribers watch TF1's live broadcasts without leaving the Netflix app. [16] The experiment underway in France will be watched closely by media executives across Europe. After all, if you can't beat the biggest streamers, joining them may be the next best thing.

The BBC and Channel 4 find global success with streaming partners

Once upon a time, a BBC or Channel 4 logo on a show meant it was a wholly domestic affair, but today it may also be a collaboration with the likes of Netflix, Amazon Prime Video, or HBO Max.

By tapping streamers' deep pockets, international distribution networks, and appetite for prestigious UK storytelling, co-productions allow PSBs to mount projects that would otherwise strain their finances.[17]As an industry trade group noted, third-party funding (through co-production deals, international pre-sales, tax credits, etc.) now supplies an estimated £400 million a year toward British PSB commissions.[18] In effect, platforms such as Netflix or Amazon may foot a large share of the bill in exchange for rights to stream the finished show globally. The arrangement reduces risks for UK broadcasters and helps ensure that a national hit can reach far beyond dear old Blighty.

The BBC has pursued such alliances aggressively, especially for lavish drama series. *His Dark Materials*, a fantasy epic based on Philip Pullman's novels, was a collaboration between the BBC and HBO that reportedly cost an estimated £50 million for its first series.[19] HBO's cash enabled the BBC to realize a truly cinematic vision—and in return HBO got a ready-made prestige show for the US market.[20] Even quintessentially British period pieces have benefited: The moody post-World War I crime drama *Peaky Blinders* was broadcast on the BBC in the UK with Netflix taking over international distribution, turning a parochial show into a worldwide hit. Similarly, Channel 4 has enjoyed other international successes partnering with global streamers.[21]

By collaborating with Netflix, Amazon, and others, the BBC and Channel 4 are ensuring that British public service content not only survives in the 21st-century mediascape, but thrives—liked, shared and binge-watched around the globe.[22] Like Canada's CBC, Channel 4 has also seen incremental growth across its offerings by publishing full episodes on YouTube.[23] In July 2025, Disney+ and UK broadcaster ITV announced a partnership to give each other's audiences a "taster" of content. Under the deal, ITV's streaming service (ITVX) will host a rotating selection of hit Disney+ titles, while Disney+ in the UK will in turn carry a curated slate of ITV's popular shows. Both sides termed it a mutually beneficial experiment —and an indicator that the streaming wars are shifting toward strategic alliances.

Pitfalls and perils: What public broadcasters risk in these partnerships

Alliances with global platforms offer visibility and funding, but they also pose significant risks. If not carefully managed, partnerships can weaken broadcasters' autonomy, dilute their brand, and undermine their public service mission. As PSBs tread into the terrain of big tech and big money, they should consider the risks:

- **Loss of control and independence.** When distribution and revenue flow through external platforms, PSBs risk becoming captive to algorithm changes or shifting corporate strategies. A single contract reversal could leave them without access to audiences or content rights. If a platform's algorithm decides to demote a broadcaster's videos, the PSB's reach could decline overnight with little recourse.
- **Erosion of direct audience ties.** Audiences who consume PSB content on other services may stop visiting broadcasters' own platforms, reducing brand visibility and access to valuable viewing data. For ad-supported PSBs, this also means reduced monetization potential.
- **Editorial and mission risks.** Chasing clicks, streamer funding, and global appeal can push public broadcasters away from their core remit. The temptation to tailor news or documentaries for algorithms or platform business goals may erode editorial independence and local cultural nuance.
- **Editorial independence and national sovereignty concerns.** Deals with foreign streaming giants have raised political eyebrows, prompted some producers and lawmakers to grumble about a US company gaining influence over a country's public content.[24] Heavy reliance on external commercial funding could also weaken the case for license fee or taxpayer funding.
- **Financial reliance and sustainability challenges.** Heavy dependence on funding from streaming platforms creates a vulnerability if those platforms pivot or pull back. Public broadcasters risk building budgets on unstable ground, even for flagship shows.

While public broadcast services are innovating to keep up with changes in audience behaviors, engagement, and funding, they face new potential risks with streamer partnerships around *control*, *identity*, and *sustainability*. Public broadcasters are, to some extent, trading a measure of independence and direct reach for the short-term gains of money and audience. The *gamble* is that they can manage this trade-off—that they can ride the beast without being thrown off or subsumed by it.

To manage these trade-offs, public broadcasters should:

1. **Protect branding and visibility.** Ensure logos and attribution remain prominent on third-party platforms to sustain trust and recognition.
2. **Secure data and proportional revenue sharing.** Negotiate access to viewing stats and proportional compensation to retain leverage.
3. **Form alliances.** Engage in joint initiatives like the UK's Freely that platform help PSBs stay competitive against global streamers.
4. **Stay true to public-value content.** Keep investing in local news, education, culture, and minority-language programming, even if they aren't global hits.
5. **Innovate with purpose.** Use partnerships to learn from global platforms' technology and apply those lessons to strengthen in-house digital offerings.

Lessons for young US streaming channels

For US media companies working to shift their declining linear TV offerings into competitive streaming services, UK and EU public broadcasters offer ways to be more flexible and innovative. While considering many of the above risks, the PSB journey offers a playbook for US streamers struggling against bigger competitors:

Embrace strategic partnerships to extend reach with legacy and niche content. Rather than cannibalizing viewership or eroding intellectual property (IP), partnering with the largest platforms can revive dormant audiences and bring the brand and IP to new audiences that would never have subscribed to a niche service.

Leverage prized content to anchor valuable partnerships that can expand visibility. Broadcasters have used their local content as a bargaining chip to gain global distribution.[25] Commercial studios can capitalize on prized IP to reach new subscribers. For example, ITV utilized *Love Island* (a local hit) to get *The Bear* on its platform; Disney leveraged *The Mandalorian* fandom to entice ITV viewers to Disney+. Commercial networks might also consider *co-producing* more frequently with global streamers, much as PSBs do.

Guard the brand and data—but be pragmatic. Like PSBs, commercial media companies are learning the need to maintain relevance by following audiences, the need to partner to achieve scale, and the importance of preserving one's identity even while operating on someone else's platform. As linear ratings and ad revenues decline—and many streaming services remain unprofitable—these partnerships may evolve from tactical experiments to core strategy. The experiences of PSBs show that, done right, alliances can be additive and financially savvy.

The bottom line: Key considerations for PSBs and US streamers

Far from being killed off by the streaming revolution and social video, many PSBs are reinventing themselves through it—by pushing content onto social platforms, by co-producing with the biggest streamers, by even letting streamers carry their channels. Done right, this could lead to a richer media ecosystem where public service content coexists with commercial content on every platform, thus injecting some local and ethical balance into global channels.

Although PSBs face significant challenges around fulfilling and defending their public mission in the face of for-profit partnerships, the innovative examples shown here apply equally to second-tier and niche US studios and streamers that are facing the same pressures to adapt. Streaming video has deconstructed and disrupted TV, and social video platforms are drawing audiences away from both TV and streaming

services. The largest video distribution platforms continue to reshape and redefine TV. Public broadcasters and private media alike have little choice but to experiment and adapt.

BY

Jeff Loucks
United States

Chris Arkenberg
United States

Tim Bottke
Germany

Duncan Stewart
Canada

ENDNOTES

1. Chris Arkenberg et al., "2025 media and entertainment outlook," *Deloitte Insights*, April 23, 2025.

2. Ofcom, "Transmission critical: The future of public service media," July 21, 2025.

3. Elsa Keslassy, "How streamers and broadcasters' cross carriage deals could disrupt the TV business in Europe," *Variety*, July 11, 2025.

4. Knight Foundation, "Public broadcasting: Its past and its future," accessed Oct. 29, 2025.

5. Once hugely profitable, in most countries the audience for traditional "linear" TV is getting smaller and older. Globally, multiple public service traditional TV broadcasters (PSBs) are seeing their viewership erode, especially among younger audiences. In the United Kingdom, for example, less than half (48%) of 16 to 24-year-olds watched any broadcast TV in a given week in 2023, down from 76% five years earlier [*CSI*, "Gen Z abandons traditional broadcast TV: Ofcom," July 31, 2024], and only 55% of children between 4 and 15 tuned in weekly, down from 81% in 2018. Still in the United Kingdom, young adults who do watch traditional TV spend barely half an hour per day on it, versus 93 minutes on video-sharing platforms like YouTube and TikTok. Even overall reach is shrinking: the weekly audience for any broadcast TV fell to 75% of Britons in 2023 (down from 79% the year prior)—the steepest decline on record. The same pattern is found in nearly every advanced country's media market and is generally true for both PSB and commercial broadcasters.

