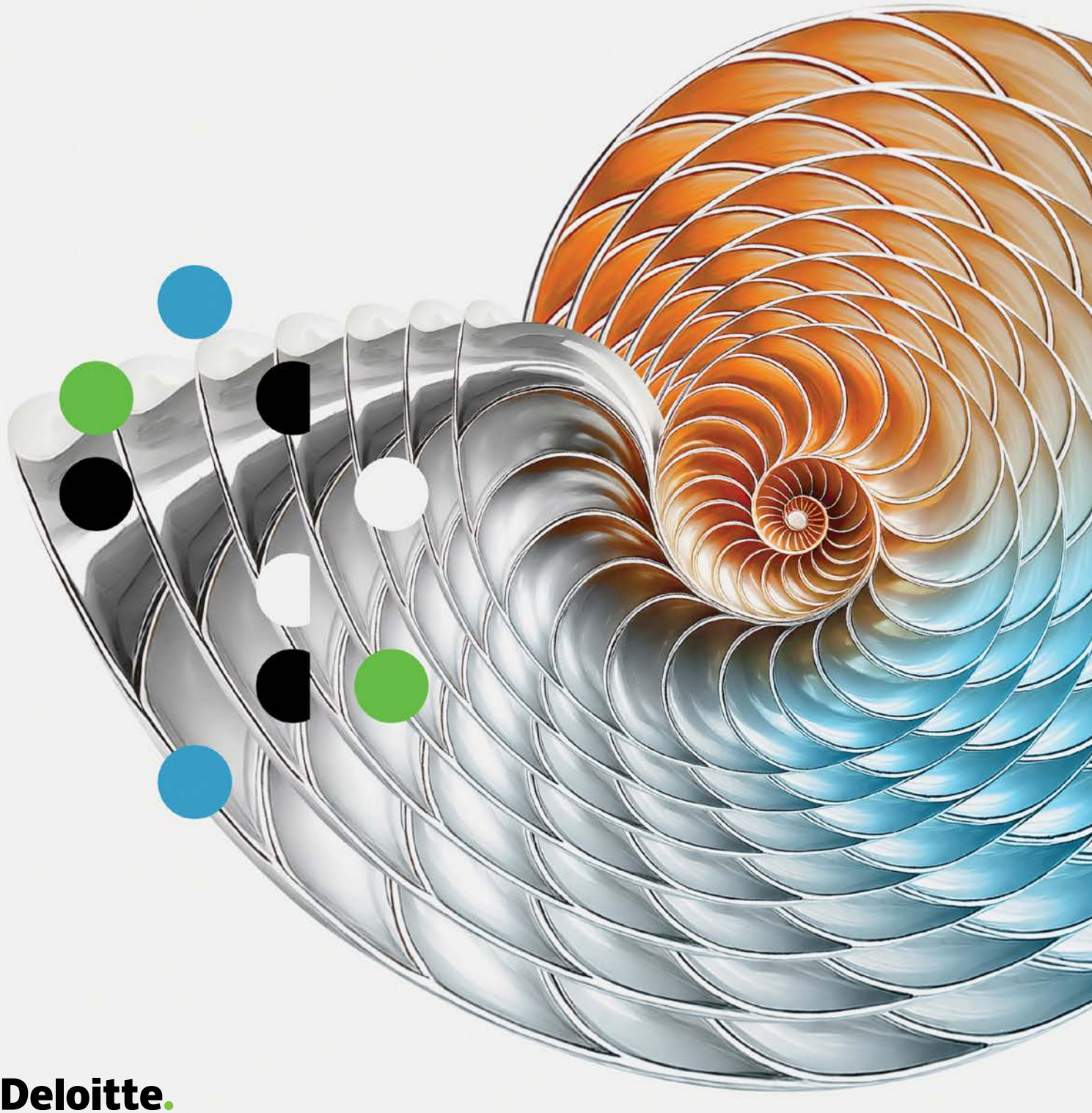
Tech Trends 2026

As technology innovation and adoption accelerate, five trends reveal how successful organizations are moving from experimentation to impact



Deloitte.
Insights

Table of contents

02 . . . **Executive summary**

04 . . . **Innovation compounds**

09 . . . **AI goes physical: Navigating the convergence of AI and robotics**

21 . . . **The agentic reality check: Preparing for a silicon-based workforce**

33 . . . **The AI infrastructure reckoning: Optimizing compute strategy in the age of inference economics**

43 . . . **The great rebuild: Architecting an AI-native tech organization**

53 . . . **The AI dilemma: Securing and leveraging AI for cyber defense**

62 . . . **Cutting through the noise: Tech signals worth tracking as AI advances**

Executive summary

Last year's Tech Trends report predicted that artificial intelligence would become akin to electricity, a foundational element that's seamlessly baked into an incredibly broad range of products and services. This year's report, the 17th annual edition of Tech Trends, proves that hypothesis. No corner of enterprise technology is untouched by AI as the demand for intelligent operations informs decisions on everything from computing hardware to physical robotics. And, while last year's focus was on building proof-of-concept projects and exploring the art of the possible, this year is all about scaling. Enterprises across industries are operationalizing AI-driven processes. The reason is simple: Leaders have realized that the key to competitive differentiation will be using AI to drive automation, innovation, and acceleration.

Innovation compounds

Technology leaders face a critical shift from AI experimentation to measurable impact. Innovation now compounds exponentially: Generative AI reached approximately 100 million users in just two months versus 50 years for telephones to reach 50 million users. This is creating a multiplying flywheel effect where improvements in technology, data, investment, and infrastructure simultaneously accelerate each other. Traditional infrastructure and sequential improvement processes can't keep pace. Success requires more than sophisticated technology. Organizations must redesign rather than merely automate processes, connect investments to business outcomes, and execute rapidly.

AI goes physical: Navigating the convergence of AI and robotics

Physical AI is evolving robots from preprogrammed machines into adaptive systems that perceive, learn, and operate autonomously in complex environments. These capabilities show up in industrial robots, autonomous vehicles, drones, and other systems. Current challenges include gaps in training, safety concerns, and cybersecurity risks, but falling costs are extending adoption beyond smart warehousing and supply chain operations into the mainstream. Humanoid robots are the next frontier, with projections of 2 million workplace humanoids by 2035. Future developments may include bio-hybrid robots and quantum robotics.

The agentic reality check: Preparing for a silicon-based workforce

Despite early enthusiasm, many businesses have yet to see significant transformation from agentic AI implementations because most simply automate existing processes rather than fundamentally redesigning operations. Only 11% of surveyed organizations have deployed agentic systems in production, with challenges including legacy system integration, data architecture constraints, and inadequate governance frameworks. Leading organizations are adopting agent-first process redesign, implementing multiagent orchestration using emerging protocols, and treating agents as a silicon-based workforce requiring specialized management frameworks. This includes agent onboarding, performance tracking, and FinOps cost management. The future points toward graduated autonomy levels, hybrid human-digital workforces, and leveraging agent-generated data for continuous learning, transforming how enterprises operate and compete.

The AI infrastructure reckoning: Optimizing compute strategy in the age of inference economics

As AI moves from experimentation to production, enterprises face an infrastructure dilemma. While token costs have dropped substantially, overall AI spending is exploding due to massive usage growth. Organizations are hitting a tipping point where cloud services become cost-prohibitive for high-volume workloads, with monthly bills reaching tens of millions. Leading enterprises are adopting strategic hybrid architectures: cloud for variable workloads, on-premises for consistent production inference, and edge for latency-critical applications. This can require purpose-built AI data centers featuring hardware optimized for graphics processing units, advanced networking, and specialized cooling. Future challenges include workforce reskilling, AI agents managing infrastructure, and sustainable computing innovations like renewable-powered and potentially orbital data centers.

The great rebuild: Architecting an AI-native tech organization

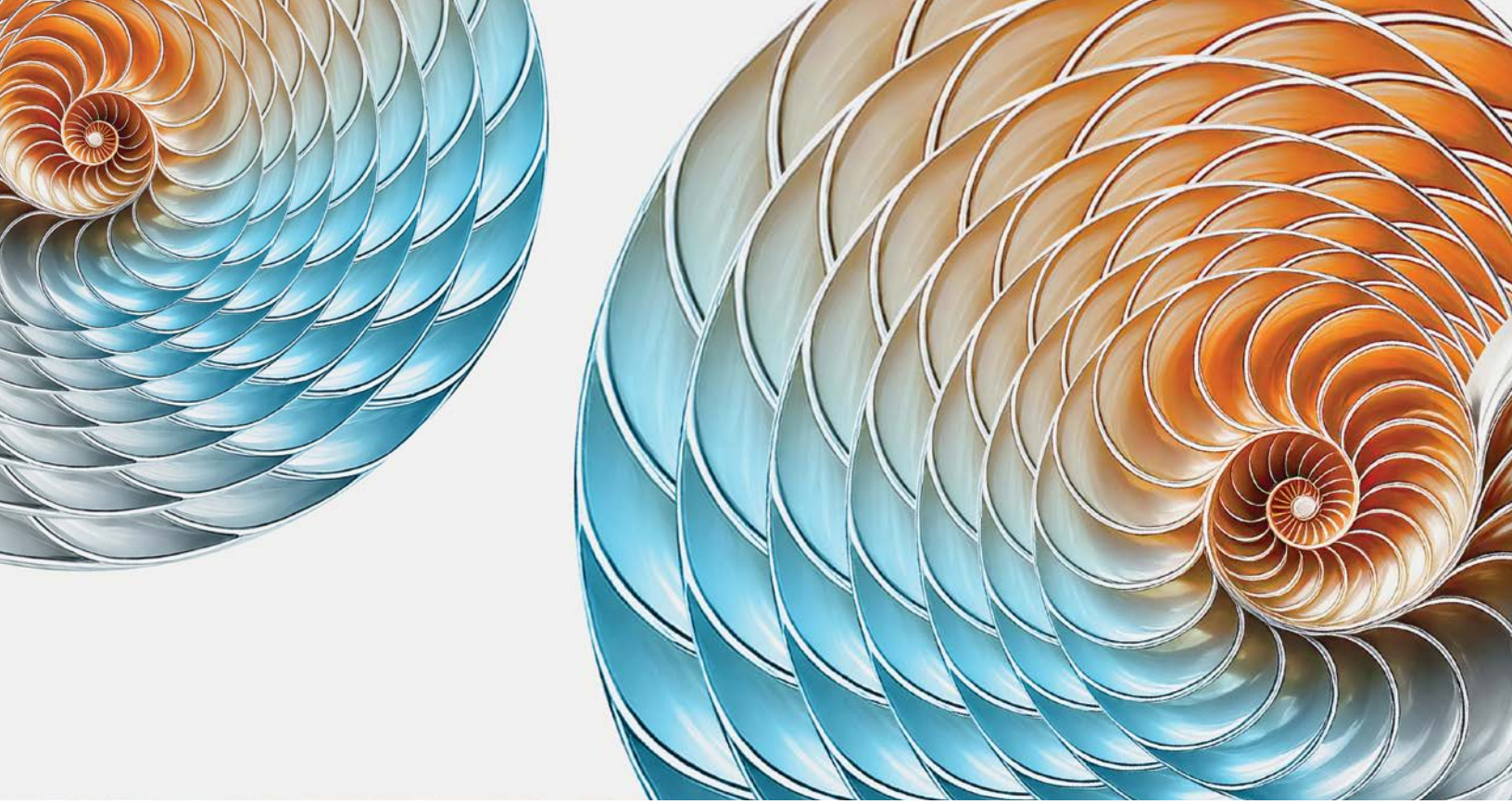
AI is fundamentally restructuring technology organizations beyond simple automation. With 64% of organizations increasing AI investments and tech budgets for AI rising, priorities are shifting from infrastructure maintenance to strategic leadership. Leading organizations are anchoring AI initiatives to measurable business outcomes, designing modular architectures for flexibility, and redefining talent strategies around human-machine collaboration. New roles are emerging, such as AI collaboration designers, edge AI engineers, and prompt engineers, while chief information officers are evolving from tech strategists to AI evangelists and orchestrators. Future tech organizations will feature agentic architectures, lean product-led teams, blended human-agent workforces, adaptive governance, and ecosystem-oriented innovation. Success requires embracing continuous evolution and boldly reimagining operations rather than incremental change.

The AI dilemma: Securing and leveraging AI for cyber defense

AI creates a cybersecurity paradox: The same capabilities driving business innovation are also introducing new risks. Organizations face threats from shadow AI deployments, adversarial attacks, and intrinsic AI system weaknesses across four domains: data, models, applications, and infrastructure. Existing security practices can be adapted to address AI-specific risks through robust access controls, model isolation, and secure deployment architectures. And AI offers powerful new capabilities to counter the very vulnerabilities it creates. Leading organizations are leveraging AI defensively through red teaming with AI agents, adversarial training, and automated threat detection at machine speed. Future challenges include AI-physical infrastructure convergence, autonomous cyber warfare, and quantum and space security threats. Success requires embedding security into AI initiatives from inception, treating it as an enabler rather than a constraint on innovation.

Cutting through the noise: Tech signals worth tracking as AI advances

Tech Trends takes a deep dive into five technology developments that are reshaping how businesses operate, but there are far more than five trends impacting organizations at any given moment. Eight adjacent "signals" also warrant monitoring. They include whether foundational AI models may be plateauing, the impact of synthetic data on models, developments in neuromorphic computing, emerging edge AI use cases, the growth in AI wearables, opportunities for biometric authentication, the privacy impact of AI agents, and the emergence of generative engine optimization. Some of these signals may mature into dominant forces. Others may fade. But all of them reflect the same underlying reality: The pace of technological change has fundamentally shifted, and the organizations that recognize these patterns early will have time to adapt.



Innovation compounds

As technology innovation and adoption accelerate, five trends reveal how successful organizations are moving from experimentation to impact

Kelly Raskovich

I spend most of my year in conversations with technology leaders, asking what's working, what isn't, and what keeps them up at night. Lately, those conversations have taken on a different quality.

The question used to be "What can we do with AI?" Now it's "How do we move from experimentation to impact?" The focus has moved from endless pilots to real business value, and there's a sense of urgency behind it all. Not because the technology is getting better—though it is—but because the pace of change itself has accelerated.

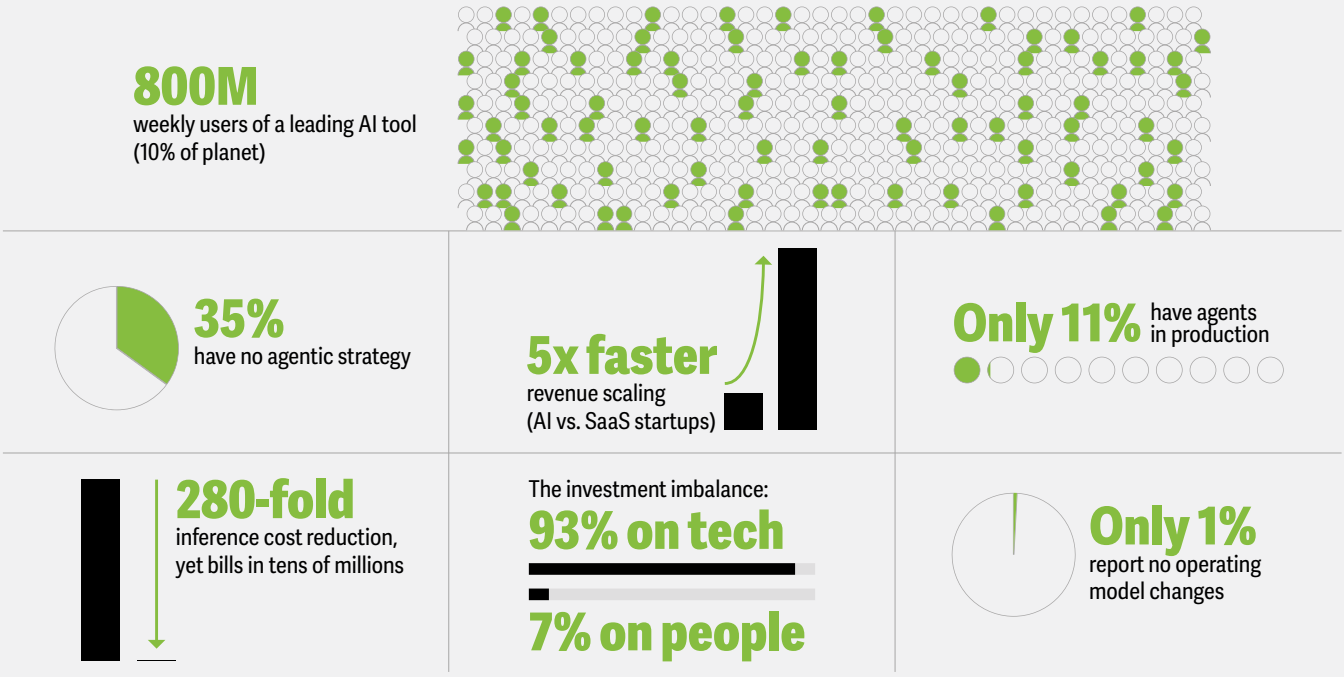
The numbers tell the story (figure 1). The telephone took 50 years to reach 50 million users. The internet took seven years. A leading generative AI tool reached about twice that many in two months.[1] As of this writing, that tool has over 800 million weekly users—roughly 10% of the planet's population.[2]

But rapid adoption is only the surface. Innovation is compounding; forces aren't simply additive, but multiplicative. Think of it as a flywheel: Better technology enables more applications. More applications generate more data. More data attracts more investment. More investment builds better infrastructure. Better infrastructure reduces costs. Lower costs enable more experimentation. Each improvement simultaneously accelerates all the others.

It's why AI startups scale from US$1 million to US$30 million in revenue five times faster than SaaS companies did.[3] It's why the knowledge half-life in AI has shrunk to months from years.[4] And it's why one chief information officer (CIO) told me, "The time it takes us to study a new technology now exceeds that technology's relevance window."

Figure 1

The AI transformation in numbers



Sources: Rebecca Bellan, "Sam Altman says ChatGPT has hit 800M weekly active users," TechCrunch, Oct. 6, 2025; Deloitte Emerging Technology Trends in the Enterprise Survey; Wing Venture Capital, "AI growing faster than SaaS"; Stanford Human-Centered AI Institute, "AI Index Report 2025"; Deloitte research on AI investment allocation patterns, 2025; Deloitte 2025 Tech Spending Outlook.

Every organization we studied is discovering the same truth: What got them here won't get them there.

The infrastructure built for cloud-first strategies can't handle AI economics. Processes designed for human workers don't work for agents. Security models built for perimeter defense don't protect against threats operating at machine speed. IT operating models built for service delivery don't drive business transformation.

This isn't only about enhancement. It's about rebuilding.