6. Funk, "Funk Bericht 2024," Dec. 13, 2024.

7. Ibid.

8. Evan Shapiro and Marion Ranchet, "TESTING & LEARNING: The CBC Case Study," The Media Odyssey, audio podcast episode, April 24, 2025.

9. Ibid.

10. Social Blade, "CBC News YouTube channel statistics," accessed Oct. 29, 2025.

11. Shapiro and Ranchet, "TESTING & LEARNING."

12. Ibid.

13. K.D. with AFP, ""Un accord historique": après TF1 et Netflix, France Télévisions s'associe à Prime Video pour diffuser ses contenus sur Amazon," BFM Tech & Co, July 3, 2025.

14. Ibid.

15. Ibid.

16. AFP, "Netflix breaks new ground with global launch of French TV content," *ForbesIndia.com*, June 19, 2025.

17. Mark Sweney, "BBC and ITV slash big-budget TV spend as US streamers pour money into UK," *The Guardian*, Feb. 16, 2025.

18. Pact, "Submission to Ofcom consultation on the proposals for the new Channel 4 licence," February 2024.

19. BBC, "His Dark Materials: Critics heap praise on 'ravishing' dramatisation," Nov. 4, 2019.

20. Sheena Scott, "'His Dark Materials' is BBC's most expensive series and promises to be A faithful adaptation," *Forbes*, Oct. 31, 2019.

21. BBC, "How The End of the F***ing World became a cult TV phenomenon," Nov. 4, 2019; Daniel D'Addario, "'It's a Sin' is a transporting and tragic tale of the AIDS epidemic: TV review," *Variety*, Feb. 21, 2018.

22. Travis Clark, "8 great Netflix original TV series that show how well its British strategy is working," *Business Insider*, April 2, 2019.

23. John Moulding, "ITV and C4 happy to let viewers watch long-form content on YouTube," *The Media Leader*, March 13, 2025.

24. Max Goldbart, "Streamers Will Not Be Regulated Fully In UK For Another Two Years," *Deadline*, Feb. 26, 2025.

25. Lucas Manfredi, "Netflix, France's TF1 strike landmark distribution deal," *TheWrap*, June 18 2025.

ACKNOWLEDGMENTS

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

Next-gen satellite internet is transforming pricing, capacity, and regulation worldwide

Satellite connectivity sees direct-to-device growth but often faces monetization hurdles, while low Earth orbit data expansion and tech advancements help reshape deployment and resilience, and create regulation complexities

ARTICLE • 11-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

Satellite connectivity appears to be expanding faster than ever. Direct-to-device satellites are often proliferating, but struggling to monetize, while the number of low-Earth-orbit broadband constellations is growing, requiring telecom providers to address opportunities and disruptions. Alongside these developments, technological advancements are reshaping the industry, helping to enable faster deployment, greater resilience, and reduced costs. Regulatory challenges and spectrum management are also emerging as potentially pivotal factors in helping to ensure sustainable growth and integration with terrestrial networks.

Some analysts expect low-Earth-orbit (LEO) satellite constellations to generate around US$15 billion in annual revenues in 2026,[1] and **Deloitte predicts** that global subscribers will surpass 15 million by the year's end.[2] We further predict that the number of communications satellites in LEO will expand to five constellations[3] made up of over 15,000 to 18,000 satellites by the year-end.[4] We further predict that spending on direct-to-device (D2D) satellite capacity will be US$6 to US$8 billion in 2026, with over 1,000 D2D-capable satellites in orbit by the year-end; however, since monetization and business models for D2D are currently unclear, we are not forecasting D2D revenues.

D2D and LEO satellite services are deeply interconnected. At first, starting in 2019, there were multiple satellites launched into LEO, which provided data services to small satellite dishes on Earth, allowing consumers and enterprises to get low-latency broadband services in areas with little or no terrestrial coverage at a reasonable price.[5] However, those signals were not accessible without pizza-sized dishes.[6] In 2023, new satellites, mainly in LEO, with new equipment, new antennas, and using new regulatory permissions, were able to connect directly with devices such as smartphones. Instead of 50 Mbps or more connections over LEO with dish services, D2D in 2023 was low-bit-rate, simple messages.[7] Going forward, D2D may offer higher connection speeds, but still not as fast as dish speeds.[8] LEO and D2D can overlap in terms of the orbits used, and even some of the satellites used,[9] but they are not identical. In September 2025, one major LEO player purchased blocks of 5G spectrum for D2D.[10] It likely won't happen in 2026: New smartphones need new chips to send and receive on that spectrum, new satellites need to be launched to use those bands, and it's unclear what density of simultaneous users can be supported and how well the service will work indoors. But by 2028 or so, smartphone users might be able to stream video directly from space to their phones.[11]

D2D can connect the unconnected, but monetization remains elusive

D2D technology helps enable satellites to directly communicate with standard consumer devices like smartphones, bypassing more traditional ground-based infrastructure, offering essential, typically low-bandwidth connectivity services.[12] This capability can be especially critical in remote or rural areas (including the oceans) that may lack terrestrial cellular coverage.

In 2023, Deloitte predicted the D2D satellite communication market would have approximately US$3 billion in spending on network infrastructure, mainly satellites.[13] Actual spending surpassed predictions, reaching around US$4 billion in 2024.[14] Current company road maps and publicly announced investment plans indicate a total capital requirement of approximately US$6 billion to US$8 billion in 2026. Of this amount, we expect around 85% to 90% to fund new satellite deployments, with the remaining 10% to 15% dedicated to replacing existing satellites.[15] Many handset makers and chip vendors have integrated satellite connectivity into smartphones: Deloitte anticipated that more than 200 million satellite-capable phones will be sold in 2024 (with nearly US$2 billion in specialty chips),[16] and most major smartphone manufacturers have flagship devices that can message via satellite.[17]

Second, a flurry of partnerships between mobile network operators and satellite firms helped expand D2D service availability.[18] These collaborations helped enable basic connectivity (emergency SMS, low-bandwidth data) in underserved regions with no cellular coverage, tapping into a large unconnected population.

In many markets characterized by low gross domestic product per capita, rural and remote regions often remain commercially unattractive for terrestrial telecom operators. These operators typically prioritize urban centers, where higher income levels can support more substantial average revenue per user. In low-income, sparsely populated areas of some countries, revenues for terrestrial cell networks are about 10 times lower per base station while incurring two to three times higher capital and operating costs than in cities.[19] Consequently, many telecom companies tend to avoid investing in infrastructure in these regions.[20] At the end of 2024, an estimated 350 million people (4% of the global population) lived in largely remote areas without mobile internet networks, underscoring the D2D opportunity, although the low incomes in these areas may make many D2D or LEO data services unaffordable to consumers.[21]

Third, global regulators and industry standards bodies have moved quickly to help accommodate non-terrestrial networks, allocating spectrum and finalizing 5G non-terrestrial network specifications, so ordinary phones can seamlessly connect to satellites.[22]

LEO satellite constellations: Rapid expansion, affordable connectivity, and emerging competition

LEO has delivered high-speed, low-latency broadband through a combination of special satellites in special (low) orbits and specialized ground equipment and continues to grow: There are more satellites in LEO every week or so, more constellations of these satellites are being built, more subscribers, and more revenues.

In 2026, Amazon's Kuiper plans to enter the market,[23] potentially competing directly with terrestrial broadband providers in developing markets at lower price points than other LEO solutions. Additional LEO constellations from China's Guowang broadband mega constellation, the Qianfan (G60/Spacesail)

project,[24] Canada's Telesat Lightspeed,[25] and the European IRIS² are either placing satellites in orbit in 2026 or plan to in the coming years.[26] Regional initiatives are also emerging, such as the United Arab Emirates–based Orbitworks venture.[27] Established operators such as Eutelsat OneWeb in Europe are upgrading their constellations to help expand capacity, improve latency, and enable direct-to-device connectivity.[28]

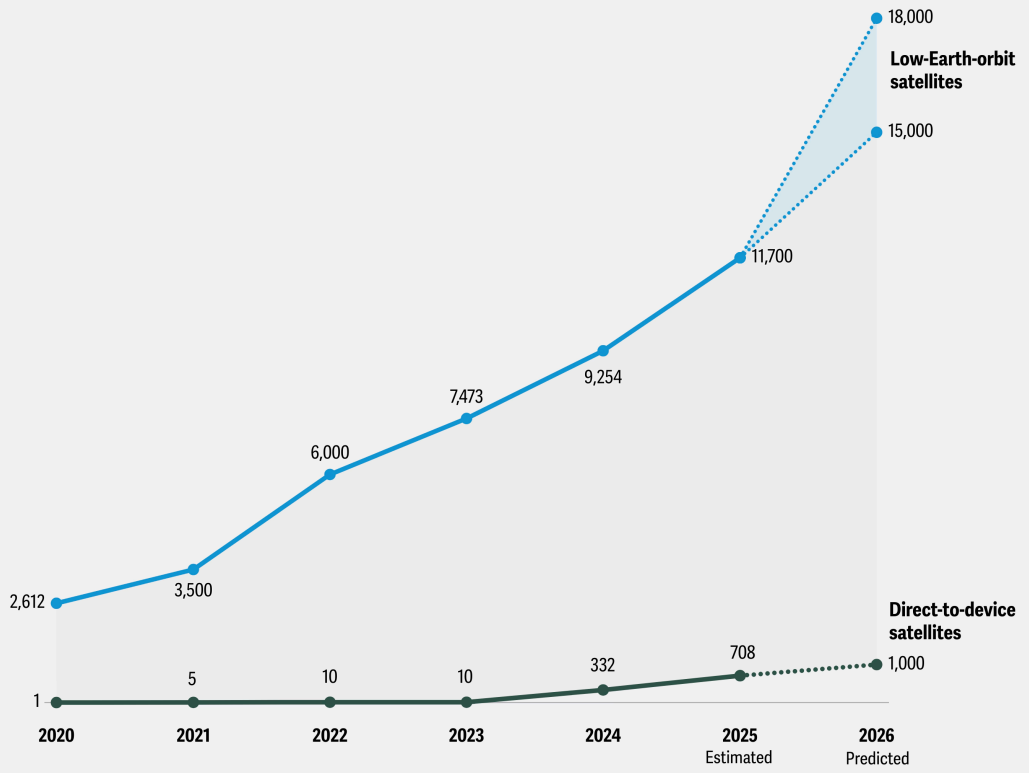
In 2022, Deloitte predicted that, by the end of 2023, more than 5,000 broadband satellites would be in LEO, providing high-speed internet to nearly a million subscribers.[29] We were too conservative; by the end of 2023, there were around 7,473 active broadband-capable LEO satellites.[30]

LEO satellites can offer lower latency and faster connection speeds compared to traditional geostationary satellites, and some portion of the growth in LEO subscribers has come at the expense of **geostationary equatorial orbit** internet providers.[31] LEO primarily targets users lacking traditional terrestrial connectivity options, although, so far, at prices usually higher than equivalent terrestrial broadband services.[32]

The distribution model for LEO satellite services is mixed, with some providers either selling directly to customers, selling through terrestrial telecom partners, or using a hybrid approach. Some may start to compete directly with terrestrial telcos in 2026 by offering more affordable subscription models, particularly in emerging markets.[33]

Figure 1

Estimated low-Earth-orbit and direct-to-device enabled satellites



Source: Deloitte analysis of publicly available satellite industry data and reported deployment milestones for 2020–2025; estimates reflect annual counts of active LEO satellites derived from industry trackers, and D2D-enabled satellites based on disclosed launch activity and service demonstrations.

Deloitte. Insights | deloitteinsights.com

Divergent marketing strategies for providers

As the LEO broadband and D2D satellite markets evolve, Deloitte predicts two distinct distribution strategies will emerge: cooperation and competition. LEO providers frequently partner with local telcos in

certain geographies, and we have seen similar partnerships in D2D in Japan, Australia, and the Philippines, for example.[34]

The coming competition from LEO providers

Alternatively, Deloitte predicts that certain satellite operators will pursue direct competition strategies, particularly in developing regions. These operators will offer services at substantially lower price points than terrestrial providers, aiming to capture underserved market segments through aggressive pricing and simplified service offerings. Although multiple new LEO constellations are planned or under construction, Amazon's Kuiper is expected to be the next to start providing significant service. They are developing a monthly low-cost pricing model.[35] Kuiper plans to target regions with substantial unserved or underserved populations directly, which could present a challenge to terrestrial telcos.[36]

That said, not all terrestrial broadband markets are equally vulnerable to disruption from space-based providers. Average broadband prices in selected developed markets range from US$33 to US$80 per month, while some developing markets have broadband prices under US$10 per month,[37] suggesting that even a relatively low-cost satellite would struggle to garner significant market share in those more affordable markets. Meanwhile, other developing markets such as Brazil or South Africa, with prices in the US$21 to US$48 per month range,[38] may see higher rates of take up, especially if LEO providers price at the lower end, or subsidize pricing in order to gain customers or to provide low-cost connectivity so that consumers can better utilize other services such as shopping or streaming. It should be noted that the ground station terminals needed for LEO cost US$200 to US$500 each,[39] and in the developing world, this relatively high-cost consumer equipment would be unaffordable for many, even with the relatively low monthly costs (approximately US$15),[40] so getting the price of the ground terminals down will also be an important factor.

Transforming satellite capacity is likely essential to unlocking next-gen connectivity

The capacity of individual satellites and the constellations they belong to can play a vital role in ensuring the effectiveness, reliability, and commercial viability of satellite-based communications, including LEO broadband services and D2D basic connectivity. The projected growth in global satellite data traffic is expected to increase 20 times by the end of 2025, presenting significant challenges in terms of satellite capacity.[41] Improved satellite capacity is essential for providing wide-area coverage, supporting high-speed data transmission, and enabling connectivity in remote or underserved regions for multiple simultaneous users: Current networks are often pretty good at connecting a few dozen subscribers in the middle of nowhere, but can struggle in a moderately densely populated area. LEO needs new tech to take it to the next level.[42]

The availability of satellite capacity is influenced by multiple technical and regulatory factors. Technically, capacity depends on the number of satellites deployed, their individual performance capabilities, and their orbital positions.[43] While theoretical models imply that LEO could support up to 10 million to 12 million satellites under ideal conditions,[44] practical constraints derived from collision risks, tracking inaccuracies, and regulatory delays can limit sustainable operations to roughly 100,000 active satellites.[45] Moreover, regulatory requirements, particularly related to spectrum availability in frequency bands like Ku-band and Ka-band, can limit the ability of satellite providers to expand their capacity.[46] For instance, despite its global reach, Starlink has occasionally faced network congestion, leading to temporary disruptions in service availability, underscoring the importance of sufficient capacity.[47] They also have needed to limit the number of subscribers in certain areas, such as southeast England.[48]

Many LEO companies are adopting more sophisticated technological solutions such as adaptive beamforming, dynamic spectrum sharing, inter-satellite links, and artificial intelligence-driven network

optimization.[49] Moreover, investments in more advanced, higher-capacity satellites, improved satellite architectures, and coordinated regulatory efforts for effective spectrum allocation will be important in balancing satellite capacity with increasing user demand and technological advancements.