For 17 years, Tech Trends has explored emerging technologies poised to reshape business in the next 18 to 24 months. Our research is based on trend sensing from conversations with Deloitte subject matter experts and external technology leaders, as well as Deloitte's proprietary research on emerging technologies. This year, the data reveals five interconnected forces.

AI goes physical: Navigating the convergence of AI and robotics

Amazon deployed its millionth robot, and its DeepFleet AI coordinates the entire robot fleet, improving travel efficiency within warehouses by 10%.[5] BMW's factories have cars driving themselves through kilometer-long production routes.[6] Intelligence isn't confined to screens anymore; it's embodied, autonomous, and solving real problems in the physical world.

The agentic reality check: Preparing for a silicon-based workforce

Only 11% of organizations have agents in production, despite 38% piloting them. The gap between pilot to production tells you everything. Forty-two percent are still developing their strategy, while 35% have no strategy at all.[7] Gartner predicts that 40% of agentic projects will fail by 2027[8]—not because the technology doesn't work, but because organizations are automating broken processes instead of redesigning operations. HPE's chief financial officer captured what works: "We wanted to select an end-to-end process where we could truly transform, not just solve for a single pain point."[9] Redesign, don't automate. That's the pattern separating success from failure.

The AI infrastructure reckoning: Optimizing compute strategy in the age of inference economics

Token costs have dropped 280-fold in two years;[10] yet some enterprises are seeing monthly bills in the tens of millions. Usage exploded faster than costs declined. Organizations are discovering their existing infrastructure strategies aren't designed to scale AI to production-scale deployment. They're shifting from cloud-first to strategic hybrid: cloud for elasticity, on-premises for consistency, and edge for immediacy.

The great rebuild: Architecting an AI-native tech organization

AI is restructuring tech organizations, making them leaner, faster, and more strategic. Only 1% of IT leaders surveyed by Deloitte reported that no major operating model changes were underway.[11] Leaders are shifting from incremental IT management to orchestrating human-agent teams, with CIOs becoming AI evangelists. Success requires bold reimagination: modular architectures, embedded governance, and perpetual evolution as core capabilities.

The AI dilemma: Securing and leveraging AI for cyber defense

The technology meant to give businesses an advantage is becoming the target used against them. AT&T's chief information security officer captured the challenge: "What we're experiencing today is no different than what we've experienced in the past. The only difference with AI is speed and impact."[12] Organizations must secure AI across four domains—data, models, applications, and infrastructure—but they also have the opportunity to use AI-powered defenses to fight threats operating at machine speed.

Throughout this year's report, you'll meet technology leaders successfully navigating this sea change. They don't have all the answers, but there are noticeable patterns as they light the way forward.

- **They lead with problems, not technology.** Broadcom's CIO: "Without focusing on a specific business problem and the value you want to derive, it could be easy to invest in AI and receive no return."[13]
- **Specifically, their biggest problems.** UiPath CEO: "Rather than getting stuck in a cycle of perpetual proofs of concept, consider attacking your biggest problem and going for a big outcome."[14]
- **They prioritize velocity over perfection.** Western Digital's CIO: "We'd rather fail fast on small pilots than miss the wave entirely."[15]
- **They design *with* people, not just *for* them.** Walmart involved store associates in building its scheduling app, which includes shift swapping, schedule visibility, and employee control. The result: Scheduling time dropped from 90 minutes to 30 minutes, and people actually used the app.[16]
- **They treat change as continuous.** Coca-Cola's CIO described their journey as moving from "What can we do?" to "What should we do?"[17] That shift—from capability-first to need-first—is what separates productive experimentation from pilot purgatory.

I've tracked technology evolution long enough to recognize the patterns. The internet changed everything. Mobile reshaped consumer behavior. Cloud computing was transformative.

But this moment is different.

It's not just that AI is powerful. It's that the S-curves are compressing. The distance between emerging and mainstream is collapsing.

Organizations built for sequential improvement can't compete with those operating in continuous learning loops. The traditional playbook assumed you had time to get it right. That assumption no longer holds.

The organizations that succeed will probably not be those with the most sophisticated technology. They'll be those with the courage to redesign rather than automate, the discipline to connect every investment to business outcomes, and the velocity to execute before the window closes.

Innovation compounds. The gap between laggards and leaders grows exponentially. How you respond determines which side of that gap you're on.

But you don't have to navigate this alone. We hope this year's publication reminds you that everyone's facing this rapid pace of change, and together, we can shape what comes next.

Kelly Raskovich
Executive editor, Tech Trends

Endnotes

1. Jeff Desjardins, "In the race to 50 million users there's one clear winner - and it might surprise you," World Economic Forum, June 26, 2018; Alexandra Garfinkle, "ChatGPT on track to surpass 100 million users faster than TikTok or Instagram: UBS," Yahoo Finance, Feb. 2, 2023.
2. Rebecca Bellan, "Sam Altman says ChatGPT has hit 800M weekly active users," TechCrunch, Oct. 6, 2025.
3. Zach DeWitt, "AI growing faster than SaaS," Wing Venture Capital, November 7, 2024.
4. Based on Deloitte analysis of technology adoption cycles and AI capability evolution timelines.
5. Scott Dresser, "Amazon deploys over 1 million robots and launches new AI foundation model," Amazon, July 1, 2025.
6. Brad Anderson, "Who needs factory drivers when cars drive themselves at BMW plants," CarScoop, Nov. 26, 2024.
7. Deloitte 2025 Emerging Technology Trends in the Enterprise Survey. From June to July 2025, Deloitte conducted an online survey of 500 US technology leaders to quantify the prevalence, engagement, and perceptions surrounding the adoption of emerging technologies across industries.
8. Gartner, "Gartner predicts over 40 % of agentic AI projects will be canceled by end of 2027," press release, June 25, 2025.
9. Marie Myers (executive vice president and chief financial officer, HPE), interview with Deloitte, March 1, 2025.
10. Stanford Institute for Human-Centered Artificial Intelligence, "The AI Index report 2025," accessed Nov. 12, 2025.
11. Deloitte 2025 Tech Spending Outlook. From June to July 2025, Deloitte conducted an online survey of 302 IT procurement leaders, heads of IT, and non-IT executives with technology spending oversight to understand how US enterprises in key industries are managing technology budgets.
12. "A no-nonsense approach to secure AI enablement at AT&T," *Deloitte Insights*, Nov. 21, 2025.
13. Katherine Noyes, "Broadcom CIO: 'Modernization should be driven by the business'," CIO Journal, *The Wall Street Journal*, and Deloitte, Sept. 10, 2025.
14. Katherine Noyes, "UiPath CEO: Agentic automation will 'usher in a new era of work'," CIO Journal, *The Wall Street Journal*, and Deloitte, Feb. 21, 2025.
15. Katherine Noyes, "Western Digital CIO: In the AI era, 'Play offense or get left behind'," CIO Journal, *The Wall Street Journal*, and Deloitte, Sept. 6, 2025.
16. Walmart, "Walmart unveils new AI-powered tools to empower 1.5 million associates," June 24, 2025.
17. Katherine Noyes, "Coca-Cola CIO on scaling AI: From 'What can we do?' to 'What should we do'," CIO Journal, *The Wall Street Journal*, and Deloitte, Jan. 18, 2025.

AI goes physical: Navigating the convergence of AI and robotics

Powered by artificial intelligence, traditional robots are becoming adaptive machines that can operate in—and learn from—complex environments, unlocking safety and precision gains

Jim Rowan, Tim Gaus, Franz Gilbert, and Caroline Brown

Robots powered by physical AI are no longer confined to research labs or factory floors. They're inspecting power grids, assisting in surgery, navigating city streets, and working alongside humans in warehouses. The transition from prototype to production is happening now.

Physical AI refers to artificial intelligence systems that enable machines to autonomously perceive, understand, reason about, and interact with the physical world in real time. These capabilities show up in robots, vehicles, simulations, and sensor systems. Unlike traditional robots that follow preprogrammed instructions, physical AI systems perceive their environment, learn from experience, and adapt their behavior based on real-time data. Automation alone doesn't make them revolutionary; rather, it's their capacity to bridge the gap between digital intelligence and the physical world.

In the nascent but rapidly evolving category of robots, physical AI turns robots into adaptive, learning machines that can operate in complex, unpredictable environments. The combination of AI, mobility, and physical agency enables robots to move through environments, perform tasks, and interact with the world in ways that fundamentally differ from enhanced appliances. Embodied in robotic systems, physical AI is quite literally on the move.

Today, AI-enabled drones, autonomous vehicles, and other robots are becoming increasingly common, particularly in smart warehousing and supply chain operations. The industry, regulatory bodies, and potential adopters are working to break down barriers that hinder the deployment of solutions at scale. As organizations overcome these challenges, AI-enabled robots will likely transition from niche to mainstream adoption. Eventually, we'll witness physical AI's next evolutionary leap: the arrival of humanoid robots that can navigate human spaces with unprecedented capability.

From prototype to production

Unlike traditional AI systems that operate solely in digital environments, physical AI systems integrate sensory input, spatial understanding, and decision-making capabilities, enabling machines to adapt and respond to three-dimensional environments and physical dynamics. They rely on a blend of neural graphics, synthetic data generation, physics-based simulation, and advanced AI reasoning. Training approaches such as reinforcement learning and imitation learning enable these systems to master principles like gravity and friction in virtual environments before being deployed in the real world.

Robots are only one embodiment of physical AI. It also encompasses smart spaces that use fixed cameras and computer vision to optimize operations in factories and warehouses, digital twin simulations that enable virtual testing and optimization of physical systems, and sensor-based AI systems that help human teams manage complex physical environments without requiring robotic manipulation.

Whereas traditional robots follow set instructions, physical AI systems perceive their environment, learn from

experience, and adapt their behavior based on real-time data and changing conditions. They manipulate objects, navigate unpredictable spaces, and make split-second decisions with real-world implications. Robot dogs process acoustic signatures to detect equipment failures before they become catastrophic. Factory robots recalculate their routes when production schedules shift mid-operation. Autonomous vehicles use sensor data to spot cyclists sooner than human drivers. Delivery drones adjust their flight paths as wind conditions change. What makes these systems revolutionary isn't just task automation but their capacity to perceive, reason, and adapt, which enables them to bridge the gap between digital intelligence and the physical world.[1]

Tech advancements drive physical AI–robotics integration

Physical AI is ready for mainstream deployment because of the convergence of several technologies that impact how robots perceive their environment, process information, and execute actions in real time.

Vision-language-action models. Physical AI adopts training methods from large language models (LLMs) while incorporating data that describes the physical world. Multimodal vision-language-action (VLA) models integrate computer vision, natural language processing, and motor control.[2] Like the human brain, VLA models help robots interpret their surroundings and select appropriate actions (figure 1).

Onboard computing and processing. Neural processing units—specialized processors optimized for edge computing—enable low-latency, energy-efficient, real-time AI processing directly on robots. Onboard capability allows physical AI systems to run LLMs and VLA models, process high-speed sensor data, and make split-second, safety-critical decisions without cloud dependency—essential for autonomous vehicles, industrial robotics, and remote surgery.[3] It can also transform robots from isolated machines into autonomous systems that can share knowledge and coordinate actions across intelligent networks.

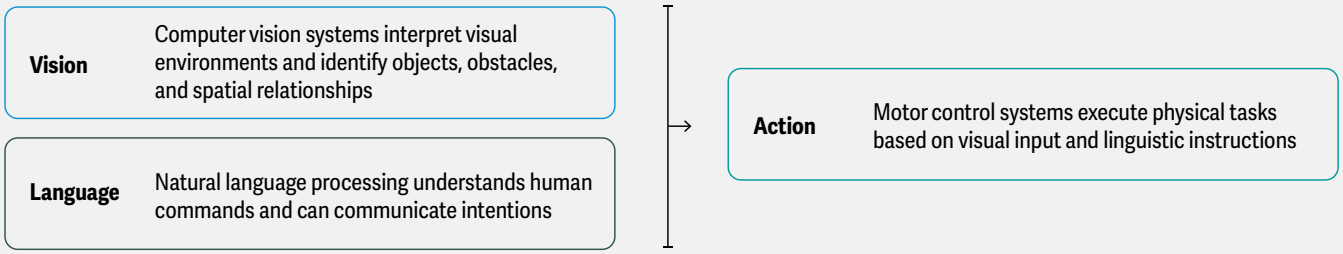
Robotics advancements have made robots more accessible and capable:[4]

- **Computer vision** for "seeing" and understanding surroundings
- **Sensors** for capturing information such as sound, light, temperature, and touch
- **Actuators** for movement, inspired by human muscles
- **Spatial computing** for navigating 3D environments
- **Improved batteries** that enable longer operation without frequent recharging

Training and learning. In reinforcement learning, robots develop sophisticated behaviors through trial and error

Figure 1

How vision-language-action models work



Source: Deloitte analysis.

by receiving rewards or penalties. In imitation learning, robots mimic expert demonstrations. Both approaches can be applied in simulated environments or in the physical world with real hardware.[5] A blend of these techniques, starting with simulation-based reinforcement training and then fine-tuning with targeted physical demonstrations, can create continuous learning loops. This helps robots continue to improve by feeding real-world data back into their training policies and simulation spaces.[6]

Compelling economics boost industrial adoption

As technology advances, costs have been coming down, and many real-world applications have emerged.

Advanced manufacturing infrastructure now supports the production of complex robotics and physical AI systems at enterprise scale. This means that physical AI robots can now be produced with the reliability and quality control of smartphones or cars, making them practical for everyday industrial use.

Component commoditization and open-source development are reducing entry costs for physical AI systems. However, because these robots need advanced AI chips and processors, they remain more expensive than traditional industrial robots. For now, this cost gap is likely to persist, even as overall prices gradually decline.

These economics are driving the adoption of physical AI and robotics in select use cases. Autonomous vehicles and drones are the most visible robotic form factors (figure 2). Waymo's robotaxi service has completed over 10 million paid rides, while Aurora Innovation has launched the first commercial self-driving truck service with regular freight deliveries between Dallas and Houston.[7]

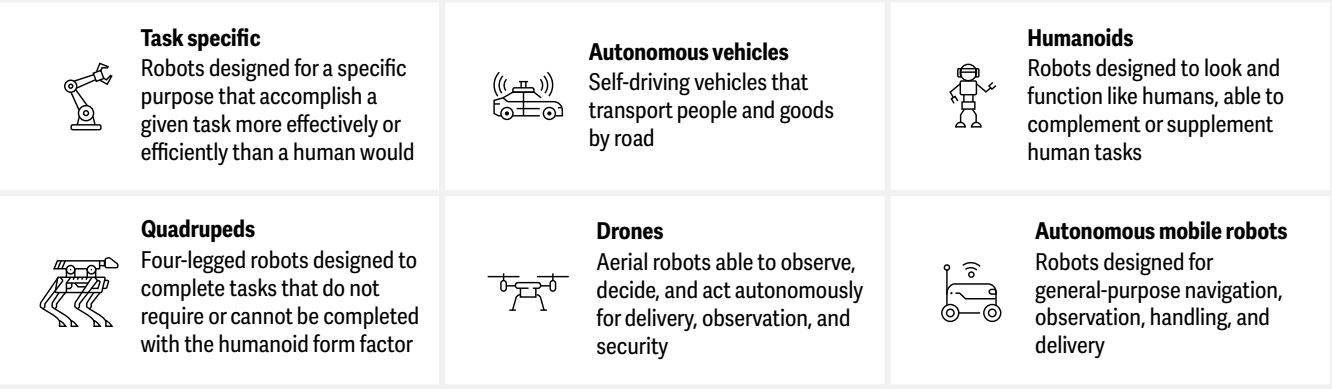
AI-enabled drones are fundamentally changing consumer expectations around speed and convenience, while also serving as powerful commercial tools. Equipped with advanced cameras and sensors, drones now manage warehouse inventory autonomously by navigating between shelves and scanning products with barcode and QR code readers.[8]

In the enterprise, warehousing and supply chain operations are the earliest adopters of physical AI robotic systems, likely due to labor market pressures.[9]

Many organizations now use these systems at scale. For example, Amazon recently deployed its millionth robot,

Figure 2

Six key form factors for robotics and physical AI



Source: Mark Osis, Raquel Buscaino, and Caroline Brown, "Robotics & physical AI," Deloitte, 2025.

part of a diverse fleet working alongside humans.[10] Its DeepFleet AI model coordinates the movement of this massive robot army across the entire fulfillment network, which Amazon reports will improve robot fleet travel efficiency by 10%.[11]

Similarly, BMW is integrating AI automation into its factories globally. In one novel deployment, BMW uses autonomous vehicle technology—assisted by sensors, digital mapping, and motion planners—to enable newly built cars to drive themselves from the assembly line, through testing, to the factory's finishing area, all without human assistance.[12]

The physical AI inflection point

As technologies advance and converge, costs decrease, and viable use cases emerge, physical AI–driven robots are poised to transition from niche to mainstream adoption—provided that technical, operational, and societal challenges can be overcome.

Breaking through implementation barriers

As organizations seek to scale physical AI, they're encountering a set of complex, interrelated implementation challenges. The technology works, but making it work at scale requires solving problems that span technical, operational, and regulatory domains. Organizations that tackle these challenges head-on will define the next wave of deployment.

Training and learning. Simulation environments offer critical advantages in speed, safety, and scalability, but there's a persistent gap between simulated and real-world performance caused by approximated physics models.[13] "Visual images in simulated environments are pretty good, but the real world has nuances that look different," says Ayanna Howard, dean of the College of Engineering at The Ohio State University. "A robot might learn to grab something in simulation, but when it enters physical space, it's not a one-to-one match."[14] (See the sidebar for the full Q&A.)

THE HUMAN FACTOR: AYANNA HOWARD ON PHYSICAL AI AND THE FUTURE OF ROBOTICS

Ayanna Howard is the dean of the College of Engineering at The Ohio State University and a prominent roboticist and advocate for AI safety and alignment. Previously, she was a senior robotics researcher at NASA's Jet Propulsion Laboratory and later chaired Georgia Tech's School of Interactive Computing and founded the Human-Automation Systems Lab.

Q: What technology challenges are holding back progress in physical AI and robotics?

A: One of the fundamental challenges is that the physical world is inherently dynamic. I can walk into my office every day, but there's always some difference—maybe someone vacuumed, moved things around, or my computer doesn't boot up. The question is, how do you simulate all these variations so robots can learn to adapt, walk, lift, and interact with uncertainty the way humans do? You can't just practice endlessly in the real world because you'll break things.

There's also a hardware limitation I paraphrase as the "manipulation-to-physical-body ratio." Some humans can lift their own weight or more, but conventional robots—even heavy ones—often can't lift half their weight due to actuator limitations. They don't have muscles like we do to offset rigid actuation, which limits what they can interact with and move.

Finally, there's the real-time processing challenge. Large language models and vision-language models typically function in what I call "human time": We're ok waiting a second or two for a response. But if a robot is walking and needs to make a decision, a one- or two-second delay means it drops something, crashes, or potentially hurts someone. We're getting better at real-time processing, but we're not quite there yet.