LEO can be great for those who have no alternative, but it is unlikely to be a material competitor to terrestrial incumbents in most developed world markets. For example, in parts of the United Kingdom, subscriber density is already approaching its limit at approximately 0.35 customers per square km, and one analyst reports that the current Starlink network can support only around 200,000 UK homes (approximately 0.7% market share). The same report suggests that the penetration achievable with Starlink's existing infrastructure ranges from 0.4% in Germany to 1.4% in Spain. Even with a full V2 refresh of their constellation (as opposed to the current mix of V1.5 and V2 satellites), UK penetration would increase only modestly to around 1.4% (with a stretch goal of 3% to 4%, given the full proposed 15,000 satellite constellations). While a future constellation of V3 satellites might achieve a penetration of 8% to 10%, this would likely require over a decade and substantial technical progress. There are no V3s in orbit as of August 2025.[50]

One other factor necessary for growing LEO and D2D overall capacity, and capacity within a given area, is ground stations, also called gateways. Ground stations are a critical part of the infrastructure, relaying data between large data centers and the satellites, managing the satellite network, and sending signals to the satellites. There are already over 100 ground stations for LEO in 2025, and a hundred more will be needed to support multiple constellations.[51] Although many LEO satellites are starting to use laser communications so that satellites in orbit can communicate with other satellites in orbit (rather than having to relay everything through ground stations), this is not expected to eliminate the need for more ground stations.[52] Finally, ground stations should be connected to data centers over fiber to help maximize capacity and minimize latency, which could be a revenue opportunity for terrestrial fiber providers, usually telcos.

Navigating regulatory considerations in spectrum management

As satellite communication markets grow, regulatory considerations around spectrum allocation will become increasingly important.

Deloitte predicts that LEO satellite networks offering D2D services will face significant regulatory challenges, primarily due to their need to operate within frequency bands already allocated to terrestrial cellular services. These complexities are particularly pronounced in regions such as the United States and Europe, where national and regional authorities strictly regulate cellular frequency allocations to help prevent interference and ensure equitable spectrum usage.[53] In the United States, the Federal Communications Commission implemented initiatives like the "Supplemental Coverage from Space" framework, designed to integrate satellite operators with terrestrial networks, facilitating D2D connectivity.[54] Additionally, the National Telecommunications and Information Administration's policy notice for the US$42.5-billion Broadband Equity, Access, and Deployment program represents a shift that expands funding opportunities for LEO satellite providers.[55] The tech-neutral approach eliminates fiber preference and establishes performance-based criteria that put LEO satellites on equal competitive footing with traditional broadband technologies, potentially increasing LEO funding to US$10 billion to US$20 billion from approximately US$4 billion.[56]

In Europe, regulatory management is fragmented, with each national regulator overseeing spectrum allocations within frameworks established by the European Union and the European Conference of Postal and Telecommunications Administrations (CEPT).[57] CEPT is actively evaluating the technical and regulatory challenges of integrating satellite services with terrestrial mobile networks.[58]

In Asia, similar regulatory dynamics exist but are even more complex due to diverse national policies and differing stages of infrastructure development. Countries like India, China, and Japan are actively assessing regulatory frameworks to help harmonize terrestrial and satellite frequency use, ensuring interference-free coexistence while fostering innovation and competition. India, for example, is working through the Telecom Regulatory Authority of India to outline comprehensive guidelines for effectively managing spectrum allocations.[59] In China, significant regulatory reforms are being implemented to accommodate satellite communications. The Ministry of Industry and Information Technology (MIIT) has proactively developed policies to help streamline frequency allocations, manage spectrum interference, and encourage innovation in satellite communications. Recent initiatives by the MIIT include comprehensive frameworks aimed at facilitating the integration of satellite services with terrestrial mobile infrastructure and supporting China's strategic objective of achieving widespread digital connectivity.[60] Similarly, Japan is refining its regulatory framework through the Ministry of Internal Affairs and Communications.[61]

The bottom line: Capex, regulation, and market dynamics

One implication of growth in D2D and LEO to consider concerns the capital expenditures for both space companies and terrestrial connectivity providers. Deloitte predicts that, by the end of 2026, the cumulative investment in D2D satellites and in LEO broadband constellations will reach approximately US$10 billion[62]—and some of those constellations will have D2D capability on some of their satellites. That US$10 billion has been spread over multiple years since 2019, but even if all of it had been spent in a single year, it's startlingly small compared to annual global telco capex, which is about US$300 billion per year, as of 2025.[63]

One reason D2D and LEO partnerships matter for many terrestrial telcos is that they are "capex-lite" ways of meeting the ongoing pressure to connect 100% of populations, no matter how remote or rural. Serving those populations with wired or wireless technologies would cost orders of magnitude more than partnering with space-based solution providers (which have no capex requirement) or even investing in

them directly. AST SpaceMobile has raised money from global players such as Vodafone, AT&T, and Verizon, for example, and the total amount raised is a tiny fraction of those companies' annual capex.[64]

As constellations age, and with the average LEO satellites having a four- to five-year life expectancy, that capex will likely stay high over time, with 20% to 25% of the constellation needing to be replaced annually.[65]

Satellite-based broadband is emerging as a strong alternative to certain traditional terrestrial services, especially in developing regions. For instance, in Nigeria, a LEO satellite provider is now the second-largest internet service provider just two years after entering the market.[66] It seems possible that a single LEO provider, or more likely LEO providers as a group, could become the largest provider in many emerging markets with current low levels of terrestrial broadband connectivity.

Regulatory landscapes will likely evolve significantly, with governments balancing innovation and market competition against national security and sovereignty concerns. New international regulations and standards for spectrum allocation, orbital debris management, and cybersecurity will likely emerge to help address the complex challenges arising from the rapidly expanding LEO environment. Public Emergency Communications System and Emergency Services and Public Safety Requirements exist and vary from country to country. The device makers will have to follow various emergency communications and public safety regulatory frameworks in different countries.

LEO operators will need to navigate complex agreements with terrestrial mobile network operators, employing strategies such as spectrum-sharing or leasing arrangements under conditions designed to prevent interference. Critical regulatory concerns include sophisticated interference management strategies like dynamic spectrum allocation and geographic beam shaping.[67] Balancing terrestrial operator rights with satellite-enabled connectivity enhancements will likely require extensive regulatory oversight and robust collaborative models between satellite operators and terrestrial providers.

Although we focus on the consumer LEO broadband market, a large and robust enterprise market is likely to emerge over the next few years, with the number of enterprise subscribers growing nearly tenfold by 2030 to 3.4 million.[68] While that is a smaller number of subscribers than the consumer market today, enterprise customers are likely to have much higher monthly revenues and lower churn than consumers.

BY

Prashant Raman
India

Duncan Stewart
Canada

Gillian Crossan
Global

Tim Bottke
Germany

Jody McDermott
Canada

Ben Stanton
United Kingdom

ENDNOTES

1. Gartner, "**Gartner forecasts LEO satellite communications services spending to hit $14.8bn globally in 2026**," press release, July 30, 2025.
2. Deloitte analysis of publicly available market research and forecasts, combining current adoption trends, planned service launches, and demand in underserved regions to assess the feasibility of future subscriber growth.
3. Deloitte analysis of global low Earth orbit (LEO) satellite deployment trends indicates five major constellations—Starlink, Kuiper, Guowang, Honghu-3, and G60—will account for a significant proportion of the estimated 15,000 to 18,000 LEO satellites expected in orbit by the end of 2026. This projection aggregates operator-specific deployment targets, launch rate trends, and industry growth forecasts, referencing broker research and company filings.
4. This is based on a Deloitte analysis of publicly available industry data and forecasts, including current deployments as of mid-2025, announced launch schedules from major operators, and long-term projections from leading research providers. Estimates were derived by combining existing satellite counts with confirmed launch plans and aligning them with independent analysts' projections.
5. Yarnaphat Shaengchart and Tanpat Kraiwanit, "**Starlink satellite project impact on the Internet provider service in emerging economies**," Research in Globalization, May 4, 2023.
6. Nick Cowell, "**Satellite-based internet connectivity LEO Satellite Broadband**," Fujitsu, May 22, 2023.
7. Karen L. Jones and Audrey L. Allison, "**The great convergence and the future of satellite-enabled direct-to-device**," Center for Space Policy And Strategy, September 2023.
8. Joe Madden, "**The difference between NTN/D2D and satellite broadband – Madden**," Fierce Network, Jan. 16, 2024.
9. Christopher Baugh, "**Satellite direct-to-device: The characteristics of D2D constellations will limit SpaceX's ability to dominate**," Analysys Mason, July 22, 2024.
10. Mike Robuck, "**Musk outlines SpaceX D2D spectrum strategy**," Mobile World Live, Sept. 10, 2025.
11. Ibid.
12. These services include emergency messaging, basic data transmission, and sometimes voice calls.
13. David Jarvis, Duncan Stewart, Raghavan Alevoor, and Kevin Westcott, "**Signals from space: Direct-to-device satellite phone connectivity boosts coverage**," *Deloitte Insights*, Nov. 29, 2023.
14. Deloitte analysis of publicly available data on satellite industry investments for 2023–2024; investment amounts reflect disclosed funding rounds, commercial agreements, and reported capital commitments related to direct-to-device satellite communication.
15. Deloitte analysis of global direct-to-device satellite communication capital requirements, based on company filings, investor presentations, earnings call transcripts, government announcements, press releases, research reports, and expert interviews.
16. David Jarvis, Duncan Stewart, Raghavan Alevoor, and Kevin Westcott, "**Signals from space.**"
17. Aamir Siddiqui and Andrew Grush, "**Android and iPhone satellite connectivity: What is it and what are your options right now?**" Android Authority, Feb. 11, 2025.
18. Arun Menon, "**Satellite industry trends to watch in 2024**," TM Forum, Jan. 31, 2024.
19. GSMA, "**Open consultation for the council working group on international internet related public policy issues**," August 2020.

20. Ibid.

21. GSMA, "New GSMA report shows mobile internet connectivity continues to grow globally but barriers for 3.45 billion unconnected people remain," press release, Oct. 23, 2024.

22. 5G Americas, "New developments and advances in 5G and NT," February 2025.

23. Amber Jackson, "Project Kuiper explained: Australia's bid to improve internet access with Amazon," Capacity, Aug. 5, 2025.

24. Ling Xin and Victoria Bela, "China launches first satellites for GuoWang project to rival SpaceX's Starlink," *South China Morning Post*, Dec. 16, 2024.

25. Mark Holmes, "Telesat's Lightspeed is now fully funded, MDA to build constellation," *Via Satellite*, Aug. 11, 2023.

26. Connectivity and Secure Communications, "ESA confirms kick-start of IRIS² with European Commission and SpaceRISE," Dec. 16, 2024.

27. *SatNews*, "Loft Orbital and Marlan Space to create the Middle East's first private manufacturing space company of commercial satellite constellations for LEO," Aug. 26, 2024.

28. *Reuters*, "Eutelsat announces contract with Airbus for 100 satellites," Dec. 17, 2024.

29. David Jarvis, Duncan Stewart, Kevin Westcott, and Ariane Bucaille, "Too congested before we're connected? Broadband satellites will need to navigate a crowded sky," *Deloitte Insights*, Nov. 30, 2022.

30. CCIA, "Low earth orbit (LEO) satellite broadband facts and stats," March 5, 2025.

31. Rick Mur, "Low-earth orbit (LEO) networks in your global connectivity strategy," GNX, Jan. 22, 2025.

32. Ibid.

33. Garinder Shankrowalia, "Amazon's ambitions: Project Kuiper and the complex future of satellite broadband," Omdia, May 20, 2025.

34. Rakuten.Today, "Moshi moshi? Space calling: Rakuten Mobile and AST SpaceMobile achieve Japan first satellite-to-mobile video call," May 2, 2025; Cameron Page, "Australia's TPG completes first D2D satellite trials with Lynk," TelcoTitans, May 8, 2025; John Tanner, "Globe kicks off Lynk Global D2D SMS tests in Zambales," Developing Telecoms, Oct. 7, 2024.

35. Nadine Hawkins, "Amazon to launch Project Kuiper satellites next week," Capacity Media, April 3, 2025.

36. Amazon, "Here's how Project Kuiper's satellite network can help telecom partners like Vodafone and Vodacom enhance reliability and extend reach," Sept. 5, 2023,

37. World Population Review, "Internet cost by country 2025," accessed Oct. 30, 2025.

38. Ibid.

39. Jack Kuhr, "Starlink Mini Impact and Rapid Terminal Iteration: Payload Research," Payload, June 26, 2024.

40. Hawkins, "Amazon to launch Project Kuiper satellites next week."

41. Elton Chang, "Satellite network capacity and scalability," TelecomWorld101, Jan. 17, 2025.

42. Ibid.

43. Kim Larsen, "The next frontier: LEO satellites for internet services," techneconomyblog, March 12, 2024.

44. Andrea D'Ambrosio, Miles Lifson, and Richard Linares, "The capacity of low earth orbit computed using source-sink modeling," *arxiv*, June 10, 2022.

45. Harry Baker, "How many satellites could fit in earth orbit? And how many do we really need?" LiveScience, May 30, 2025.

46. Kelly Hill, "FCC revisits satellite spectrum power levels," *RCR Wireless News*, May 1, 2025.

47. Dan Heming, "Starlink waitlists return, network congestion on the rise and finally, a customer support phone #," Mobile Internet Resource Center, Nov. 21, 2024.

48. Mark Jackson, "Starlink's satellite broadband hits capacity limit in South East England," ISPreview, Dec. 31, 2024.

49. Marcin Frąckiewicz, "Artificial intelligence in satellite and space systems," Tech Stock 2, June 12, 2025. Luis Manuel Garcés-Socarrás et al., "Artificial Intelligence implementation of onboard flexible payload and adaptive beamforming using commercial off-the-shelf devices," *arXiv*, May 3, 2025.

50. James Ratzer, "Starlink: What impact might it have on the telcos?" New Street Research, June 9, 2025.

51. Stella Linkson, "Starlink ground stations: What they are and how they work," Starlink Info, March 21, 2025; Shankrowalia, "Amazon's ambitions."

52. Linkson, "Starlink ground stations: What they are and how they work."

53. Kim Larsen, "Will LEO satellite direct-to-cell networks make terrestrial networks obsolete?" techneconomyblog, January 20, 2025.

54. Federal Communications Commission, "FCC advances supplemental coverage from space framework," March 15, 2024.

55. K. C. Halm, John C. Nelson Jr., and Kasey McGee, "NTIA revamps federally funded $42.5 billion broadband deployment subsidy program," Davis Wright Tremaine LLP, June 12, 2025.

56. David Shepardson, "US Senate panel advances Trump nominee to oversee $42-billion government internet fund," *Reuters*, April 9, 2025.

57. European Conference of Postal and Telecommunications Administrations, "An introduction to the European regulatory environment for radio equipment and spectrum," Feb. 5, 2024.

58. Commission for Communications Regulation, "Radio spectrum management operating plan for 2025–2028," Dec. 13, 2024.

59. ITU-APT Foundation of India, "Recommendations on telecom regulatory authority of India consultation paper on assignment of spectrum for space-based communication services," April 6, 2025.

60. Cetecom Advanced, "China introduces first regulatory framework for radar radio management," March 24, 2025.

61. Ministry of Internal Affairs and Communications, "Progress on the WX promotion strategy action plan," May 29, 2025.

62. Deloitte analysis of publicly reported or analyst-modelled 2026 investment details of major companies in the LEO space.

63. Peter Chahal, Avinash Naga, Courtney Munroe, Bruno Teyton, and Nikhil Batra, "Worldwide telecommunications capex forecast, 2025–2029," IDC Research, June 2025.

64. Newsroom, "AST SpaceMobile secures strategic investment from AT&T, Google and Vodafone," *Business Wire*, Jan. 18, 2024; Hema Kadia, "Verizon's $100 million investment in AST SpaceMobile for satellite connectivity," TeckNexus, May 29, 2024.

65. Inside GNSS, "The case for LEO GNSS at C-Band," Feb. 3, 2025.

66. Damilare Dosunmu, "How Starlink took over Africa's largest internet market," Rest of world, April 15, 2025.

67. Larsen, "Will LEO satellite direct-to-cell networks make terrestrial networks obsolete?"

68. Pablo Tomasi, "Space to grow: Enterprise LEO forecast 2025–30," Omdia, Sept. 9, 2025.

ACKNOWLEDGMENTS

The authors would like to thank **Santosh Anoo**, **Dan Littmann**, **Jack Fritz**, **Sai Tarun Dronamraju**, **Girija Krishnamurthy**, **Michael Greco**, **Kathryn Walby**, **Paul Lee**, and **David Jarvis** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

Deloitte.
Insights

Gifts beat gigabits: Some mobile users rank rewards over network upgrades

Some consumers in developed markets struggle to perceive improvements in network performance. Telecom companies should consider more creative offerings to increase market share.