Q: You've done extensive research on trust and overtrust in AI systems. Can you explain how both extremes pose challenges?

A: It turns out that the difference between stated trust and behavioral trust is significant. In other words, people often say they don't trust AI, but if you ask whether they use a phone or computer, or even leave their house, guess what? They're using AI.

My research on overtrust focuses on behaviors, not what people say. We've conducted studies where people interact with robotic systems that were programmed to make mistakes. When surveyed, participants said they didn't trust the systems because they had seen them make errors. But when we analyzed their actual behaviors, we saw something different: Their actions showed they did trust the robot.

With physical embodiments of AI, this behavioral overtrust becomes dangerous because these robots apply physical forces in the environment. When they do things, the consequences can be irreversible. With today's AI, you still need human

THE HUMAN FACTOR: AYANNA HOWARD ON PHYSICAL AI AND THE FUTURE OF ROBOTICS, CONTINUED

actuation for most tasks, though agentic AI is starting to change that landscape.

Q: What are the most critical research areas that need investment?

A: Learning in physical spaces without causing harm. We still need to figure out how to translate simulation to the physical world safely. Visual images in simulated environments are pretty good, but the real world has nuances that look different. A robot might learn to grab something in simulation, but when it enters physical space, it's not a one-to-one match.

In research, robots do adapt after moving from simulation to physical environments, but they learn around tasks, not holistic environmental interactions. They might learn to grasp balls on different surfaces with varying friction coefficients. But they're not learning how close to get to people in a mall or on a college campus while juggling those same grasped balls based on simulated social interactions. That kind of comprehensive environmental adaptation doesn't exist yet.

Q: Do you have any hot takes that go against conventional wisdom?

A: I fundamentally believe there should always be a human in the loop somewhere. Always. And I'm a roboticist saying this. It doesn't ensure safety 100%, but it helps mitigate overtrust. Maybe it's the CEO doing annual reviews of the robots. Without that feedback loop, this can get away from us.

Advances in physics engines, synthetic data generation, and approaches that blend virtual training with real-world applications should help organizations achieve the quality of physical training at the scale and safety of simulation.

Trustworthy AI and safety. The smallest error rates can have cascading effects in physical systems, potentially leading to production waste, product defects, equipment damage, or safety incidents. If AI systems hallucinate, errors could be perpetuated and amplified across entire production runs, creating compounding downstream effects on costs and operations.

AI-powered machines can behave unpredictably even after extensive safety testing. The stakes rise significantly in public spaces, where autonomous systems must navigate unpredictable human behavior. To scale physical AI systems across various industries, comprehensive safety strategies that integrate regulatory compliance, risk assessments, and continuous monitoring are necessary.[15]

Regulatory environment. Companies must navigate overlapping and sometimes contradictory requirements across jurisdictions.[16] As robots move from controlled factory environments into public spaces, regulatory bodies are likely to develop new frameworks for safety certification, liability, and operational oversight.

Data management. Organizations must capture and manage massive amounts of sensor data, 3D environmental models, and real-time information. High-fidelity digital twins of physical assets are essential for effective training and deployment, requiring extensive data on physical characteristics, object properties, and interactions. Organizations will also need to integrate multimodal data from disparate sources, ensure data security, and manage data infrastructure costs.

Human acceptance. While most workers are generally comfortable with predictable, rule-based robots, physical AI systems that learn and adapt introduce new uncertainties, especially worries about job displacement. However, experts predict that most roles will evolve toward collaboration rather than replacement.[17] The goal is to create environments where robots handle repetitive or dangerous tasks while humans focus on creative problem-solving and complex decision-making.

Cybersecurity vulnerabilities. As discussed in "The AI dilemma," physical AI systems create new attack surfaces that bridge digital and physical domains. Connected fleets increase cyber risks, with vulnerabilities potentially leading to unauthorized access, data breaches, or even malicious robot control. The stakes are even higher when security breaches can affect physical safety and operational continuity.

Robot fleet orchestration. As physical AI systems mature, organizations will increasingly deploy heterogeneous fleets of robots, autonomous vehicles, and AI agents from multiple vendors, each with proprietary protocols. This creates interoperability challenges that can lead to accidents, downtime, system congestion, and operational inefficiency.[18] Autonomous fleet management and orchestration systems can help resolve these issues.

Over the coming 18 to 24 months, resolving these foundational issues will likely enable physical AI and robotics to expand beyond traditional industries. Warehousing and logistics may have served as physical AI's proving ground, but sector boundaries do not limit the technology.

Crumbling sector boundaries

As leading organizations across the public and private sectors are laying the groundwork for physical AI at scale, adoption is accelerating exponentially. Applications are emerging wherever physical AI solves real problems.

In health care, a sector facing global staffing shortages, medical technology companies are developing AI-driven robotic surgery and digital imaging devices. GE HealthCare is building autonomous X-ray and ultrasound systems with robotic arms and machine vision technologies. Other medtech companies are designing intelligent robotic assistants that can help with patient care and automate surgical tasks.[19]

Restaurants are also deploying robots to help address labor shortages. Sidewalk-crawling delivery robots travel at pedestrian speeds; inside restaurants, robots handle tasks like flipping burgers and preparing salads, while service robots seat customers and serve food.

Naturgy Energy Group, a Spanish multinational natural gas and electrical energy utilities company, currently uses drones for inspection purposes. Rafael Blesa, Naturgy's chief data officer, envisions an expanded role for physical AI as the technology hardens, particularly in dangerous field operations involving high voltage or open gas pipes. "Many operations related to grid maintenance could be performed by robots in the long term," he explains. "My expectation is that in three to four years, we'll have robots performing physical operations, which could save lives."[20]

Similarly, the city of Cincinnati is using AI-powered drones to autonomously inspect bridge structures and road surfaces, reducing costs, keeping human inspectors out of hazardous situations, and condensing months of analysis into minutes. "This type of technology is going to be the nuts and bolts of what's going to allow [mayors] to do their jobs better and provide better information, decisions, and cost efficiencies for their constituents," said Cincinnati's mayor, Aftab Pureval.[21]

In 2024, the city of Detroit launched a free autonomous shuttle service designed for seniors and people with disabilities whose mobility was severely limited by traditional transit systems. Known as Accessibili-D, the self-driving vehicles were equipped with wheelchair accessibility and a trained safety operator. Three autonomous vehicles operated within an 11-square-mile section of Detroit, offering 110 different stops.[22]

Regardless of the sector, these deployments share a common characteristic: They augment human capabilities in situations where safety, precision, or accessibility are most critical.

Humanoid robots and beyond

We've all seen the viral videos of humanoid robots with their fluid, not-quite-human-but-pretty-darn-close movements. They're the most compelling robotic form factor, not because they have the most efficient design, but because our world is built for human bodies. This means they can navigate existing infrastructure—doorways, staircases, factory floors, and home kitchens—without costly modifications to accommodate specialized robotic systems.[23]

"People are very compliant in how they interact with the world and constantly make contact with their environment. That's very hard for a commercial robot," says Jonathan Hurst, a robotics researcher at Oregon State University and cofounder of Agility Robotics. "Typically, robots are very position-controlled devices. They're good for things like CNC machining [precision manufacturing requiring exact, repeatable positioning] or spot welding, but they're not good for assembly, manipulation, or locomotion in nonstructured spaces."[24] (See sidebar for the full Q&A.)

Several companies have developed and continue to refine bipedal robots with more precise finger control. With the recent introduction of chain-of-thought reasoning abilities comparable to human cognition, the technological foundation continues to advance.[25]

During the next decade, the intersection of agentic AI systems with physical AI robotic systems will result in robots whose "brains" are agentic AIs. Robots of all form factors should increasingly be able to adapt to new environments, plan multistep tasks, recover from failure, and operate under uncertainty. The impact of this technology convergence will be particularly profound for humanoid robots.

Instead of custom robotics for each domain, more general agentic modules may be reused across warehouses, homes, health care, agriculture, and other areas. Agentic humanoids could one day function as assistants, coworkers, or health care aides with more intuitive interaction, reasoning, and negotiation capabilities.

Mass adoption of humanoids is likely several years away. Still, UBS estimates that by 2035, there will be 2 million humanoids in the workplace, a number it expects to increase to 300 million by 2050. The firm estimates the total addressable market for these robots will reach between US$30 billion and US$50 billion by 2035 and climb to between US$1.4 trillion and US$1.7 trillion by 2050.[26]

Enterprise applications like warehousing and logistics remain the proving ground for humanoid deployment, driven by labor shortages. BMW is testing humanoid robots at its South Carolina factory for tasks requiring dexterity that traditional industrial robots lack: precision manipulation, complex gripping, and two-handed coordination.[27] For similar reasons, humanoids could play a role in health care. One health care company is testing humanoids in rehabilitation centers to assist therapists by guiding patients through exercises and providing weight support.[28]

The larger long-term opportunity lies in consumer markets, where the vision extends to comprehensive household tasks such as elderly and disability care, cleaning and maintenance, meal preparation, and laundry. The Bank of America Institute projects that the material costs of a humanoid robot will fall from around US$35,000 in 2025 to between US$13,000 and US$17,000 per unit in the next decade, and Goldman Sachs reports that humanoid manufacturing costs dropped 40% between 2023 and 2024.[29]

FROM THE LAB TO THE REAL WORLD: JONATHAN HURST ON HUMANOID ROBOTS

Jonathan Hurst is a robotics professor at Oregon State University and cofounder of the school's Robotics Institute, where his research focuses on legged locomotion. He's also the cofounder and chief robot officer of Agility Robotics, which develops and deploys humanoid robots that operate alongside human workers in commercial applications.

Q: Were you trying to solve a specific problem by building a robot with a humanoid form factor?

A: We wanted to make machines that move like animals or people and that could also exist in human spaces. People are very compliant in how they interact with the world and constantly make contact with their environment. That's very hard for a commercial robot. Typically, robots are very position-controlled devices. They're good for things like CNC machining [precision manufacturing requiring exact, repeatable positioning] or spot welding, but they're not good for assembly, manipulation, or locomotion in nonstructured spaces.

With our robot, we've gotten pretty close to a normal human-like leg configuration—bipedal, upright torso, bimanual. The most important thing is that each of these features has a purpose. We are capturing the function that underlies that form.

Q: How did you figure out what the humanoid could do?

A: From the beginning, we aimed to build a human-centric, multipurpose robot. We looked at hundreds of use cases. It turned out that the simple task of lifting and moving bins and totes is a good match for the technology. This task requires something with a narrow footprint to operate in hallways, go through doors, and be in human spaces. It needs to be able to lift something heavy—like 25 kilograms—to the top of a two-meter shelf.

For this, you need something that's dynamically stable—a robot that maintains its balance while in motion. A statically stable base will just tip over if you're trying to lift these things. Therefore, a bipedal pair of legs is the most effective way to be dynamically stable and not fall.

That's the starting point. From there, we got it to be bimanual because, to pick up big things, you need a grasp on both sides. You need a reachable workspace to pick something off the ground and

FROM THE LAB TO THE REAL WORLD: JONATHAN HURST ON HUMANOID ROBOTS, CONTINUED

lift it up high, and an upright torso to do that, so it's a particularly good match for the technology.

That's very hard to do with existing automation. There is a fair amount of flexibility needed because all the workflows are distinctive. Different types of totes go to different places, for example. You stack the totes, palletize them, put them on conveyor belts, or take them off AMRs [autonomous model robots]. This kind of variety makes it hard for traditional automation, yet it's still quite structured. You might say semi-structured. It's in an industrial environment that's well-controlled and process-automated, which makes it a really good starting point for a humanoid.

Q: How can the use of humanoids scale?

A: The market for humanoids is going to be twice the size of the automotive industry in 25 years. There's a lot of scaling to be done to get to that point because that's millions of robots. Today, there are only hundreds of robots.

There is a massive market for the functionally safe humanoid—the robot that doesn't have to be confined to its own work cell. That's when you'll be able to start deploying robots in the thousands. How do you support them in the field? How do you have your robot fleet management software work over all the unique bandwidth limitations and everything else? That's a hard thing to do in robotics, but it can be done. Waymo has deployed what are basically robots on the roads, so it's definitely feasible. It's not like something has to be invented, but the organization has to execute really, really capably to do that. That's the journey we're on, once the robot is safe enough to warrant the scale.

Beyond humanoids?

Humanoid robots capture public imagination with their familiar bipedal form. Where do we go from there?

In terms of physical form factors, boundary-pushing engineers are increasingly experimenting with machines that blur biological lines. Imagine robots powered by living mushroom tissue, those that mimic movements using rat muscle tissue, or machines that can transition between solid and liquid states using magnetic fields. In innovative laboratories today, scientists are integrating living organisms into mechanical systems, developing robots that can navigate complex environments through multiple modes of locomotion, and creating machines that adapt their physical form to match the task.[30]

Quantum robotics—the combination of quantum computing and AI-powered robotics—also holds promise, though it's still in the very early stages. Superposition, entanglement, quantum algorithms, and other quantum computing principles could allow robots to operate at speeds that are impossible for today's binary computers.[31] Quantum algorithms are expected to improve processing, navigation, decision-making, and fleet coordination, while quantum sensors will enhance perception and interaction.[32]

Useful quantum robots are expected to be many decades away. Hardware immaturity, integration challenges, and the extreme sensitivity of quantum states are just a few of the challenges that must be solved before quantum computing can be widely deployed.[33]

Humanoid butlers are at least a decade away, and exotic form factors and quantum capabilities remain largely experimental. But they represent a fundamental shift in how we think about robotics. As these breakthrough technologies graduate from the lab to the enterprise to the home, the field of robotics is moving beyond simply automating human tasks toward creating entirely new categories of machines.

Endnotes

1. Nvidia, "What is physical AI?" accessed Nov. 6, 2025.
2. Anony, "Vision-language-action models for embodied AI: A survey overview," Medium, May 12, 2025.
3. Josh Schneider and Ian Smalley, "What is a neural processing unit (NPU)?" IBM, accessed Nov. 6, 2025.
4. Jiefei Wang and Damith Herath, "What makes robots? Sensors, actuators, and algorithms," *Foundations of Robotics* (Singapore: Springer, 2022); Bank of America Institute, "Humanoid robots 101," April 29, 2025.
5. *MIT Technology Review*, "Training robots in the AI-powered industrial metaverse," Jan. 14, 2025.
6. Automate, "NVIDIA on what physical ai means for robotics," Aug. 5, 2025.
7. Mark Osis, Raquel Buscaino, and Caroline Brown, "Robotics and physical AI: Intelligence in motion," Deloitte, Oct. 17, 2025.
8. Ibid.
9. Ibid.
10. Michael Grothaus, "What are physical AI and embodied AI? The robots know," *Fast Company*, July 19, 2025.
11. Scott Dresser, "Amazon launches a new AI foundation model to power its robotic fleet and deploys its 1 millionth robot," Amazon, July 1, 2025.
12. Brad Anderson, "Who needs factory drivers when cars drive themselves at BMW plants," *Carscoops*, Nov. 26, 2024.
13. Erica Salvato, Gianfranco Fenu, Eric Medvet, and Felice Andrea Pellegrino, "Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning," *IEEE Access* 4, 2016.
14. Ayanna Howard, interview with Deloitte, Sept. 18, 2025.
15. Standard Bots, "Industrial robot safety standards: What you need to know," April 23, 2025.
16. Jacob Otasowie, Alexander Blum, Mohamed El Sayed Ahmed, and Mathias Brandstötter, Danger of AI in robotics: A systematic analysis of ethical, regulatory, and economic challenges," *Springer*, Sept. 2, 2025.
17. Osis, Buscaino, and Brown, "Robotics and physical AI."
18. Rexroth, "Efficient fleet management: How to successfully orchestrate heterogeneous vehicle fleets," Aug. 30, 2024.
19. Conor Hale, "Nvidia outlines new AI projects in robotic surgery, autonomous imaging," *Fierce Biotech*, March 21, 2025.
20. Rafael Blesa, interview with Deloitte, May 22, 2025.
21. Deloitte US, "Rust belt renaissance: Cincinnati's OptoAI story," YouTube video, Sept. 15, 2023.
22. Deloitte US, "Detroit gets moving on autonomous vehicles," accessed Nov. 6, 2025.
23. *Humanoid Robotics Technology*, "Top 12 humanoid robots of 2025," February 2025.
24. Jonathan Hurst, interview with Deloitte, Oct. 6, 2025.
25. Anabelle Yearsdon, "Humanoid robots guide (2025): Types, history, best models, anatomy and applications," Top 3D Shop, April 28, 2025.
26. Steve Goldstein, "300 million humanoid robots are coming - and here are the companies that will benefit," *Morningstar*, June 18, 2025.
27. BMW Group, "Humanoid robots for BMW Group plant Spartanburg," Nov. 9, 2024.
28. *News.am*, "Humanoid robot Fourier GR-1 has been introduced: What is it for?" July 14, 2023.
29. Goldman Sachs, "The global market for humanoid robots could reach US$38 billion by 2035," Feb. 27, 2024; Bank of America Institute, "Humanoid robots 101."
30. Future Today Strategy Group, "2025 tech trends report," accessed Nov. 6, 2025.
31. Matt Swayne, "What is quantum robotics? Researchers report the convergence of quantum computing and AI could lead to Qubots," *The Quantum Insider*, May 9, 2025.
32. Fei Yan, Abdullah M. Iliyasu, Nianqiao Li, Ahmed S. Salama, and Kaoru Hirota, "Quantum robotics: A review of emerging trends," *Quantum Machine Intelligence* 6, no. 86 (2024).
33. Swayne, "What is quantum robotics?"

About the authors

Jim Rowan
jimrowan@deloitte.com

Jim Rowan is the US head of AI at Deloitte and collaborates with external technology organizations, clients, and Deloitte's business leaders to help our clients achieve their AI ambitions. Beyond his client work, Rowan is a principal in Deloitte Consulting LLP. His experience spans the life sciences, health care, and telecommunications industries, with a strong focus on applying analytics, planning, forecasting, and digital transformation to enhance finance functions.

Tim Gaus
tgaus@deloitte.com

Tim Gaus is a principal and the smart manufacturing business leader with Deloitte Consulting LLP. He brings over 25 years of supply chain experience with a focus on value chain optimization using emerging technology. He has led multiple supply chain transformations, spanning supply chain strategy, manufacturing optimization, supply chain planning, inventory optimization, operating model design, and operational excellence for domestic and multinational corporations.