ARTICLE • 8-MIN READ • 18 NOVEMBER 2025 • Deloitte Center for Technology Media & Telecommunications

Deloitte predicts that in 2026, mobile operator reward schemes may matter to mainstream consumers in developed markets as much as, or even more than, network performance. Over the remainder of the decade, as network upgrades continue, non-network benefits may become increasingly critical to attract users or suppress churn: A slice of margherita pizza may hold more allure than a slice of stand-alone 5G (the more complete version of the 5G standard).[1] The former is tangible, and the latter often beyond the understanding of mainstream consumers.

This trend toward rewards appears to reflect the growing maturity of mobile networks in developed markets. Demand, particularly from the perspectives of network speed and latency (the speed at which a network responds), is largely satiated. Coverage is typically imperfect—there are not spots (no coverage) and overly busy hot spots (too many users relative to available capacity)—but comparing coverage between network operators is often too challenging a chore for consumers who may lack the tools, understanding, and patience to contrast thoroughly.

As a result, network upgrades that are marketed for their higher downlink or uplink speeds, or improved latency, may have diminishing impact on loyalty to a network, as many users can neither perceive nor value such upgrades. Similarly, while important, users may struggle to comprehend the benefit to them of sunsetting 2G and 3G networks and reallocating spectrum to 4G and 5G.

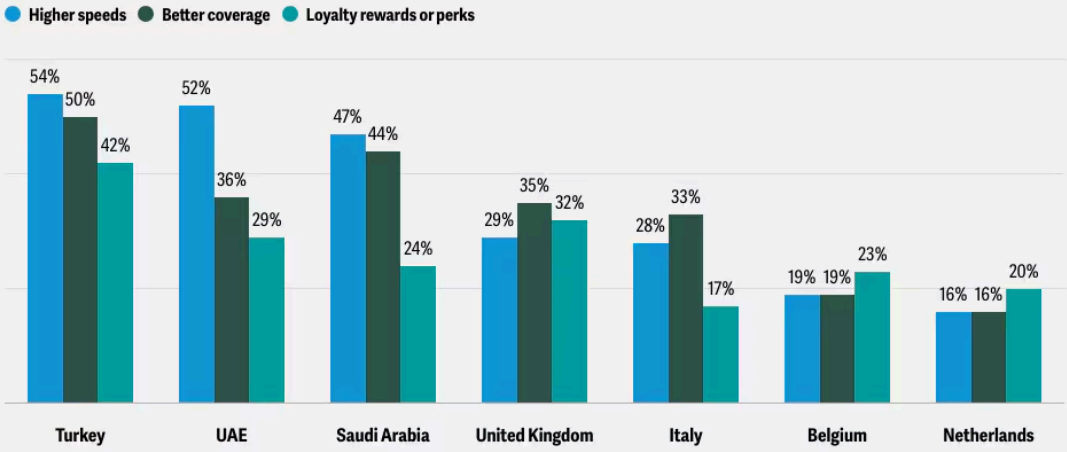
The shift from network upgrades to rewards-based differentiation

Deloitte's view is that each market is likely to be at different points in the journey to rewards-based differentiation (figure 1). But most are likely heading in the same direction. As of 2024, rewards were reportedly the No. 1 factor that could cause churn in the Netherlands and Belgium, and No. 2 in the United Kingdom (note that pricing was excluded as a factor, as would typically be the leading claimed factor). In other markets, however, higher speeds or better coverage were more important.[2] Over the medium term (through 2030), Deloitte predicts that non-network differentiation via offerings such as rewards is likely to become increasingly important.

Figure 1

Some consumers in multiple markets may switch carriers for rewards

Factors that would encourage surveyed consumers to switch mobile networks, multiple markets, percentage of those who chose a given factor, 2024



Notes: Question: Which, if any, of the following would encourage you to switch mobile network provider? Weighted base: all respondents who have a phone or smartphone; aged 18 to 75: UK (3,866), Netherlands (1,944), Italy (1,913), Belgium (978), Turkey (973); aged 18 to 50: UAE (915), Saudi Arabia (874).

Source: Deloitte Digital Consumer Trends 2024.

Deloitte. Insights | deloitteinsights.com

At some point, network upgrades may exceed need, and all carriers in a market may offer what users perceive as roughly equivalent network performance. This is a contrast to the historical situation that had prevailed from the late 1970s in which almost every generational upgrade was meaningful and evident.[3] For example, in the early 2010s, the 4G upgrade delivered an instantly notable performance improvement relative to any 3G network.[4] The technology unlocked what consumers equated to "Wi-Fi like" speeds and latency (response times) when out and about, and applications like search or navigation that faltered on 3G could thrive on 4G.[5]

As of late 2025, however, there are few if any mainstream applications that only work on a 5G network.[6] As such there may be far less motivation to switch to another network with a claimed superior 5G network than was the case with 4G. Some year-over-year improvements may be imperceptible. For example, latency on UK mobile networks over the 2024 to 2025 period showed a 0.7 millisecond improvement (decline) to 18.2 milliseconds.[7] (A millisecond is one thousandth of a second.) A 0.7 millisecond variation is not discernible by a human; even elite athletes react in about 140 milliseconds.[8] Further, almost no mainstream application is likely to benefit from it (latency of 150 milliseconds on a voice call is barely noticeable).[9] Real-time applications such as Voice over IP, for example, need 100 millisecond latency; in the United Kingdom, the slowest-performing network technology, 3G, had 42.3 millisecond average latency in 2025.[10] Deloitte UK's research from that year found that two-thirds of surveyed mobile customers in the United Kingdom noticed no difference in their network over the prior year (figure 2).

Figure 2

While networks improved, less than one in three surveyed users noticed

Two in three (66%) of surveyed mobile customers noticed no difference in their network in the past 12 months

Notes: Question: Over the past 12 months, would you say that the quality of your mobile internet service has got better or worse, or has it stayed about the same? Weighted base: all respondents who have a phone or smartphone, aged 16 to 75, UK (4,023). UK data, additional country data to be added in as it becomes available.

Source: Deloitte Digital Consumer Trends, 2025.

Deloitte. Insights | deloitteinsights.com

Additionally, some consumers may struggle to compare mobile networks in their market, and this may blunt the impact of marketing campaigns urging consumers to switch to a better-performing network.

Most users' network usage is unique, with differing travel patterns and different preferred applications. Coverage maps exist for mobile coverage, but they do not reflect intensity of demand in each location at each point in time.[11] A user could compare two networks side by side by maintaining two SIMs, but this would likely be too tedious a task for most users.

Peak personal connectivity may nearly be here

It has taken more than four decades to satiate demand, but the transformation of consumer connectivity may be nearing completion.[12]

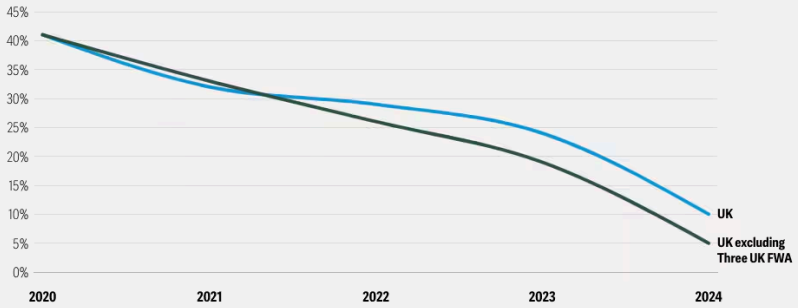
While a prediction should never say never, there is a reasonable probability that no further new fundamentally revolutionary devices that connect to a mobile network will emerge in the medium term (the next five years, through 2030). Similarly, there may not be any transformative applications running over these networks—a mainstream migration to a metaverse could be possible, albeit improbable. And finally, the connectivity demands per major application may remain steady or decline.[13]

The stability and predictability of usage patterns appear to be signaled by data usage trends. Over the past five years, in many major markets, the rate of growth in gigabytes (GB) per SIM has declined. In 10 developed markets, the rate of growth had fallen to single-digit levels by 2024;[14] where growth is at double-digit levels, this is often attributable to a modest growth in cellular mobile connections being used for home broadband, either via a dedicated fixed wireless access (FWA) device or via smartphone tethering. For example, in the United Kingdom, year-over-year increases in mobile data consumption declined to 10% by 2024; however, when excluding the impact of dedicated FWA, growth declined to 5% by 2024 (figure 3).

Figure 3

UK growth in bandwidth consumption has significantly declined, a trend mirrored in other markets

Year on year increases in mobile data consumption, UK market, GB per month



Source: Deloitte analysis based on Enders Analysis.

Deloitte Insights | deloitteinsights.com

Differentiation in the advent of 6G

If consumers struggle to perceive the benefits of 5G, then marketing some elements of 6G may be even harder. The general rule of thumb for every new mobile network generation is a 10-fold (or better) improvement in performance.[15] This would include capacity, which may be needed in specific places at specific times (such as the largest music festivals or the busiest shopping seasons). It may also reduce the cost per gigabyte (GB) carried, as has been the case with 5G versus 4G.[16] But it would also include factors such as higher speeds. The specification for 6G may be finalized in 2026, but there have already been tests of the technology that have demonstrated speeds of 100 gigabits per second (Gbps),[17] which is about 20 times faster than 5G's peak speed of about 5 Gbps (this is the total speed per cell, which would then be shared among users within it).

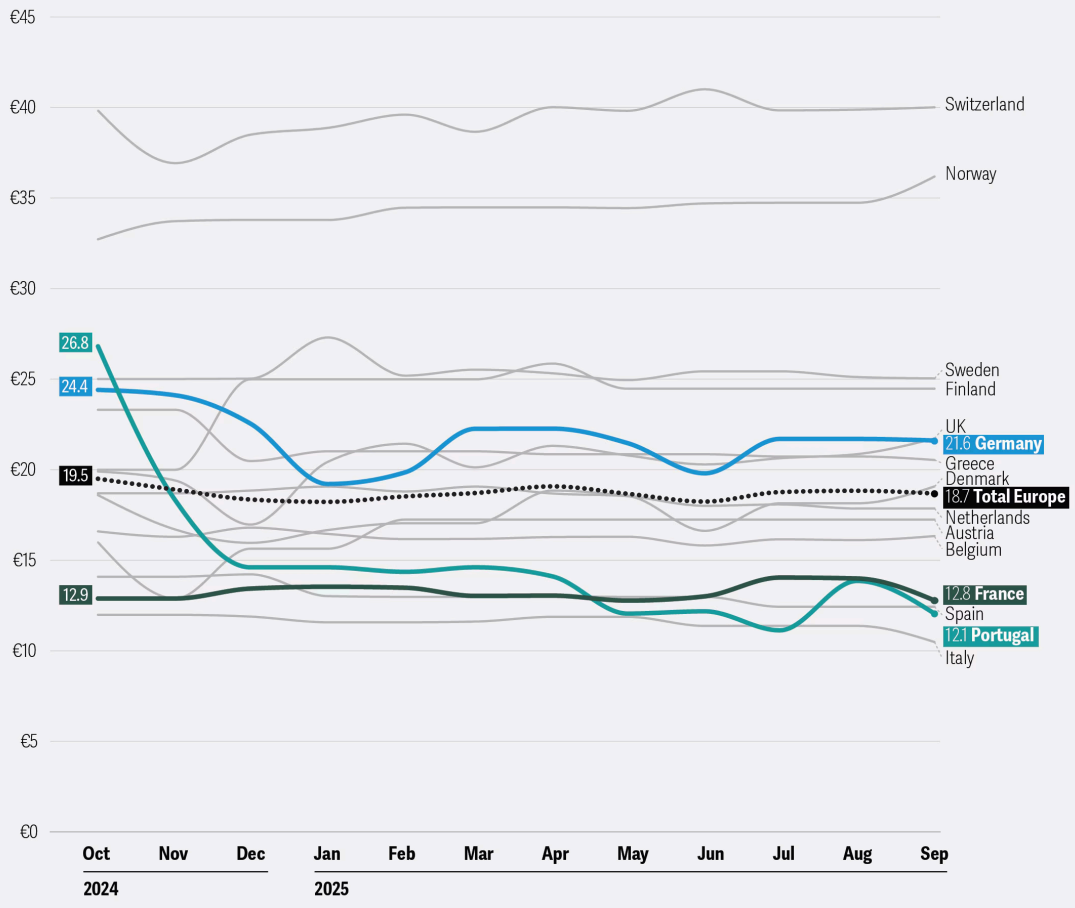
While 6G may offer higher peak speeds, demand may remain static. A typical high-definition video stream delivered to a smartphone often requires under 5 megabits per second (Mbps) per connection. Over the coming years this may well remain constant, or more likely, decline further, as compression and other factors reduce the average bit rate. If demand remains static, then the return on capital from an *extensive* network upgrade to 6G may be challenging, unless the primary intent of an upgrade is to reduce operational costs.

An additional reason for upping the focus on rewards may be to help lessen comparisons that are focused primarily on cost (also known in the industry as a tariff) per a given bundle; for example, 10 GB per month. Between 2024 and 2025, prices for mobile declined by up to 50% in some markets (figure 4).[18] If a bundle also includes elements such as complimentary coffee and pizza, this may make like-for-like comparison less probable.

Figure 4

Mobile prices in some markets are experiencing greater year on year declines

Average mobile tariff price (EUR/month)



Source: New Street Research, 2025.

Deloitte. Insights | deloitteinsights.com

Bottom line: Loyalty rewards may be a promising path forward

Leaders at carriers should consider how their networks are going to be selected in the future and ask if this reflects a major variance to the past. And if so, they should adjust for it. Capital allocation should always matter, and the next decade might look very different for telecom companies. Right now, return on invested capital is 7.3%, but weighted average cost of capital is 6.9%.[19] So telecom investment just about breaks even in economic terms. A slowdown in data usage across fixed and mobile may therefore be a blessing, allowing telecom companies to forgo significant spending on network upgrades for propositions that may deliver greater return on capital.

Telecom companies should note that many other industries have embraced rewards as a differentiator as their core offering has matured. The airline industry—which at one point in time regarded supersonic speeds as a value add—appears to have pivoted substantially to rewards as a sales tool. Airline loyalty schemes have been valued at more than US$100 billion, with just three airlines' schemes being value at more than US$20 billion.[20] In the United States, more than 90% of general credit card spending since 2019 has been on a reward credit card.[21]

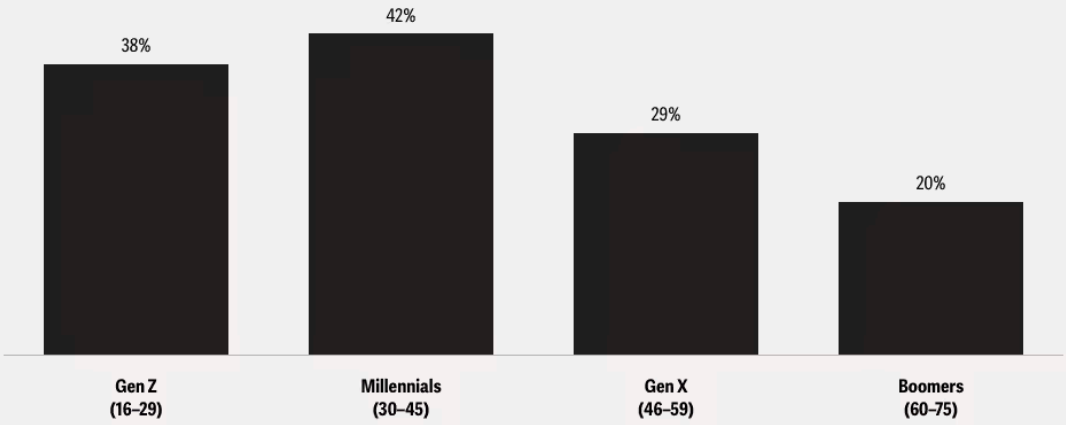
As telecom companies invest in non-network benefits, they should market them judiciously. High-budget, above-the-line campaigns are already used by some telecom companies to showcase rewards across screen, print, radio, social media, and billboards. T-Mobile US has celebrated 1 billion total "thank-you" gifts claimed, which included food, movies, gas, and trips.[22] Vodafone UK counted 175 million rewards via its VeryMe scheme.[23] O2 UK claimed its customers saved £23 million in one year via its Priority scheme.[24]

Operators should consider that Generation Z and millennial subscribers may be more amenable to the offer of perks versus network performance. A customer in their mid-20s may be unfamiliar with the sluggishness of 3G (the latest technology in the 2000s) and have mostly used 4G connections and perceived little difference from 5G. A customer in their 40s may have not had to try to browse on a 2G (the latest network in the 1990s) data connection. And so, some groups may be more likely to look for certain differentiators than others. According to Deloitte UK's research, surveyed Gen Z and millennials have a higher propensity to switch networks based on loyalty rewards than older age groups (figure 5).