Franz Gilbert
frgilbert@deloitte.com

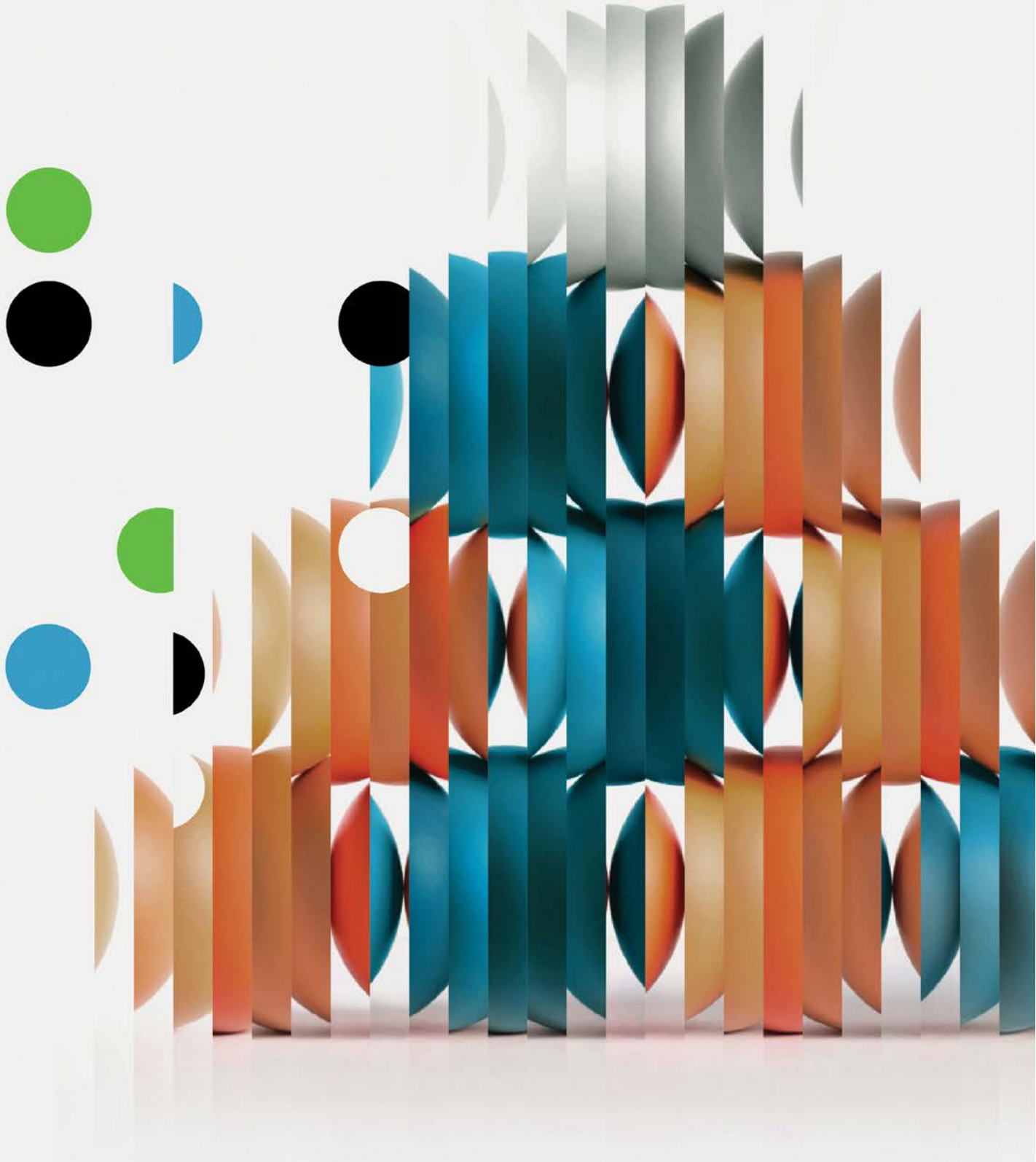
Franz Gilbert is a managing director at Deloitte Consulting LLP, where he is the Human Capital Strategy and Innovation leader, and serves on the Human Capital Management Committee. He and his team are responsible for developing and driving the Human Capital Growth strategy, incubating new and emerging businesses, and stewarding alliances to bring innovative solutions and deliver more valuable outcomes for clients. Gilbert serves on the board of directors for the Human Resource Certification Institute.

Caroline Brown
carolbrown@deloitte.com

Caroline Brown is a senior manager within Deloitte's Office of the CTO. She leads a cross-functional editorial and design production team in developing thought leadership. She serves as the editor of Tech Trends, Deloitte's flagship technology report. A writer and researcher, Brown earned undergraduate and graduate degrees in English and journalism from the University of North Carolina at Chapel Hill.

Acknowledgments

Much gratitude goes to the many subject matter leaders across Deloitte who contributed to our research for this chapter: **Mahesh Chandramouli** and **Ryan Kaiser**.



The agentic reality check: Preparing for a silicon-based workforce

Despite its promise, many agentic AI implementations are failing. But leading organizations that are reimagining operations and managing agents as workers are finding success.

Jim Rowan, Nitin Mittal, Parth Patwari, and Ed Burns

Enterprises are moving quickly toward agentic AI, but many are hitting a wall. They're trying to automate existing processes—tasks designed by and for human workers—without reimagining how the work should actually be done.

Leading organizations are discovering something different: True value comes from resigning operations, not just layering agents onto old workflows. This means building agent-compatible architectures, implementing robust orchestration frameworks, and developing new management approaches for digital workers.

It also means rethinking work itself. As organizations embrace the full potential of agents, not only are their processes likely to change but so will their definition of a worker. Agents may come to be seen as a silicon-based workforce that complements and enhances the human workforce. Getting the fundamentals right—from microservice-based agent architectures to silicon-workforce management—can prepare enterprises for whatever shape the future of workflow automation takes and position them to compete effectively in an agent-native business environment.

Henry Ford put it perfectly: "Many people are busy trying to find better ways of doing things that should not have to be done at all. There is no progress in merely finding a better way to do a useless thing."[1] He was writing about building automobiles in 1922, but he could just as easily have been describing enterprise AI in 2025.

The agent reality check

Agentic AI has captured the attention of enterprises with its compelling promises of autonomous operation and intelligent execution. The momentum is undeniable: Gartner predicts that 15% of day-to-day work decisions will be made autonomously through agentic AI by 2028, up from none in 2024, while 33% of enterprise software applications will include agentic AI by the same timeframe, compared with less than 1% today (figure 1).[2]

Yet despite this enthusiasm, enterprises are encountering significant obstacles in translating agentic pilots into production-ready solutions. Deloitte's 2025 Emerging Technology Trends in the Enterprise study notes that while 30% of surveyed organizations are exploring agentic options and 38% are piloting solutions, only 14% have solutions that are ready to be deployed and a mere 11% are actively using these systems in production. Furthermore, 42% of organizations report they are still developing their agentic strategy road map, with 35% having no formal strategy at all.[3]

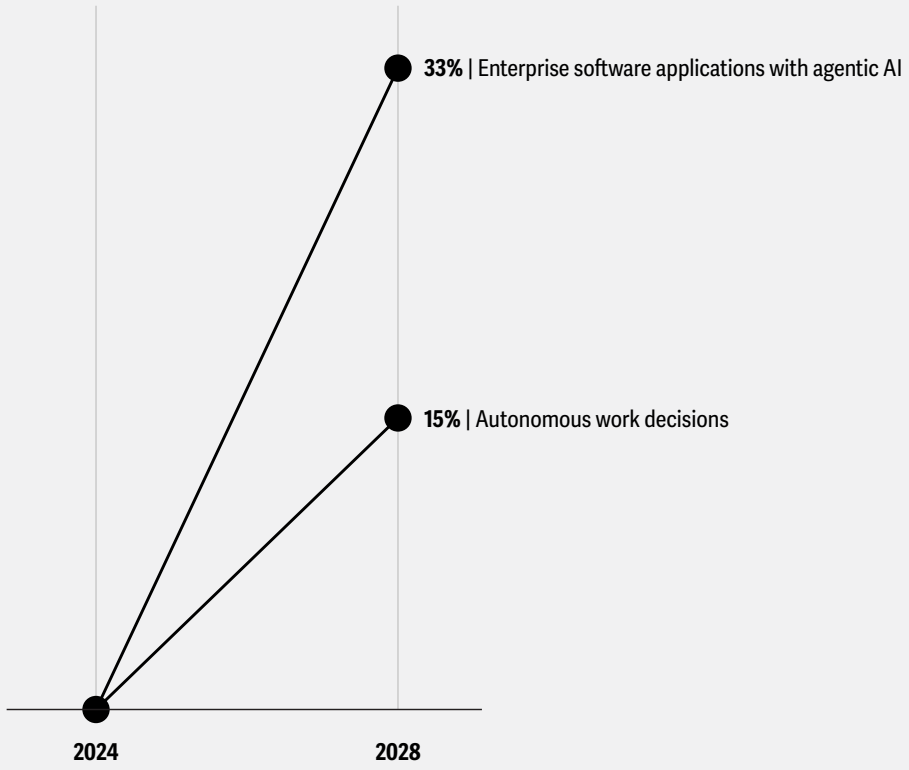
The agentic reality gap

Three fundamental infrastructure obstacles may prevent organizations from realizing the full potential of agentic AI.

Legacy system integration: Traditional enterprise systems weren't designed for agentic interactions. Most agents still rely on application programming interfaces (APIs) and conventional data pipelines to access enterprise systems, which creates bottlenecks and limits their autonomous capabilities. Gartner predicts that over 40%

Figure 1

Projected agentic AI adoption



Source: Gartner analysis. In January 2025, Gartner polled 3,412 webinar attendees about their companies' plans for agentic AI implementation.

of agentic AI projects will fail by 2027 because legacy systems can't support modern AI execution demands. These systems lack the real-time execution capability, modern APIs, modular architectures, and secure identity management needed for true agentic integration.[4]

Data architecture constraints: Current enterprise data architectures, built around extract, transform, load (ETL) processes and data warehouses, create friction for agent deployment. The fundamental issue is that most organizational data isn't positioned to be consumed by agents that need to understand business context and make decisions. In a 2025 Deloitte survey, nearly half of organizations cited searchability of data (48%) and reusability of data (47%) as challenges to their AI automation strategy.[5]

The solution involves a paradigm shift from traditional data pipelines to what can be described as enterprise search and indexing—similar to how Google made the World Wide Web discoverable. This approach involves contextualizing enterprise data through content and index stores built on knowledge graphs, making information discoverable without requiring extensive ETL processes.

Governance and control frameworks: Enterprises struggle to establish appropriate oversight mechanisms for systems designed to operate autonomously. Traditional IT governance models don't account for AI systems that make independent decisions and take actions. The challenge extends beyond technical control to fundamental questions about process redesign: Many organizations

attempt to automate current processes rather than reimagine workflows for an agentic environment.

Additionally, many so-called agentic initiatives are actually automation use cases in disguise. Enterprises often apply agents where simpler tools would suffice, resulting in poor ROI. This "agent washing" compounds the problem, with vendors rebranding existing automation capabilities as "agents."[6] Furthermore, poorly designed agentic applications can actually add work to a process, with some enterprises finding agentic "workslop" can make processes even less efficient.[7]

At their core, AI agents represent a new paradigm in how work gets done, but most enterprises today simply aren't set up to take advantage of the opportunities for automation that agents present. However, we're starting to see signs at leading organizations that these challenges can be surmounted through strategic process redesign, architectural modernization, and new governance frameworks.

The architecture of autonomous operations

Forward-thinking organizations are moving beyond pilot projects to implement systematic approaches for agentic transformation. Their success stems from recognizing that effective agentic AI requires more than deploying individual agents. Instead, it requires thoughtful approaches to integrating agents into systems and workflows, and carefully managing agents once they're rolled out.

Redesigning processes to be agent-native

Leading enterprises don't simply layer agents onto existing workflows. Instead, they redesign processes to leverage the unique strengths of agents. This requires taking a step back and examining end-to-end processes rather than finding automation opportunities within current operations. Agents can handle a range of transactions, communicate with each other, and collaborate to achieve a business outcome, but only when the underlying processes are structured to support these capabilities.

"Now is an ideal time to conduct value stream mapping to understand how workflows should work versus the way they do work," says Brent Collins, head of global SI alliances and former vice president of AI strategy at Intel. "Don't simply pave the cow path. Instead, take advantage of this AI evolution to reimagine how agents can best collaborate, support, and optimize operations for the business."[8]

Most businesses' existing processes were designed around human staff. Agents operate differently. They don't need breaks or weekends. They can complete a high volume of tasks continually. When organizations realize this, the opportunities for process redesign become compelling. That's why enterprises that are succeeding with agentic AI are looking at their processes from end to end.

Enterprise software and services company HPE is developing an AI agent with exactly this kind of process redesign in mind. "We wanted to select an end-to-end process where we could truly transform rather than just solve for a single pain point. We wanted to operate differently," says Marie Myers, executive vice president and chief financial officer.[9]

Her team led the creation of an AI agent called Alfred that helps complete internal operational performance reviews. Myers says the process of conducting the review is very time-consuming, but it's also developed from large data sets, making it ripe for agentic automation. The agent developed by the team consists of an agentic front-end user interface that works with four separate underlying agents. These agents break down queries into multiple elements for processing, conduct data analysis on SQL data, build charts and graphs to present data, and translate AI insights into user-friendly structured reports. The agents pull data from the company's data warehouse, which sits on top of its enterprise resourcing planning and customer relationship management systems.

Myers says she believes the project holds lessons for those outside of her team and even beyond HPE: "That's why we chose this use case, because it applies across functions and industries. We wanted to be able to drive change across the various levels of the organization."

DIGITIZING SKILLS AT SCALE: JOHN ROESE ON USING AI AGENTS TO TRANSFORM BUSINESS PROCESSES

John Roese is the global chief technology officer and chief AI officer at Dell Technologies, where he leads the company's global technology strategy and AI transformation initiatives. With decades of experience in enterprise technology, he focuses on driving practical AI implementation that delivers measurable business value while maintaining rigorous governance and security standards.

Q: What are enterprises missing when it comes to AI agents?

A: If we think of agents as digital skills, their real value emerges when they start operating as a collective. First-generation AI tools, like chatbots and coding assistants, are very good at dealing with single-dimensional processes, like presenting sales information or writing code. But the minute you get into a process that's composite—that doesn't wholly exist within a single domain—agents are the better tool. Agents have the ability to pass context between each other, to reason across boundaries, and to interact over protocols like agent-to-agent.

Most composite processes don't exist solely within the enterprise. Third parties, software vendors, and SaaS providers are part of that workflow. Trustworthy, secure interworking between agents is critical. Otherwise, we can never digitize those processes across boundaries. Most enterprises have barely tapped into applying AI to monolithic singular processes. Imagine the productivity if you apply AI to the composite processes that run your organization.

Q: How are you putting this into practice internally?

A: We now have a dozen agentic proofs of concept, all going after composite problems like quoting or end-to-end remediation of a customer issue across domains, including entitlements, billing, and logistics. We're very focused on ROI. We don't do science projects. We have agentic technology emerging across sales, services, supply chain, and engineering, areas that have a material impact on the company's financial performance.

We've probably tapped into 20 digitized enterprise processes. Before the end of 2025, we will have live autonomous systems that are more than likely working across domains as first-generation tools, which sets us up for a very good year next year to significantly expand the use of agents.

Q: How have you helped the organization think about the cost and infrastructure investments required?

A: In the front end of our process, we require material ROI signed off by the finance partner and the head of that business unit. That discipline has kept experiments as experiments, and production only happens if there is solid ROI.

We also realized that you apply AI to processes, not to people, organizations, or companies. We expect you to be very clear about the processes you're improving.

As we continue to improve, we've become very disciplined in our processes. As a result, we stopped allowing people to design their own AI solutions, and instead, we created an architectural review board that evaluates and approves AI investments and solutions.

Q: Were you already documenting and measuring existing business processes?

A: AI is a process improvement technology, so if you don't have solid processes, you should not proceed. Figure that out first, because otherwise, you'll be guessing where to apply this technology.

We cleaned up our data and gained clarity on the processes we have in place. Without that, we would have been trying to apply AI to something that wasn't quantifiable and might not be accurate.

With this approach in mind, our services organization has digitized every process. We brought all their data together into a single assistant that sits in every digital and human channel to predict the next best action. The result has been double-digit improvements on every metric around cost and customer satisfaction.[10]

Legacy system replacement

When an organization examines its end-to-end processes, it will likely discover workflows that span multiple systems, including legacy software. This has implications for core modernization strategies. As we discussed in last year's Tech Trends report, AI is increasingly able to learn and understand the essential business rules and workflows that define a business' operations. Organizations should carefully consider what constitutes their true core systems and determine whether to use traditional application modernization when agents can effectively bridge legacy system gaps.

At Toyota, teams are using an agentic tool to gain better visibility into the estimated time of arrival of vehicles at dealerships and will soon start using agents to resolve supply issues. The process used to involve 50 to 100 mainframe screens and significant hands-on work from supply chain team members. Now, an agent delivers

real-time information to staff on vehicles from pre-manufacturing through delivery to the dealership, all without anyone having to interact with the mainframe.

Going forward, the team plans to empower agents to identify delays in vehicle shipments and draft emails to try to resolve the issue.

"The agent can do all these things before the team member even comes in in the morning," says Jason Ballard, vice president of digital innovations at Toyota. "We've made that critical decision to just go ahead and invest in this area a bit further. We feel like that's where the differentiator is going to be going forward."[11]

Managing the mixed silicon- and carbon-based workforce

Perhaps the most significant shift when implementing AI agents involves recognizing that agents represent a new form of labor, one that may share some similarities with the human (or carbon-based) workforce. Some organizations are beginning to think beyond using agents as simple automation tools and are starting to explore ways to integrate them with their human workforce.

This evolution represents a fundamental reimagining of what work means, how it's performed, and who performs it. At the heart of this shift is a recognition that AI agents and human workers have different skill sets. While agents excel at defined processes, humans remain essential for navigating the shifting ground of business requirements and complex problem-solving scenarios.

This transformation creates two primary areas that human workers are moving toward.

- **Compliance and governance:** Humans increasingly focus on validation, oversight, and building guardrails for agent operations.
- **Growth and innovation:** They also concentrate on reimagining operations and identifying new opportunities that emerge from agent capabilities.

At insurance company Mapfre, AI agents are used across the organization, including in claims management, where agents handle routine administrative tasks like damage assessments. And when it comes to more sensitive tasks like customer communication, a person is always in the loop. Maribel Solanas Gonzalez, Mapfre's group chief data officer, says she carefully considers which tasks to delegate to agents, ensuring that they are tasks that agents can complete safely and efficiently. Anything that may carry risk still goes through a human worker. This is beginning to change the nature of jobs. The company has published an AI manifesto that prioritizes well-governed, respectful, and safe AI.

"It's hybrid by design," she says. "With the high level of autonomy of these agents, it's not going to substitute for people, but it's going to change what [human workers] do today, allowing them to invest their time on more valuable work."[12]

Other enterprises are going even further. Biotech company Moderna recently named its first chief people and digital technology officer, essentially combining its technology and HR functions. The move was a strategic step to evolve Moderna's operating model by integrating people and technology to accelerate how work gets done.

"The HR organization does workforce planning really well, and the IT function does technology planning really well. We need to think about work planning, regardless of if it's a person or a technology," says Tracey Franklin, chief people and digital technology officer at Moderna.[13]

Specialized vs. broad automation

Successful deployments focus on specific, well-defined domains rather than attempting enterprise-wide automation. Broad automation remains possible but requires multiple specialized agents working in an orchestrated fashion rather than single, monolithic solutions.

Organizations face critical build-versus-buy decisions that often depend on technical maturity and specific use case requirements. Research indicates that pilots built through strategic partnerships are twice as likely to reach full deployment compared to those built internally, with employee usage rates nearly double for externally built tools.[14]

Multiagent orchestration

The first wave of generative AI in the enterprise consisted largely of general-purpose chatbots, which, while

useful as productivity tools, often don't deliver the kind of opportunities to automate that businesses need to drive new efficiencies. With AI agents, organizations can develop highly specialized tools that automatically execute specific tasks. When these specialists are deployed in an orchestrated manner, they can automate entire workflows. This approach is enabled by evolving standards and protocols that facilitate agent interaction.

Model Context Protocol (MCP): Developed by Anthropic, MCP standardizes how AI systems connect to data sources and tools, providing a universal interface for agents to access enterprise resources.[15] While promising, MCP faces limitations in handling complex enterprise security requirements and integrating legacy systems.

Agent-to-Agent Protocol (A2A): Google's protocol enables direct communication between different AI agents across platforms, handling agent discovery, task delegation, and collaborative workflow.[16]

Agent Communication Protocol (ACP): This is an open protocol that enables agents to communicate with each other through a RESTful API, allowing agents to collaborate regardless of the environment in which they were built.[17] ACP may face hurdles due to limitations on the number of agents that can coordinate in a single network and the complexity of integrating with existing enterprise tools.[18]

These protocols represent the foundational layer for what experts describe as a "microservices approach to AI": deploying numerous smaller, specialized agents across various platforms closer to where workflow instructions and data reside. This approach offers several advantages, such as reduced complexity (because smaller agents are easier to debug, test, and maintain); scalable orchestration, where multiple specialized agents can be combined for complex tasks; and platform flexibility that allows agents to run on different systems while maintaining interoperability.

FinOps for agents

As agents operate continuously, poorly configured agent interactions can trigger cascading actions like unpredictable resource consumption and ballooning costs, making cost management critical. Organizations need specialized financial operations frameworks (or FinOps) to monitor and control agent-driven expenses and account for token-based pricing models. These frameworks help track costs in detail through resource tagging, real-time monitoring, automated resource management including autoscaling and rightsizing, and strong governance frameworks to manage AI-specific expenditures.[19]

Five questions to drive agentic AI implementations

As organizations begin their agentic journey, they can consider five strategic questions to help drive their adoption, both now and into the future.

- What agents will be deployed, and what functions will they perform?
- What are the cost profiles relative to human employees?
- Which processes will be automated and at what level of efficiency?
- What will be the optimal mix of human and digital workforce over the next four years?
- Will agents eventually take over entire operational areas beyond the five-year horizon?

Most enterprises ready to implement AI agents today are likely to have prepared answers for the first three questions. However, things get hazier as they consider the latter two. A lot depends on how agentic technology and the underlying generative AI models develop in the future and how this development drives changes in workforce makeup and operational priorities.

Human-digital collaboration drives differentiation

The future enterprise is likely to experience significant changes in the fundamental nature of work, extending beyond traditional carbon-based workforces to include digital agents that autonomously handle entire job functions. As we discussed, companies are already beginning to develop hybrid human-digital workforces. If organizations get this balance right, it may become the primary competitive differentiator in most industries going forward.

The autonomy spectrum

Organizations should define clear boundaries for agent decision-making through graduated autonomy levels, with appropriate human oversight triggers. The autonomy spectrum progresses through three distinct phases.

- **Augmentation**: Today's reality where agents enhance human worker capabilities
- **Automation**: An emerging capability where agents automate tasks within processes defined by humans
- **True autonomy**: A future state in which artificial general intelligence enables agents to work with minimal oversight

Success requires deploying "agent supervisors"—humans who enter workflows at intentionally designed points to handle exceptions requiring their judgment. This isn't simply about checking agents' work, but about strategic handoffs of work at critical decision points. Over the coming years, as AI technology improves, potentially to the point of reaching artificial general intelligence, organizations should be able to let agents work more independently. Leaders should continually assess the state of AI capabilities to ensure they are delegating responsibilities that agents are suited to handle.

HR for agents

As agents mature within job functions, organizations will need equally mature approaches to managing them. This will likely require an entirely new framework for managing agents that not only leans on traditional human resource management concepts for areas where agents share similarities with human workers, but also diverges to account for their unique characteristics. Some areas of focus for HR, such as workplace culture, employee loyalty, and worker motivation, won't be applicable to agents but will remain key pillars of how organizations manage their human staff. Other features of worker management can be extended to apply to agents, even if they look slightly different.

Onboarding: Just as with human workers, agents will require onboarding processes that train them in the enterprise's unique data and operations. At the same time, the human supervisor of the agent should receive training and education on how to leverage the new agents. This will require a new two-pronged approach to onboarding digital labor that prepares both the agent and the human staff for collaboration.

Performance management: This may be one of the areas where managing agents diverges most from traditional human resource management. Organizations will need systems to prove what agents did, why they made specific decisions, and under whose authority they acted. This requires digital identity systems, cryptographic receipts for transactions, and immutable logs for every agent action. As agents roll out across businesses' operations, they will create too much data for human managers to evaluate, which may drive a need for additional agents that manage performance.

Life cycle management: Agents will require ongoing training updates, redeployment to priority areas, and potentially even retirement planning. Organizations are beginning to assign individual names to agents to track productivity contributions, recognizing that digital workers may eventually be subject to taxation similar to human employees.[20]

Zero trust architecture: Implementing ephemeral authentication systems ensures that agent actions are continuously verified and authorized, just as human workers must periodically complete authentication tasks to access enterprise resources.[21]

When calibrated properly, the framework for managing agents will drive strong collaboration between human and digital workers. However, taking the analogy of agents as digital workers too literally may limit the potential of agents. Holding them to standards developed for measuring human performance risks misaligning their activities to functions better left to human workers.

Data as digital exhaust

In an agent-driven environment, systems generate vast amounts of data describing actions taken and outcomes. Today, most agents do not train on their own output data, but in the future, this digital exhaust of silicon workers can become a valuable trove of insights that allow agents to learn and improve. Going forward, the key differentiation lies in how organizations channel this byproduct to reinforce agent learning and capabilities.

This represents a fundamental mindset shift. Every act of inference by agents generates tokens, and those tokens constitute data that can reinforce learning systems. What is likely to matter most in the future is the sophisticated use of this continuous data stream.

The agent-native future

Examining the future of system modernization, early evidence suggests that a hybrid approach is most likely to prevail, where agents extend the useful life of legacy systems, while organizations pursue the selective modernization of critical business processes. This approach allows organizations to realize immediate value from agentic capabilities while maintaining strategic flexibility for future technology decisions.

The transition to agentic AI represents more than technological evolution—it's an organizational transformation that is likely to reshape how enterprises operate, compete, and create value. Organizations that master the foundational elements of agent-native process design, multiagent orchestration, and silicon workforce management will be positioned to thrive in an increasingly automated business environment.

The key to success lies in recognizing that agentic transformation is not about replacing humans with machines, but about creating new forms of human-AI collaboration that leverage the unique strengths of both human and silicon-based workers. The organizations that figure out how to drive this collaboration effectively will define the future of work itself.

THE JAGGED FRONTIER: ETHAN MOLLICK ON AI AGENTS IN THE WORKFORCE

Ethan Mollick is a professor at the Wharton School of the University of Pennsylvania and author of *Co-Intelligence: Living and Working with AI*. A leading voice on the practical applications of AI in business and education, he is known for his research on how organizations can effectively adopt and integrate AI into their operations.

Q: What does the transition from AI as a tool to AI as a workforce look like in practice?

A: Leaders in many organizations aren't clear on what this means. There tends to be a lot of hand-waving and statements like "AI will do stuff" or "you'll manage a bunch of agents." But that doesn't happen without rethinking and redoing the way organizations operate.

I find it's not actually a technology problem. It's a process problem. It means you have to understand the jagged frontier. AI has gotten very good at math and coding, which has an obvious impact on math and coding tasks, but also a less apparent impact on tasks like analysis or meeting with people. Workers will have to adjust their time in their jobs to do different things. It's not that AI agents do everything; they do the basic grunt work, so I can call more organizations to interview them instead. Leaders have to be able to articulate that future.

Q: What do organizations need to consider in terms of agent-first process redesign?

A: You need three things to do AI work: leadership, lab, and crowd. First, you need the crowd: everyone in the organization using these systems. Second, you need the lab, which is actively doing 24/7 experimentation, taking ideas from the crowd, and turning them into real products. And finally, you need aligned leadership. Leaders have to think about organizational design. For example, if you can code 10 times or 100 times faster than you did before, are you still doing Agile development? Agile doesn't work at that speed, so you don't need to be doing it.

Q: What workforce skills are the most important?

A: There's a "using AI" skill that we don't exactly know how to measure or train for yet. It probably involves agency and willingness to experiment, being incentivized properly, and being a subject matter expert in your field.

Q: When do you expect agents to take over operations?

A: I don't know, but agents are already better than people think. True agents are already here. You're just not using them. And you have to build them. But it's doable today. There's no future timeframe. Because you absolutely can build economically valuable agents right now with current technology, and companies are building agentic workflows that do a lot of work autonomously at high accuracy levels. Do they replace all work yet? No, nor do I want them to. But if you're waiting until the technology is more mature, you're going to be in trouble because it's already there.[22]

Endnotes

1. The Henry Ford, "Henry Ford quotations," accessed Nov. 6, 2025.
2. Gartner, Inc., "Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027," press release, June 25, 2025.
3. 2025 Deloitte Emerging Technology Trends in the Enterprise Survey, publication in process.
4. Gartner, Inc., "Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027."
5. 2025 Tech Value Survey by Deloitte Center for Integrated Research, fielded June 2025.
6. Gartner, Inc., "Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027."
7. Bruce Gil, "'Workslop': AI-generated work content is slowing everything down," Gizmodo, Sept. 23, 2025.
8. Deloitte On Cloud podcast interview with Brent Collins, vice president of AI strategy, Intel, Aug. 27, 2025.
9. Marie Myers, executive vice president and chief financial officer, HPE, Deloitte interview, March 1, 2025.
10. John Roese (chief technology officer and chief AI officer, Dell Technologies), interview with Deloitte, Sept. 29, 2025.
11. "Reimagining operations with agentic AI at Toyota," *Deloitte Insights*, Dec. 3, 2025.
12. Maribel Solanas Gonzalez, group chief data officer, Mapfre Insurance, Deloitte interview, June 18, 2024.
13. Tracey Franklin (chief people and digital technology officer, Moderna), interview with Deloitte, Sept. 26, 2025.
14. Aditya Challapally, Chris Pease, Ramesh Raskar, and Pradyumna Chari, "The gen AI divide: State of AI in business 2025," July 2025.
15. Anthropic, PBC, "Introducing the model context protocol," Nov. 25, 2024.
16. Rao Surapaneni, Miku Jha, Michael Vakoc, and Todd Segal, "Announcing the Agent2Agent Protocol (A2A)," Google for Developers, April 9, 2025.
17. AgentCommunicationProtocol.dev, "Welcome," accessed Nov. 6, 2025.
18. Saad Merchant, "ACP: Future of offline AI agent collaboration," Alumio, Oct. 24, 2025.
19. Kearney, "FinOps for AI and AI for FinOps," Jan. 28, 2025.
20. Jake Latimer, "Will AI be taxed? The debate over AI-powered businesses: The 2025 tech-tax tussle," *Medium*, March 13, 2025.
21. Ken Huang, "Agentic AI identity management approach," Cloud Security Alliance, March 11, 2025.
22. Ethan Mollick (professor, Wharton School of the University of Pennsylvania), Deloitte interview, Jan. 1, 2025.

About the authors

Jim Rowan
jimrowan@deloitte.com

Jim Rowan is the US head of AI at Deloitte and collaborates with external technology organizations, clients, and Deloitte's business leaders to help our clients achieve their AI ambitions. Beyond his client work, Rowan is a principal in Deloitte Consulting LLP. His experience spans the life sciences, health care, and telecommunications industries, with a strong focus on applying analytics, planning, forecasting, and digital transformation to enhance finance functions.

Nitin Mittal
nmittal@deloitte.com

Nitin Mittal is a principal at Deloitte and leads Deloitte's global AI program. He advises organizations on AI applications and the implications of emerging technologies on the strategy and competitive positioning of businesses. At Deloitte, Mittal is responsible for shaping the AI market and creating new business opportunities by harnessing emerging technologies. He is also leading Deloitte's own effort to be a global AI-fueled organization and transform how they deliver professional services.

Parth Patwari
ppatwari@deloitte.com

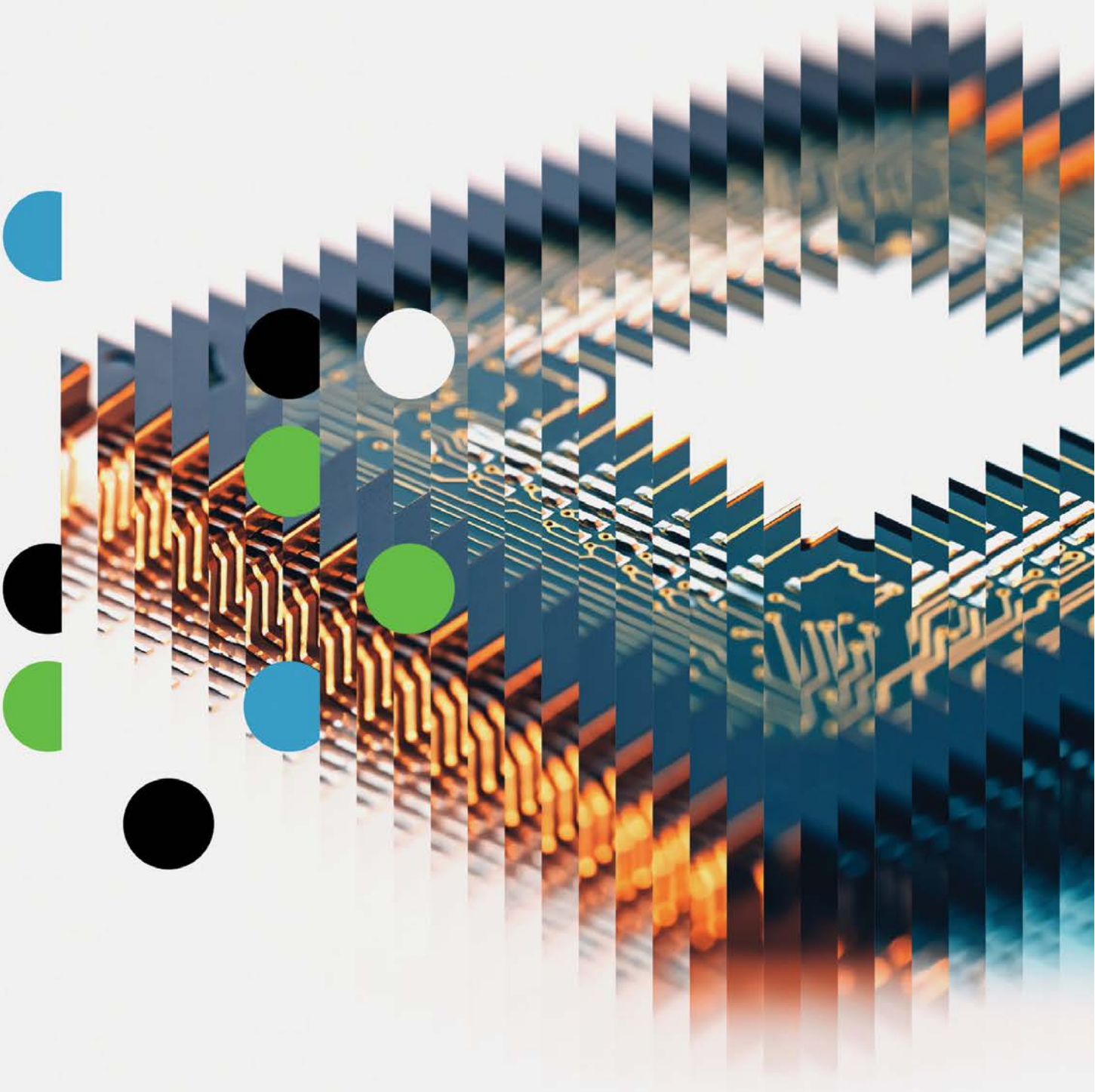
Parth Patwari leads Deloitte's AI and Data practice across all industries in the United States. He has extensive experience working with capital markets and payments institutions to architect, design, and implement large-scale systems that support mission-critical customer, finance, risk, regulatory, and compliance functions. Patwari drives efficiency agendas using AI, analytics, and data management capabilities.

Ed Burns
edburns@deloitte.com

Ed Burns leads the client stories initiative within the Office of the CTO known as Trend Lines. This project serves as a key research input to Tech Trends and other eminence. Prior to his current role he led a tech news publication that covered all things AI, analytics, and data management.

Acknowledgments

Much gratitude goes to the many subject matter leaders across Deloitte who contributed to our research for this chapter: **Jinlei Liu, Baris Sarer, Kate Fusillo Schmidt, Prakul Sharma, Akash Tayal,** and **Ashish Verma.**



The AI infrastructure reckoning: Optimizing compute strategy in the age of inference economics

As AI moves from proof of concept to production-scale deployment, enterprises are discovering their existing infrastructure may be misaligned with the tech's unique demands

Nicholas Merizzi, Chris Thomas, and Ed Burns

When generative artificial intelligence exploded on the scene, businesses got busy dreaming up next-generation products and services. Today, AI has grown up. But as it moves from proof of concept to production-scale deployment, enterprises are discovering their existing infrastructure strategies aren't designed for AI's demands.

Recurring AI workloads mean near-constant inference, which is the act of using an AI model in real-world processes. When using a cloud-based AI service, this can lead to frequent API hits and escalating costs, prompting some organizations to rethink the compute resources used to run AI workloads. But the problem isn't just cost; it's data sovereignty, latency requirements, intellectual property protection, and resilience. The solution isn't simply moving workloads from cloud to on-premises or vice versa. Instead, it's building infrastructure that leverages the right compute platform for each workload.

While exploring AI-optimized infrastructure, organizations will find that advances in chipsets, networking, and workload orchestration can address critical needs across the enterprise. Organizations that act now, addressing both infrastructure modernization and workforce readiness, can define the competitive landscape of the computation renaissance ahead.

The inference economics wake-up call

The mathematics of AI consumption is forcing enterprises to recalculate their infrastructure at unprecedented speed. While inference costs have plummeted, dropping 280-fold over the last two years,[1] enterprises are experiencing explosive growth in overall AI spending.[2] The reason is straightforward: Usage, in the form of inference, has dramatically outpaced cost reduction.

Large language model (LLM) tools based on application program interfaces (APIs) work for proof-of-concept projects but become cost-prohibitive when deployed across enterprise operations.[3] Some enterprises are starting to see monthly bills for AI use in the tens of millions of dollars. The biggest cost contributor is agentic AI, which involves continuous inference, which can send token costs spiraling.

Why organizations are rethinking compute

Rising bills are forcing organizations to reconsider where and how they deploy AI workloads, but there are other factors.

Cost management: Organizations are hitting a tipping point where on-premises deployment may become more economical than cloud services for consistent, high-volume workloads. This may happen when cloud costs begin to exceed 60% to 70% of the total cost of acquiring equivalent on-premises systems, making capital

investment more attractive than operational expenses for predictable AI workloads.[4]

Data sovereignty: Regulatory requirements and geopolitical concerns are driving some enterprises to repatriate computing services, with organizations reluctant to depend entirely on service providers outside their local jurisdiction for critical data-processing and AI capabilities. This trend is particularly pronounced outside the United States, where sovereign AI initiatives are accelerating infrastructure investment.[5]

Latency sensitivity: Real-time AI workloads demand proximity to data sources, especially in manufacturing environments, oil rigs, and autonomous systems, where network latency prevents real-time decision-making. Applications requiring response times of 10 milliseconds or below cannot tolerate the inherent delays of cloud-based processing.

Resilience requirements: Mission-critical tasks that cannot be interrupted require on-premises infrastructure as either primary compute or backup systems in case connection to the cloud is interrupted.

Intellectual property protection: Because the majority of enterprises' data still resides on premises, organizations increasingly prefer bringing AI capabilities to their data rather than moving sensitive information to external AI services. This allows them to maintain control over intellectual property and meet compliance requirements.

Due in part to these factors, companies in many countries are rolling out new data center capacity at an unprecedented rate. Danish property management firm Thylander is building out new data center colocations within the Nordic country, providing the rack space, networking, power, and cooling that cutting-edge graphics processing units (GPUs) and other hardware used in AI workloads require.

Anders Mathiesen, CEO of Thylander Data Centers, says all the hyperscale-sized data centers currently in Demark are owned by foreign companies. But there are growing calls from businesses for more options that will allow their data to be stored and processed by companies owned and operated within the country.

"Looking at data sovereignty and thinking about who actually owns data centers was the start of us saying that we want to do something Danish for Danish companies, but also for external [companies] who think the Danish markets are valuable," Mathiesen says.[6]

The infrastructure mismatch

While enterprises can use these factors to guide future moves, the current state of their infrastructure may create other barriers. Existing data centers feature raised floors, standard cooling systems, orchestration based on private cloud virtualization, and traditional workload management, all designed for rack-mounted, air-cooled servers. The technical specifications of AI infrastructure—from networking requirements between GPUs to advanced interconnecting technologies like InfiniBand—demand architectural approaches that don't exist in traditional enterprise environments.

The AI-optimized data center

Rather than choosing between cloud and on-premises infrastructure, leading enterprises are building hybrid architectures that leverage the strengths of each platform. This approach is a shift from the binary cloud-versus-on-premises thinking that dominated the previous decade.

This physical infrastructure mismatch could become a primary bottleneck as enterprises expand AI adoption. However, forward-looking organizations are beginning to explore the contours of the data center of the future.

The three-tier hybrid approach

Leading organizations are implementing three-tier hybrid architectures that leverage the strengths of all available infrastructure options.

Cloud for elasticity: Public cloud handles variable training workloads, burst capacity needs, experimentation phases, and scenarios where existing data gravity makes cloud deployment a logical choice. Hyperscalers provide access to cutting-edge AI services, simplifying the management of rapidly evolving model architectures.

On-premises for consistency: Private infrastructure runs production inference at predictable costs for

high-volume, continuous workloads. Organizations gain control over performance, security, and cost management while building internal expertise in AI infrastructure management.

Edge for immediacy: Local processing handles time-critical decisions with minimal latency, particularly crucial for manufacturing and autonomous systems where split-second response times determine operational success or failure.

"Cloud makes sense for certain things. It's like the 'easy button' for AI," says AI thought leader David Linthicum. "But it's really about picking the right tool for the job. Companies are building systems across diverse, heterogeneous platforms, choosing whatever provides the best cost optimization. Sometimes it's the cloud, sometimes it's on-premises, and sometimes it's the edge."[7] (See sidebar for the full Q&A.)

HYBRID REALITY CHECK: DAVID LINTHICUM ON RIGHT-SIZING AI INFRASTRUCTURE

Dave Linthicum is a globally recognized thought leader, innovator, and influencer in AI, cloud computing, and cybersecurity. He provides thought leadership, architecture, and technology guidance to Global 2000 companies, new innovative companies, and government agencies.

Q: As enterprises evolve from cloud-first to hybrid models, what challenges will they face and what are the solutions?

A: The biggest challenge is complexity. When you adopt heterogeneous platforms, you're suddenly managing all these different platforms while trying to keep everything running reliably. We saw this happen with multicloud adoption: Companies went from managing 5,000 cloud services to 10,000 services overnight, and they had to run and operate all of that across different platforms.

Rather than managing each platform individually, enterprises need unified management approaches. I don't want to have to think about how my mobile platform stores data differently than my cloud environment or my desktop. You need to push all that complexity down to another abstraction layer where you're managing resources as groups or clusters, regardless of where they physically run.

Otherwise, enterprises handle complexity by hiring specialized teams and buying platform-specific tools. That's expensive and inefficient, and it drains business value because you're scaling complexity management through ad hoc processes instead of thinking strategically. It's really about reducing that operational headache so you can focus on what actually matters to your business.

Q: Deloitte research suggests that when cloud costs reach 60% to 70% of equivalent hardware costs, enterprises should seriously evaluate alternatives. What other tipping points should enterprises monitor when considering the shift from cloud-first to hybrid models?

A: All things being equal between on-premises infrastructure and public cloud, I'm going with public cloud every time because it's easier and gives me scalability and elasticity. But when cloud costs reach 60% to 70% of equivalent hardware costs, you should evaluate alternatives like colocation providers and managed service providers. That's a practical, quantifiable metric that can help you make data-driven choices about infrastructure deployment rather than defaulting to cloud-first strategies regardless of economic considerations.

Q: How does the problem of data center sustainability get solved?

A: At the end of the day, we're not going to stop data center growth. The appetite is huge. I live in Northern Virginia—there are a hundred data centers within 10 miles of where I'm sitting right now. So if we're going to move in that direction, let's try to do less damage by using clean power sources. Nuclear is one of them. That's scary for a lot of people, but there aren't any other options.

I think we'll see small nuclear power plants in data-center-concentrated areas, and everybody's going to pull power off those just like they do now. There's a power plant near me that does nothing but serve data centers—but it's not clean energy. Doing the same thing with more environmentally friendly options while still letting us sustain the business is really the only trade-off we're going to get. We can't increase the grid at any kind of speed that makes it worth the while unless we figure out some sort of power source.

Q: Do you have a hot take on this topic, or a piece of conventional wisdom that you think is wrong?

A: That's easy. Not everything is going to run on GPUs, and we need to get out of that mindset. The GPU hoarding a few years ago was just ridiculous. The reality is that most workloads using AI in ways that actually bring value back to enterprises aren't going to need specialized processors. They're going to run perfectly fine on CPUs. Now, if I'm doing an LLM and huge amounts of training, then yeah, I need specialized processors or else it's going to take 10 years instead of a few months. But those use cases are very few and far between. Most enterprises aren't going to be doing that level of AI work.

Decision framework for compute infrastructure placement

A framework for making compute infrastructure decisions (figure 1) may seem straightforward, but such choices are rarely simple in practice. Everyone wants the fastest hardware running the latest models with the fewest barriers to getting their projects up and running, but this can get expensive.

That's why Dell Technologies recently created an architecture review board. This board evaluates new AI projects and ensures they use consistent tools and the optimal infrastructure based on cost, performance, governance and risks. Dell is currently developing agentic AI use cases across its four core areas, and leaders are looking to expand use cases within these high-ROI areas. As the number of projects grows, leaders say it's critical to ensure they run on appropriate infrastructure. Sometimes that may mean calls to an AI service provider's API, but in other cases it means using entirely on-premises resources.

"Having that architectural rigor is even more necessary now that the resource intensity of these systems is so high," says John Roese, global chief technology and chief AI officer at Dell. "When you start talking about things like reasoning models and agents, and the costs associated with them, having that architectural discipline is critical."[8] (See sidebar for the full Q&A.)

The hardware architecture revolution

This moment presents enterprises with an unprecedented opportunity to move beyond thinking centered on central processing units (CPUs) toward specialized AI-optimized hardware architectures. Organizations are making deliberate decisions about processor deployment, transitioning from general-purpose computing to workload-specific optimization.

The evolution involves integrating multiple processor types within single systems: GPUs for parallel AI processing, CPUs for orchestration and traditional workloads,

Figure 1

A decision framework for computation workload placement

Cloud hyperscalers

Choose when:

- There are variable or unpredictable AI workloads that benefit from elastic scaling
- Experimentation and rapid prototyping phases prioritize speed to market
- Access to the latest AI services and managed platforms provides competitive advantage
- Regulatory flexibility exists regarding data residency requirements

On-premises infrastructure

Build when:

- Consistent, high-volume AI processing justifies capital investment
- Sensitive data requires air-gapped processing behind corporate firewalls
- Latency-critical applications demand immediate response times
- Mission-critical tasks cannot depend on external cloud connectivity
- Long-term strategic AI capabilities warrant deployment commitments of three years or more
- Industries face strict regulatory requirements

Edge infrastructure

Deploy when:

- Real-time decision-making applications demand immediate processing
- Bandwidth-constrained environments make data transmission expensive or impractical
- Local data regulations prevent information from crossing geographic boundaries
- Offline resilience needs exist where connectivity cannot be guaranteed

Source: Deloitte analysis.

neural processing units for efficient inference, and tensor processing units for specific machine learning tasks. Server refreshes now include mixed CPU/GPU configurations. Where once server racks might have had four to eight GPUs on a tray with a CPU coordinator, we're increasingly seeing two GPUs per CPU.

The custom-built AI data center: These trends are coalescing into what we might call the AI data center, which involves a higher number of GPUs relative to CPUs; new server models and orchestration layers for hybrid workloads; evolving data-center form factors that allow for rapid deployment; optical networking between processors for reduced latency; and migration and replatforming of workloads to leverage GPU capabilities.

The rise of AI factories

AI workloads are driving the emergence of "AI factories": integrated infrastructure ecosystems specifically designed for artificial intelligence processing. These environments integrate multiple specialized components into a single solution:

- **AI-specific processors:** GPUs co-packaged with high-bandwidth memory and specially designed CPUs optimized for AI orchestration rather than general computing tasks
- **Advanced data pipelines:** Specialized systems for gathering, cleaning, and preparing data specifically for AI model consumption, eliminating traditional extract, transform, load bottlenecks
- **High-performance networking:** Advanced interconnection technologies to minimize data-transfer latency, including optical networking advancements and specialized GPU-to-GPU communication protocols
- **Algorithm libraries:** Preoptimized software frameworks that align AI functionality with specific business objectives, reducing development time and improving performance
- **Orchestration platforms:** Unified management systems capable of handling multimodal AI workloads across different compute types, enabling seamless integration between various AI technologies

These AI factories can also offer excess computing capacity through service models, allowing organizations to monetize unused processing power while maintaining strategic control over critical workloads.

THE GREENFIELD ADVANTAGE: JOHN ROESE ON PURPOSE-BUILT AI INFRASTRUCTURE

John Roese is global chief technology officer and chief AI officer at Dell Technologies, where he leads the company's global technology strategy and AI transformation initiatives. With decades of experience in enterprise technology, he focuses on driving practical AI implementation that delivers measurable business value while maintaining rigorous governance and security standards.

Q: What are the biggest bottlenecks and challenges related to AI infrastructure?

A: The infrastructure many enterprises have today was designed for the pre-AI era and based on an architectural decision made when the multicloud era began. Businesses made these decisions—like which cloud to use, the topology, what to do on-prem versus off-prem—probably before the pandemic. No one is that smart or lucky to have designed their architecture for a thing that didn't exist when they designed it.

Very quickly, most of the infrastructure capacity will be in service of AI systems, not traditional workloads. AI workloads need accelerated compute; they need a knowledge layer, not just data. These workloads are distributable, and they're in a high-entropy industry with lots of different approaches. That's an entirely different type of workload at the most foundational infrastructure levels.

Q: How does the concept of AI factories figure into this?

A: AI factories are greenfield environments for AI. Retrofitting a brownfield environment designed for traditional enterprise applications and services to properly run an AI environment is difficult. You need an accelerated computing architecture, often not just CPUs [central processing units]. The topology from a networking perspective is simplified, but very high-speed. The storage environment is really a knowledge layer: vector databases, graph databases, knowledge graphs, context-aware chatbots, and data pipelines. And the actual AI applications and their environments are mostly extremely modern.

THE GREENFIELD ADVANTAGE: JOHN ROESE ON PURPOSE-BUILT AI INFRASTRUCTURE, CONTINUED

You could try to retrofit that into your brownfield infrastructure but we recommend building an optimized infrastructure for AI: an AI factory. That will allow you to move much faster, de-risk your architecture, and create an environment purpose-built for AI.

Q: How long does it typically take to spin that up?

A: It's actually faster than retrofitting. We've stood up parallel environments with GPUs and knowledge layers attached to the rest of the infrastructure. A data mesh connects all data together. The AI tools talk to the same data mesh that traditional tools talk to, but the actual physical topology—the infrastructure they live on—is accelerated servers, knowledge layer data management, AI workloads, observability, and controls.

Standing that up as a separate entity was faster than trying to retrofit the existing environment. You can buy an appliance that's basically a collection of storage, computing, and networking preassembled to run the whole agentic AI stack out of the box. By standing it up outside the brownfield, you're able to isolate it from a lot of the complexity, which allows you to move fast.

This might sound like an extra cost, but it really isn't. The operational expense of trying to weave AI capabilities into your existing legacy environment would likely be higher than building a new, dedicated infrastructure.

Q: How does sustainability play into the infrastructure decision-making process?

A: Energy efficiency is a key consideration in the planning process. Innovations in advanced cooling systems, thermal management solutions, and servers can maximize performance per watt while empowering organizations to monitor and reduce their energy use.

One of the biggest choices you have is when to use liquid cooling approaches. For example, direct liquid cooling can be at least twice as energy-efficient as free air cooling, so a single rack with direct liquid cooling can help reduce costs and footprints.

Second, focus on your legacy infrastructure. Poor utilization of IT assets is the largest cause of energy waste in data centers. If you can optimize your legacy infrastructure to reduce waste and increase efficiency, you may be able to reduce the incremental environmental impact of your AI expansion.

Finally, much of the AI workload can be pushed to the client device. An AI PC is a very energy-efficient, distributed computing environment. It's already in the energy grid and exists within your environmental footprint. A significant amount of computing tasks could be moved out of the data center and onto that device. Depending on workload needs, if you distribute parts of the functionality out to these highly efficient AI PCs, you could reduce your overall footprint.

The new frontiers of the data center

The current transformation in AI infrastructure represents only the beginning of a broader computational revolution. Over the next five to 20 years, as emerging computing paradigms mature, data centers will need to continue to evolve to accommodate increasingly specialized tools for specific applications.

Infrastructure evolution continues

Custom silicon integration is accelerating beyond general-purpose chips toward specialized processors designed for specific AI tasks. This includes neuromorphic computing for pattern recognition applications[9] and optical computing for more energy-efficient data processing, which is an increasing focus for AI.[10]

Quantum computing integration is likely to fundamentally alter data-center design requirements once the technology achieves scale.[11] Quantum systems demand specialized infrastructure, including cooling systems, advanced form factors, and extreme noise and temperature sensitivity controls that differ dramatically from current AI infrastructure needs.

Managing this hybrid architecture requires new categories of expertise and management tools. Future orchestration layers may replace legacy solutions with platforms specifically designed for AI workloads. These systems may manage not only traditional virtual machines and containers but also quantum processing units, neuromorphic chips, and optical computing arrays.

Workforce transformation requirements

The infrastructure transformation may require reskilling across IT organizations. Data center teams will likely have to transition from traditional server management to AI-optimized infrastructure operations, GPU cluster management, high-bandwidth networking, and specialized cooling systems.

Network architects face the challenge of designing for AI-first traffic patterns and high-throughput requirements that differ fundamentally from traditional enterprise networking. The networking demands of AI—including GPU-to-GPU communication, massive data-transfer requirements, and ultra-low latency needs—require expertise that many organizations lack.

Cost engineers will need to develop expertise in hybrid compute portfolio optimization, understanding not just cloud economics but also the complex trade-offs between different infrastructure approaches. This includes mastering new financial models that account for GPU utilization rates, inference economics, and hybrid cost structures.

After years of cloud migration have eliminated much internal data center expertise, many organizations struggle to find professionals who understand AI infrastructure requirements. This talent gap represents both a challenge, particularly for businesses that have shifted completely to the cloud, and an opportunity for organizations willing to invest in workforce development.

AI agents managing AI infrastructure

With the growing complexity of AI infrastructure, traditional IT playbooks are likely inadequate for the dynamic optimization required by AI workloads, leading to the emergence of custom-designed AI copilots for IT operations that can summarize alerts, propose root causes, and suggest remediation strategies.[12]

These agents are extending into capacity planning and vendor selection, with services like Amazon Web Services publishing AI patterns that auto-analyze capacity reservations and recommend actions through Amazon Bedrock agents.[13] This represents the precursor to fully autonomous agents that should be able to dynamically juggle model selection, instance-type optimization, spot versus reserved pricing, and multicloud cost and carbon optimization.

Procurement is becoming algorithmic and continuous rather than periodic and manual. Organizations are likely to increasingly rely on AI agents to make real-time infrastructure decisions based on workload demands, cost fluctuations, and performance requirements.

Sustainable data center innovation

The environmental impact of AI infrastructure is driving innovation in sustainable computing approaches. Government and private sector initiatives are exploring nuclear energy to power data centers without carbon emissions, though implementation remains limited to hyperscalers and organizations with substantial capital resources.

Microsoft's Project Natick demonstrated that underwater data center containers could provide practical and reliable computing while using ocean water as a heat sink, though the company ended the research program after completing a concept phase.[14] In contrast, Chinese maritime equipment company Highlander has deployed commercial underwater data center modules and is expanding operations with formal government backing.[15]

Renewable energy integration is accelerating with projects like Data City in Texas, which plans fully renewable energy-powered data center operations with future hydrogen integration capabilities.[16] These initiatives point toward broader trends in sustainable computing infrastructure.

Emerging concepts include orbital data centers that operate on solar power and radiate heat directly into space, eliminating the need for cooling water entirely. Companies like Sophia Space and Lonestar are developing on-orbit compute capabilities, with some achieving initial flight tests of lunar data center payloads, though practical implementation remains years away.[17]

The computation renaissance: AI infrastructure as strategic differentiator

The organizations that successfully navigate this infrastructure transformation are likely to gain sustainable competitive advantages in AI deployment and operation. Those that fail to adapt are likely to face escalating costs, performance limitations, and strategic vulnerabilities as AI becomes increasingly central to business operations.

This AI infrastructure transformation is more than a temporary market adjustment; it's a fundamental shift in how enterprises approach computing resources. Just as cloud computing reshaped IT strategy over the past decade, hybrid AI infrastructure will probably define technology decision-making for the decade ahead.

The computation renaissance has begun, and its outcomes will determine which organizations thrive in an AI-driven business environment.

Endnotes

1. Stanford Institute for Human-Centered AI, "The 2025 AI index report," Stanford University, accessed Nov. 12, 2025.
2. Sarah Wang, Shangda Xu, Justin Kahl, and Tugce Erten, "How 100 enterprise CIOs are building and buying gen AI in 2025," Andreessen Horowitz, June 10, 2025.
3. 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.
4. Thomas, Tayal, Stewart, Kearns-Manolatos, and Parveen, "Is your organization's infrastructure ready for the new hybrid cloud?"
5. Exasol, "The rise of cloud repatriation," Nov. 6, 2024.
6. "A new asset class for Danish RE investment firm: AI-ready, sustainable data centers," *Deloitte Insights*, Dec. 5, 2025.
7. David Linthicum (former chief cloud strategy officer, Deloitte) interview with Deloitte, Sept. 8, 2025.
8. John Roese (chief technology officer and chief AI officer, Dell Technologies), interview with Deloitte, Sept. 29, 2025.
9. National Institute of Standards and Technology, "Introduction to neuromorphic computing, why is it so efficient for pattern recognition, and why it needs nanotechnology," accessed Nov. 12, 2025.
10. Kazuhiro Gomi, "Optical computing: What it is, and why it matters," *Forbes*, Sept. 10, 2024.
11. Christopher Tozzi, "Assessing the state of quantum data centers: Promises vs. reality," *Data Center Knowledge*, Feb. 8, 2024.
12. ServiceNow, "Now Assist for IT operations management (ITOM)," Jan. 30, 2025.
13. Ankush Goyal, Salman Ahmed, Sergio Barraza, and Ravi Kumar, "Optimizing ODCR usage through AI-powered capacity insights," Amazon Web Services, June 5, 2025.
14. Sebastian Moss, "Microsoft confirms Project Natick underwater data center is no more," *Data Centre Dynamics*, June 17, 2024.
15. Peter Judge, "China's Highlander completes first commercial underwater data center, looks for exports," *Data Centre Dynamics*, April 4, 2023.
16. *Fuel Cells Works*, "Energy Abundance unveils Data City, Texas — World's largest 24/7 green-powered data center hub with future hydrogen integration," March 24, 2025.
17. Mandala Space Ventures, "Mandala Space Ventures launches Sophia Space: The world's first scalable data center in space," May 19, 2025; Lonestar Data Holdings, "Lunar data center achieves first success en route to the moon," PR Newswire, March 5, 2025.

About the authors

Nicholas Merizzi
nmerizzi@deloitte.com

Nicholas Merizzi is a principal at Deloitte Consulting LLP and a recognized leader in digital transformation. He is Deloitte's Silicon2Service and AI Infrastructure leader, where he works with organizations to accelerate technology modernization, unlock cloud potential, and integrate AI-driven solutions. Merizzi blends deep infrastructure experience with strategic vision and cloud innovation, guiding clients through the complexities of digital change to achieve their technology goals.

Chris Thomas
chrthomas@deloitte.com

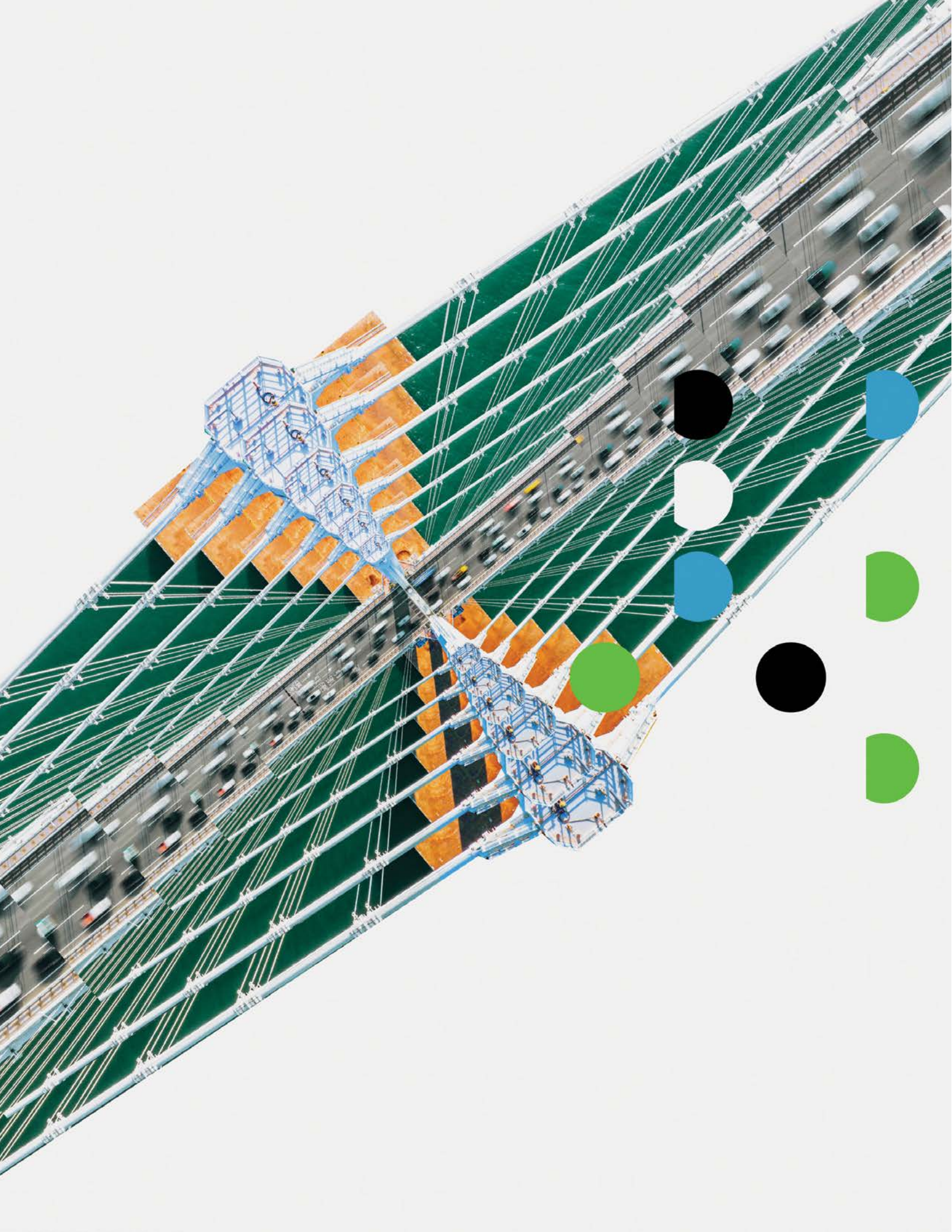
Chris Thomas is a principal in Deloitte Consulting LLP and the US Hybrid Cloud Infrastructure leader. He has over 25 years of strategy consulting and hands-on cloud transformation experience across industries to lead Deloitte's US AI & Engineering business offering for hybrid cloud infrastructure to help clients optimize hybrid cloud strategies and build future-ready organizations. He has extensive experience working with senior executives to enable business outcomes through cloud-centric operating models, large-scale technology transformations, strategic cost optimizations, global outsourcing programs, and workforce of the future initiatives.

Ed Burns
edburns@deloitte.com

Ed Burns leads the client stories initiative within the Office of the CTO known as Trend Lines. This project serves as a key research input to Tech Trends and other eminence. Prior to his current role, he led a tech news publication that covered all things AI, analytics, and data management.

Acknowledgments

Much gratitude goes to the many subject matter leaders across Deloitte who contributed to our research for this chapter: **Bernhard Lorentz**, **Baris Sarer**, **Duncan Stewart**, and **Rohit Tandon**.



The great rebuild: Architecting an AI-native tech organization

What will define tomorrow's tech org—and how can leaders start building it today?

Lou DiLorenzo, Jr., Anjali Shaikh, Michael Caplan, and Erika Maguire

The era of incremental technology change is over. In the span of a few years, artificial intelligence has leapt from automating tasks to dismantling and rebuilding the very structure of the technology organization.[1] Consider that 78% of tech leaders anticipate broad, targeted, or transformational integration of AI agents into architecture workflows over the next five years, according to Deloitte's 2025 Horizon Architecture Survey.[2]

Yet this is more than a shift in tools and headcount. AI is reengineering how technology teams are structured, governed, and led. Tomorrow's model will likely be leaner, faster, and infused with AI at every layer—from architecture to delivery—transforming the tech organization into a dynamic engine that continuously learns and optimizes.

"Agents and people will soon be completely integrated in terms of how work gets done, and it's going to happen really fast—faster than most companies are ready for," says Tracey Franklin, chief people and digital technology officer at Moderna. "Companies need to get better at constant road mapping and iteration because the era of 'build it once and forget it' is over."[3]

While there's no single, definitive blueprint for structuring a tech organization for an AI-driven world, the path forward is coming into view. Tomorrow's high performers won't just keep pace with AI, they'll let it propel them into entirely new terrain. The question for every leader today is not whether AI will transform the tech org, but how quickly they can harness its full potential.

How AI is reshaping the tech organization

Deloitte's Tech Spending Outlook finds that 64% of surveyed organizations plan to increase AI investments over the next two years[4]—a clear sign that leaders recognize the substantial value and transformative potential AI can deliver across the enterprise. While most acknowledge they're still in the exploratory phases of generative and agentic AI (figure 1), data shows how AI is reshaping the tech organization in many ways, from priorities and people to purpose.

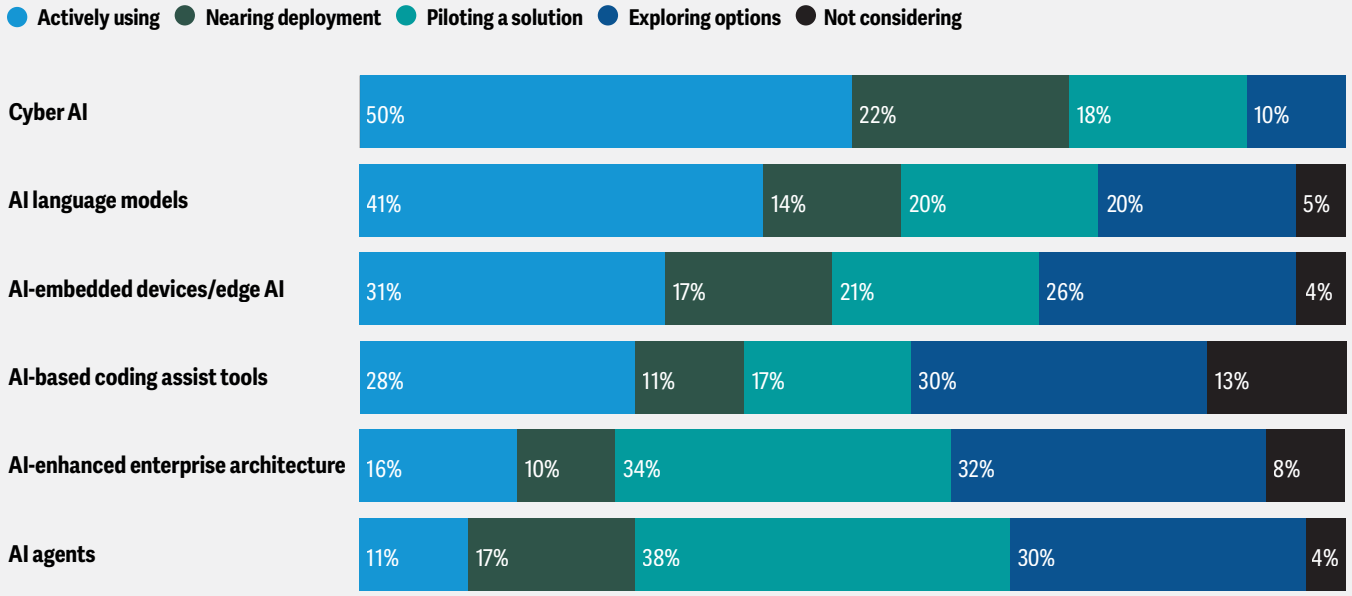
Priorities. Chief information officers (CIOs) in Deloitte's 2025 Tech Executive Survey singled out harnessing the full potential of AI, data, and analytics as the area where they're spending most of their time and energy.[5] While AI has been top of mind for many executives in the past, generative and agentic AI have placed it at the top of the tech organization's agenda—and they're investing accordingly. The percentage of tech budgets allocated to AI is expected to rise significantly over the next two years, from 8% to 13% on average, highlighting how AI is moving from experimentation to core strategy.[6]

People. Nearly 70% of tech leaders from the same survey plan to grow their teams in direct response to gen AI[7]—a clear shift from fears of job loss to a strategy of augmentation and specialization. AI's continued momentum is also creating new roles, like the chief people and digital technology officer at Moderna and forward-deployed engineers, as well as increasing the presence of others.[8] For instance, the number of AI architect roles is expected to almost double, from 30% today to 58% in the next two years.[9] AI is driving new ways of working and is no longer a "plug-and-play" tool but a

Figure 1

While organizations are still piloting AI technologies, broad experimentation trends signal strong momentum to rearchitect the technology function

Q: "What's your organization's current stage of adoption for each of the following technologies?"



Note: May not total 100% due to rounding.

Source: Deloitte 2025 Emerging Technology Trends in the Enterprise Survey. Deloitte conducted an online survey of 500 US technology leaders between June and July 2025 to quantify the prevalence, engagement, and perceptions surrounding the adoption of emerging technologies across industries.

technology that requires thoughtful design, integration, and governance—tasks that demand specialized expertise.

Purpose. As AI plays a key role in CEOs' current and future strategies,[10] the mandate of the tech organization is changing. CEOs today look to tech leaders to drive business strategy, not just run IT. Most large enterprises (66% in Deloitte's Tech Exec Survey) view their tech org as a revenue generator rather than a service center, and when asked what the tech org's role is in shaping the business, the top response from Deloitte's Tech Spending Outlook was "strategic leader: enabling the overall business strategy with a focus on technology."[11] The increasing number of CIOs reporting directly to CEOs (65% in 2025[12] versus 41% in 2015[13]) further signals that technology and AI are not just operational concerns—they're central to growth, innovation, and competitive positioning. The purpose of the tech organization is expanding from "keep the lights on" to "light the way forward."

Strategies to prepare for an AI-powered future

Organizations are actively assessing their tech operating models as AI gains momentum. In fact, when asked how they're evolving their tech op models to meet evolving business demands, just 1% of surveyed IT decision-makers said they had no major changes underway.[14]

The journey to preparing for an AI-driven future will vary depending on organizational maturity and priorities, among other factors, and will likely start with increasing the adoption of AI and automation. Beyond that, here's how organizations are planning for an AI-driven future, today.

Modernization starts with business problems, not technology

Seventy-one percent of surveyed organizations are currently modernizing core infrastructure to support AI implementation, and 23% are investing 6% to 10% of annual revenue in modernizing core enterprise systems.[15] But recognition isn't enough; the key is solving real business problems, not just upgrading technology.

"Modernization is not about technology for technology's sake; it's about addressing fundamental business problems like costs, go-to-market issues, and so on," says Alan Davidson, CIO of Broadcom. "AI is a good example. The technology is evolving at such a rate that conversations taking place about AI today are very different from those that happened six months ago, so it's important to have a tactical plan. Without focusing on a specific business problem and the value you want to derive, it could be easy to invest in AI and receive no return."[16]

Architectures designed for modularity and observability

The AI-driven, future-ready enterprise can't be built on legacy platforms patched together for survival, which may explain why 66% of surveyed organizations are piloting or exploring options around AI-enhanced enterprise architecture.[17] When emerging technologies are used in conjunction with an end-to-end enterprise focus, they can deliver amplified value. New architectures can be designed for modularity and observability,[18] which, at its core, is the ability to see, understand, and optimize a system by analyzing its external outputs.

"At Western Digital, we're developing an observability architecture to help us take a holistic approach to managing our tech landscape," says Sesh Tirumala, CIO of Western Digital. "We're not waiting for perfect AI solutions; we'd rather fail fast on small pilots than miss the wave entirely."[19]

The Coca-Cola Company is also prioritizing a modular approach. "For global organizations, one size rarely fits all. Ways of working are not the same everywhere throughout the world. Our approach has been to build a modular architecture and a set of guiding core principles supported by an agile team able to operate at speed while localizing as needed," says senior vice president and CIO Neeraj Tolmare.[20]

Human–machine collaboration at the heart of tech talent strategy

Recent Deloitte research highlights the rapid evolution of the tech talent landscape. As organizations adopt emerging technologies, the most anticipated new roles include:

- Human-AI collaboration designers, responsible for crafting seamless interactions between people and intelligent systems
- Edge AI and embedded systems engineers, who bring AI capabilities directly to devices and connected infrastructure
- Data quality specialists for synthetic data, ensuring that the training data fueling AI systems is trustworthy and representative
- AI prompt engineers and model trainers, optimizing AI outputs and tailoring models to specific business needs[21]

To make AI work in practice, consider how it can be both an upskilling engine and a tool to bridge knowledge gaps. "Even if you're not a JavaScript expert or a product manager expert, AI can help bridge that gap, or even fill that gap," says Gene Kim, researcher and coauthor of *The Phoenix Project* and the newly released book, *Vibe Coding*.[22] As organizations reconsider their tech talent strategy, it's helpful to think about what degree of functional expertise they'll need on their teams given what AI can do, he adds. That can help illuminate where to focus any upskilling or reskilling efforts. (See sidebar for the full Q&A.)

While AI can democratize capabilities and expertise, tomorrow's competitive edge will likely not simply come from adopting AI tools but from building teams that can design, manage, and evolve the way humans

and machines work together. The future isn't human *or* machine, but rather human *and* machine.[23]

Governance that enables speed while managing emerging risks

While this is not an easy feat, Vince Campisi, chief digital officer and leader of the enterprise services division at aerospace and defense company RTX, shares his strategy for adapting governance in the age of AI: "Break governance down into three M's: map, measure, and monitor. This means teams can map activities to keep tabs on progress, measure results to see if they're achieving the outcomes they want, and monitor quality to make sure the initial goals are realized. Next, focus on tactics designed to maintain alignment with strategic intent. As AI becomes more agentic, organizations can establish governance that starts with leadership's intent and builds in explainability and auditability so humans can verify and trust the results."[24]

VIBE SHIFT: GENE KIM ON AI-POWERED CODING IN ENTERPRISE IT

Gene Kim is a researcher and *Wall Street Journal* bestselling author who studies high-performing technology organizations. A former CTO, Kim is the organizer of the annual Enterprise Technology Leadership Summit. His books have sold over one million copies. His latest book, *Vibe Coding*, was co-authored with Steve Yegge (see the sidebar "From writer to director: Steve Yegge on the software developer's transformation").

Q: How are AI coding tools changing the enterprise landscape?

A: AI is creating what I call "autonomous" teams, where you don't necessarily need deep functional expertise in every area because AI can help bridge or fill those gaps. You might not be a database expert, a business expert, or a product manager, but AI can help you work more independently across these domains.

Any business-domain expert—whether in sales, marketing, or customer support—paired with a developer can now accomplish great things without a lot of oversight. A senior technology leader said to me, "I spent 20 years of my career hearing that I under-delivered, was late, and couldn't keep promises. Now it's the opposite—I'm constantly being told I'm going too fast and need to slow down." That's where all of us want to be.

Q: How should IT leaders prepare their teams beyond just saying, "go learn to code with AI"?

A: The people getting the most success are often more senior, technically minded leaders who understand the limitations but also see the potential. As a leader, you need to set the tone that time can neither be stored nor created, so if something can save us time, we need to use it.

It's interesting that many senior engineers are actually pushing back on AI coding tools. The technology is still janky and unpredictable, so many classically trained coders are resistant, thinking the way they were trained is better. Training is required precisely because the tools are janky. You can't just try once or twice and give up. You need to understand some theory and how they work internally.

One key insight from a recent report on the state of AI-assisted software development sticks out to me: Trust in AI correlates directly with usage frequency and duration. The more you use these tools, the better you understand their quirks and limitations. You start giving them bigger problems, and that's where you see huge payoffs.

Q: What advice do you have for CIOs and IT leaders facing this transition?

A: Leadership will be critical in helping senior engineers who are resistant to seeing the value and [risk getting] left behind. While hiring is down overall, the hiring that is happening will probably favor developers who use AI. From an economic standpoint, you'd choose someone leveraging AI to accelerate their work over someone insisting on writing every line by hand.

Leaders also play a crucial role in [determining] who captures the productivity surplus. If AI isn't discussed openly, people might do a day's work in an hour and not tell anyone. But in a culture where AI practices are shared, that engineer might say, "I did five days of work in an hour—here's how." The value of that knowledge sharing far outweighs the time saved, and the organization captures the benefit.

Q: Do you have any hot takes on what's happening with AI in enterprise IT?

A: Two points that might not be mainstream. First, I believe the days of coding by hand are coming to an end. No one can convince me otherwise.

Second, I don't really care if AI gets much better. Even if AI performance froze at current levels, I'd be incredibly grateful. The leverage you get from existing AI is already miraculous. We don't need major advances for it to be useful—it's already useful. That means there's no reason for any software engineer or leader to wait. Jump in now.

Bold ambition prioritizing reimagination over incremental change

Transforming the tech org demands more than a series of small, safe steps—it requires a courageous vision that reimagines what's possible. Organizations that set bold ambitions harness AI far beyond tactical automation, radically reshaping how technology, talent, and strategy intersect.

"Rather than getting stuck in a cycle of perpetual proofs of concept, consider attacking your biggest problem and go for a big outcome," says Daniel Dines, UiPath CEO and executive chairman. "With a significant success in hand, you can then prove that there are not just opportunities to rethink business processes but also potential productivity enhancements and opportunities to uncover new revenue streams. The sooner you get started, the better your position can be in the journey toward those ends."[25]

Redefining the CIO role

As AI takes hold, the CIO's mandate is expanding from tech strategist to AI evangelist. In fact, 70% of CIOs from the Tech Executive Survey say their primary role with gen AI at their organizations is either implementing gen AI across the enterprise or serving as an evangelist, helping teams see the possibilities of the technology.[26] As AI-enabled capabilities are embedded across organizations and IT is less centralized, CIOs become orchestrators and integrators rather than owners of infrastructure. In fact, almost a third of CIOs say that orchestrating fellow tech leaders is essential in the next 18 months.[27] The role now requires deeper integration with business strategy and enterprisewide transformation, making the CIO both a change agent and a responsible gatekeeper.

"CIOs were once more like chief integration officers because much of their remit was making sure SaaS and other applications worked together effectively. Today, I consider my role a combination of the traditional CIO plus chief data officer, chief AI officer, and chief digital officer," adds Western Digital's Tirumala. "This era is an opportunity for technology leaders to step up. We understand the technology, the data, and the processes. Don't wait for permission—lean in as a partner. Articulate a strong digital ambition and develop a road map for enabling top-line growth and business model shifts, along with a strategy for managing the risks. Focus on speed, agility, outcomes, and value. With the right approach, there won't be any need to ask for forgiveness later on."[28]

The markers of an AI-powered tech organization

Every enterprise's AI journey will be distinct, but successful AI-powered tech organizations share common characteristics. These markers represent the new standard for tech organizations that thrive in an AI-driven world.

AI as a core collaborator

Tomorrow's tech operating models elevate AI from an add-on tool or efficiency play to an embedded collaborator at every layer—from decision-making and operations to product development. As a co-creator, AI can accelerate road mapping, automate feedback loops, and reprioritize work in real time. Much like the revolutions of cloud and mobile before it, this shift positions AI as the next core capability for competitive advantage.

Delivering this vision requires cloud-native, platform-powered foundations. Forty-eight percent of organizations surveyed in the Tech Spending Outlook say they're currently expanding cloud-native and DevOps practices to better align tech with business needs.[29] Cloud is no longer just infrastructure. It's the engine of speed, flexibility, and innovation. Modular, API-first, self-service platforms enable rapid scaling while reducing infrastructure overhead; platform engineering and orchestration ensure consistency, governance, and reuse across product lines. In this model, the tech organization becomes the architect for enterprise AI, providing standardized, secure, and scalable building blocks so teams can adopt AI confidently and consistently.

Work reimagined for speed

In the years ahead, traditional project teams will likely shift into lean, cross-functional squads aligned to products and value streams—tightening the loop from concept to customer and hardwiring ownership of outcomes. Fifty-seven percent of organizations report that they're already shifting from project to product models to bring business and IT closer together.[30] In this model, product

lines deliver user-focused features via shared, customer-facing platforms; agile pods govern ways of working and tool choices; and forward-deployed engineers work alongside product or customer teams to shorten the path to value.[31] The result is stronger ownership, faster iteration, and a clearer line of sight to real-world impact.

AI, cognitive tools, and robotics can amplify this structure by embedding continuous planning, delivery, and experimentation into daily work. Predictive models and smart automation can replace manual handoffs, while roles like AIOps lead emerge and traditional project management fades. Organizational agility can expand beyond IT, creating an operating model that continually adapts to shifting priorities while preserving speed and accountability at the team level.

Human-agent teams at scale

The future workforce fuses human ingenuity with machine intelligence. Two-thirds of organizations are piloting, actively using, or close to deploying AI agents.[32] These future teams will likely be a seamless blend of humans, AI agents, and orchestrators, where humans contribute creativity, oversight, and ethical judgment, and AI brings speed, precision, and pattern recognition. This model fuels perpetual experimentation, rapid prototyping, and scalable innovation across products, services, and operations. As AI agents assume more complex tasks, digital fluency becomes a core skill for every role. The tech organization's future success will likely hinge on orchestrating this collaboration, ensuring that humans and machines learn and evolve together.

Embedded governance

Modern tech organizations are replacing slow, point-in-time oversight with adaptive governance cycles: continuous, AI-assisted mechanisms that protect speed without sacrificing safety. Predictive models and real-time signals are transforming decision-making from subjective, opinion-based guesswork to objective, fact-based choices, surfacing risks before they escalate and informing priorities as conditions change. Policies, processes, and controls become living assets—codified, monitored automatically, and iterated in short cycles to keep pace with emerging technologies—so compliance, security, and ethics are embedded in the flow of work rather than bolted on.

Delivering this at scale requires strong collaboration among leaders. AI outcomes won't emerge from siloed innovation; they're unlocked when the CIO, chief financial officer (CFO), and chief strategy officer (CSO) operate as a cohesive triumvirate, balancing vision, execution, and value realization. In this dynamic, the CIO drives technology integration, the CFO ensures investments deliver measurable ROI, and the CSO aligns strategy with enterprise priorities.[33] Together, they create the connective tissue between innovation and business outcomes, demonstrating that AI success is as much about shared leadership as it is about advanced technology.

Orchestrating ecosystems

The tech organization will likely evolve from service provider to ecosystem orchestrator, coordinating across startups, hyperscalers, regulators, and academia to accelerate innovation. As digital capabilities diffuse across the enterprise and tech-fluent roles become the norm, the boundaries between IT and the business may dissolve. In the years ahead, enterprises will likely operate in fluid innovation networks, running a portfolio of bets and building on what works. Success will depend less on owning all the technology and more on orchestrating an adaptive ecosystem—one that experiments continuously and embraces a "fail fast, learn faster" culture.

Continuously evolving: Always beta by design

The defining trait of tomorrow's tech orgs is perpetual evolution, where change becomes a core capability, not a one-time event. Embedding adaptability and an always-beta mindset into their structure, culture, and strategy creates organizations that learn as fast as the technology they harness.

"The way you've always done things doesn't have to be the way you do them tomorrow," says Kim. "Leverage everything you can get out of AI right now because even if performance levels freeze, what AI can do today for your organization and your teams is still miraculous. There's no time to wait. The time to jump is now."

FROM WRITER TO DIRECTOR: STEVE YEGGE ON THE SOFTWARE DEVELOPER'S TRANSFORMATION

A software engineer with more than 30 years of industry experience, Steve Yegge is the co-author of the book *Vibe Coding* (with Gene Kim; see the previous sidebar "Vibe shift: Gene Kim on AI-powered coding in enterprise IT"). Yegge has written over a million lines of production code in more than a dozen languages and has led multiple teams of up to 150 people each. He's currently an engineer at Sourcegraph, working on AI coding assistants.

Q: How is AI coding affecting the tech function?

A: IT is a layered activity. We're losing the bottom layer, code generation. Tasks or roles continuously get pushed down into hardware or software, and humans get pushed up the ladder. A lot of engineering activity involves design, merging workstreams, and leading teams. Everyone's getting pushed up in that direction because AI is writing code.

It also means nonprogrammers are entering the IT function. Roles like product managers and UX designers are helping with coding because we're using AI to produce these shared artifacts. There's a translation layer between the business and IT that we've never had before. We're seeing small teams—maybe an engineer, a financial analyst, and a marketing person—create software.

Q: How will this change the software engineer's role?

A: You can't trust everything to AI. Eventually, a human needs to look at it. It's like an old-school technical program manager who used to manage teams of engineers, but now you're managing fleets of AI agents. But agents can't solve everything. They can work much faster than a human, but our ambitions will get so much bigger. All the projects that we ever wanted to do, we'll be able to do now, but it will take this constant course-correcting and babysitting and shepherding AI.

As the AI [tools] get smarter, more nonprogrammers will be able to do this [oversight] over the next few years. But right now, it's all about programmers and their ability to be neuroplastic enough to adapt to this entirely new way of working, where they're essentially directing AI.

Q: How do you measure developer productivity in this environment?

A: Companies have been trying to figure this out since AI-powered code completions appeared back in 2022. With code completions, the AI would autocomplete the line of code you were writing, and you accepted or ignored it. The productivity metric was the acceptance rate.

That measure of productivity vanished almost overnight when chat-based coding tools came along, because [now] all you do is make a request in chat, the AI writes the code, and you copy and paste it. It was harder to find good metrics because the improvements were more varied and context-dependent than a simple acceptance rate.

Now we have coding agents, where the AI can use tools to run the code itself, see the results, and iterate without you having to manually copy-paste and relay information back and forth. People who use coding agents are 10 times more productive than people who don't, by any measure that you pick: lines of code, commits, actual business outcomes. It's so obviously an order of magnitude larger than the people who aren't using the coding agents that companies don't even try to measure it. Then the discussion becomes what to do at performance review time when you're trying to compare people who are 10 times more productive than their peers.

Q: What are your thoughts on hiring developers in the AI era?

A: It's a tough time for new entrants into the field, but my take is that people are being overcautious and they're under-hiring. They're standing in the way of having an army of brilliant junior programmers building the next-generation thing that could launch the company to the top of its category.

People who are adaptable and neuroplastic have always been needed, but now they're more important than ever. Hire people who don't have a lot of baggage, not the ones who say, "I won't do X, I won't do Y." Invest in them, train them, and give them the flexibility to make mistakes and learn from them as an organization. Companies that do this are going to be super successful.[34]

Endnotes

1. Kelly Raskovich et al., "IT, amplified: AI elevates the reach (and remit) of the tech function," *Deloitte Insights*, Dec. 11, 2024.
2. Deloitte 2025 Horizon Architecture Survey. From June to July 2025, Deloitte conducted an online survey of 250 US technology leaders across industries to understand the state of technology architectures today and how different approaches drive business value. All respondents were leaders within their organization's IT functions (director level and above) and were from commercial companies with US$1 billion or more in annual revenue.
3. Tracey Franklin (chief people and digital technology officer, Moderna), interview with Deloitte, Sept. 26, 2025.
4. Deloitte Tech Spending Outlook. From June to July 2025, Deloitte conducted an online survey of 302 IT procurement leaders, heads of IT, and non-IT executives with technology spending oversight to understand how US enterprises in key industries are managing technology budgets, making spend decisions, measuring value delivered by technology investments, and planning scenarios based on market dynamics. All respondents were from organizations with US$1 billion or more in annual revenue; 66% of surveyed organizations were publicly owned, and 34% were privately owned.
5. Deloitte US, "New Deloitte Tech Exec Survey spotlights a moment of reinvention for the tech C-suite as need for gen AI skills, cross-functional collaboration becomes critical," June 17, 2025. From March to April 2025, Deloitte conducted an online survey of 622 US technology leaders across industries to understand how senior technology leadership roles and responsibilities are evolving, as well as key challenges and priorities for 2025 and beyond. Survey respondent titles included chief information officer (33%), chief technology officer (18%), chief data analytics officer (25%), and chief information security officer or equivalent (24%).
6. Deloitte Tech Spending Outlook.
7. Deloitte 2025 Tech Executive Survey.
8. Isabelle Bousquette, "Why Moderna merged its tech and HR departments," *The Wall Street Journal*, May 12, 2025; Lee Chong Ming, "An OpenAI exec said the company is using a new engineering role to get big customers' projects going fast," *Business Insider*, July 23, 2025.
9. Deloitte 2025 Horizon Architecture Survey.
10. Benjamin Finzi, Brett Weinberg, and Elizabeth Molacek, "Spring 2025 *Fortune*/Deloitte CEO Survey," Deloitte, May 15, 2025.
11. Deloitte US, "New Deloitte Survey shows tech execs driving growth, shaping strategy, and eyeing the CEO seat;" Deloitte Tech Spending Outlook.
12. Deloitte 2025 Tech Executive Survey.
13. Deloitte CIO Program, "Many tech leaders' influence in the C-suite is growing, new Deloitte research suggests," *Deloitte Insights*, Sept. 26, 2024.
14. Deloitte Tech Spending Outlook.
15. 2025 Deloitte Emerging Technology Trends in the Enterprise Survey.
16. Katherine Noyes, "Broadcom CIO: 'Modernization should be driven by the business,'" *The Wall Street Journal*, Sept. 10, 2025.
17. 2025 Deloitte Emerging Technology Trends in the Enterprise Survey.
18. Michael Caplan et al., "The technology operating model of the future: Rise of the agentic enterprise," *The Wall Street Journal*, Aug. 23, 2025.
19. Katherine Noyes, "Western Digital CIO: In the AI era, 'play offense or get left behind,'" *The Wall Street Journal*, Sept. 6, 2025.
20. Katherine Noyes, "Coca-Cola CIO on scaling AI: From 'what can we do?' to 'what should we do?'" *The Wall Street Journal*, Jan. 18, 2025.
21. 2025 Deloitte Emerging Technology Trends in the Enterprise Survey.
22. Gene Kim (researcher and coauthor of *The Phoenix Project* and the newly released book, *Vibe Coding*), interview with Deloitte, Sept 15, 2025.
23. 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.
24. Katherine Noyes, "RTX CDO on AI: 'Value beats volume every time,'" *The Wall Street Journal*, Sept. 13, 2025.
25. Katherine Noyes, "UiPath CEO: Agentic automation will 'usher in a new era of work,'" *The Wall Street Journal*, Feb. 22, 2025.
26. Deloitte 2025 Tech Executive Survey.
27. Ibid.
28. Noyes, "Western Digital CIO: In the AI era, 'play offense or get left behind.'"
29. Deloitte Tech Spending Outlook.
30. Ibid.
31. Gergely Orosz, "What are forward deployed engineers, and why are they so in demand?" The Pragmatic Engineer, Aug. 12, 2025.
32. 2025 Deloitte Emerging Technology Trends in the Enterprise Survey.
33. Lou DiLorenzo et al., "AI's ROI triumvirate: CIO, CFO, and chief strategy officer," *The Wall Street Journal*, May 10, 2025.
34. Steve Yegge (coauthor of Vibe Coding and software engineer, Sourcegraph), interview with Deloitte, Oct. 1, 2025.

About the authors

Lou DiLorenzo Jr.
ldilorenzojr@deloitte.com

Lou DiLorenzo Jr. leads Monitor Deloitte's US Technology, AI, and Data Strategy practice. With over 25 years of experience across various sectors, DiLorenzo excels in bringing key stakeholders together to drive change, develop new capabilities, and achieve positive financial results for both large corporations and startups. DiLorenzo is a prominent voice in the technology community, frequently quoted in leading publications such as *Forbes*, *Fortune*, and *The Wall Street Journal*.

Michael Caplan
mcaplan@deloitte.com

Michael Caplan is a principal in Deloitte Consulting's Strategy practice and leader of Deloitte's Technology Operating Model Design and Enablement capability. Caplan brings over 20 years of experience advising companies in complex technology and business model transformations with a focus on aligning technology and the