Figure 5

A higher proportion of GenZ and Millennial consumers surveyed reportedly favor perks over performance more often than other generations

Percentage of UK consumers who would switch mobile network for loyalty rewards, by generation



Notes: Question: Which, if any, of the following would encourage you to switch mobile network provider? [Loyalty rewards or perks] Weighted base: all respondents who have a phone or smartphone, aged 16 to 75, UK (4,023).

Source: Deloitte Digital Consumer Trends, 2025.

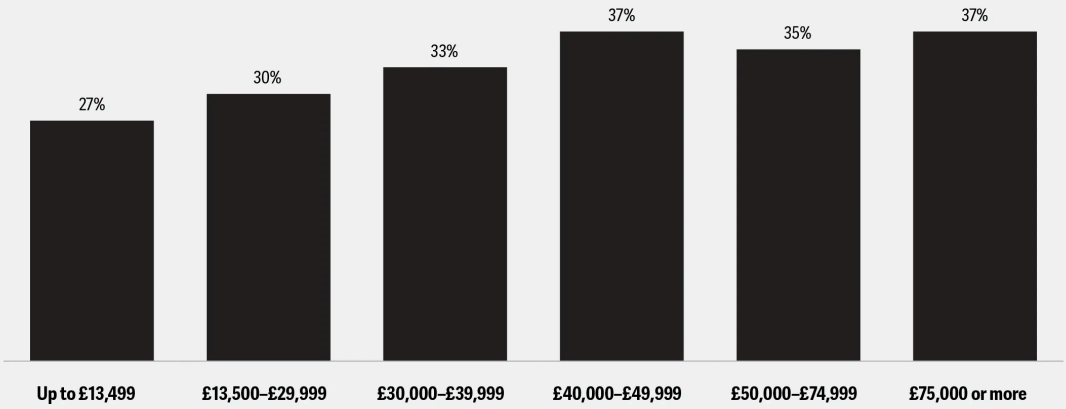
Deloitte. Insights | deloitteinsights.com

Operators should note that rewards may resonate more with higher-spend subscribers. Subscribers with higher incomes may be more inclined to switch for better offers than those on lower incomes (figure 6). The offer of a "freebie" can create a powerful, even slightly irrational, positive emotional response.[25]

Figure 6

Higher income consumers surveyed may appear to prefer perks

Percentage of UK consumers who would switch mobile network for loyalty rewards, by household income



Notes: Question: Which, if any, of the following would encourage you to switch mobile network provider? [Loyalty rewards or perks] Weighted base: all respondents who have a phone or smartphone, aged 16 to 75, UK (4,023).

Source: Deloitte Digital Consumer Trends, 2025.

If major telecom companies pursue similar strategies, one risk may be that rewards become commoditized, just like connectivity. Also, as alternate sectors like banking and utilities build their own schemes, the market can become further saturated.[26] Consumers may have a ceiling for the number of free coffees they are willing to consume—and if multiple service providers are providing the same perk, the appeal may get diluted. So creating a unique, differentiated scheme will be important to help attract customers and reduce churn. This means live events, concerts, and sports games may become particularly attractive properties,[27] but these perks can have limited reach, benefiting tens of thousands of customers among a base of tens of millions.

Paul Lee
United Kingdom

Ben Stanton
United Kingdom

Tim Bottke
Germany

Jody McDermott
Canada

Dieter Trimmel
Germany

Jack Fritz
United States

ENDNOTES

1. GSMA, 5G Network Slicing, accessed October 2025.
2. Paul Lee and Ben Stanton, *Deloitte Digital Consumer Trends 2025, UK Edition*, Deloitte LLP, June 2025; Adrie Cronje et al., *Digital Consumer Trends 2024, Netherlands Edition*, Deloitte LLP, December 2024; Vincent Frosty and Vincent Pirard, *Digital Consumer Trends 2024, Belgium Edition*, Deloitte Belgium, 2024.
3. Vodafone UK, 2G, accessed October 2025.
4. Simon Thomas, "What is the difference between 3G and 4G?," 4G.co.uk, September 29, 2014.
5. Ivor Nicholls, "LTE vs 4G: Understanding the difference between LTE and 4G," UCtel, February 10, 2025.
6. Andrew Wooden, "The telecoms industry's biggest problem? Failure to monetise 5G," *Telecoms.com*, March 14, 2024.
7. Ofcom, *Mobile matters*, July 17, 2025.
8. Espen Tønnessen, Thomas Haugen, and Shaher Ahmmad Ibrahim Shalfawi, "Reaction time aspects of elite sprinters in athletic world championships," *Journal of Strength & Conditioning Research* 27, no. 4 (2013): pp. 885–92.
9. IR, "Network latency—Common causes and best solutions," accessed October 2025.
10. Ofcom, *Mobile matters*.
11. Ofcom, "Improving your mobile phone reception," May 27, 2022.
12. William Webb, "It's time to rethink 6G," *IEEE Spectrum*, February 10, 2025.
13. Netflix, "Netflix-recommended internet speeds," accessed October 2025; Paul Lee, Dieter Trimmel, and Eytan Hallside, "No bump to bitrates for digital apps in the near term: Is a period of enough fixed broadband connectivity approaching?," TMT Predictions 2024, Deloitte, November 29, 2023.
14. Tefficient, "The demand for additional mobile data is weaker than ever—ARPU growth softens," July 31, 2025.
15. Michael Irving, "'Ultrabroadband' 6G chip clocks speeds 10 times faster than 5G," *ScienceAlert*, September 3, 2025; 4G.co.uk, "How fast are 4G and 5G?," accessed October 2025.
16. ETTelecom.com, "5G will make cost of GB lower than 4G: Experts," July 31, 2020.
17. NTT DOCOMO, "DOCOMO, NTT, NEC and Fujitsu develop top-level sub-terahertz 6G device capable of ultra-high-speed 100 Gbps transmission," press release, April 11, 2024.
18. New Street Research, Europea Tariff Tracker, accessed October 2024.
19. Jennifer Johnson, "Peak data growth is a quiet win for telcos," Reuters, June 2, 2025.
20. Evert de Boer and Xiao Yao Chin, *Top 100 most valuable airline loyalty programs*, On Point Loyalty, January 2023.
21. Consumer Financial Protection Bureau (CFPB), "CFPB takes action on bait-and-switch credit card rewards tactics," news release, last modified December 18, 2024.
22. Mike Sievert, "The power of appreciation: Taking customer loyalty to the next level," *Un-carrier blog*, T-Mobile, February 13, 2024.
23. Vodafone UK, "Spin for a chance to win £1,000 each day with VeryMe Rewards," press release, June 9, 2025.

24. Virgin Media O2, "Priority from O2 launches 'Blue Mondays' with millions of unmissable rewards, prizes and experiences for customers," April 28, 2025.

25. CI Group, "The psychology of freebies: Why small rewards yield big returns," accessed October 2025.

26. Octopus Energy, "Octoplus, our rewards programme for smart meter customers," accessed October 2025.

27. Vodafone UK, "Music festivals," accessed October 2025.

ACKNOWLEDGMENTS

The authors would like to thank **Matt Roberts**, **Pedro Goncalo Sanguinho**, **Dan Littman**, **James Brass**, **Michele Gabriel**, **Jolyon Barker**, **Dan Adams**, **Matt McDermott**, **Giles Warner**, **Jan-Piet Nelisson**, **Jonas Malmlund**, **Duncan Stewart**, **Gizem Bozdag**, and **Ralf Esser** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock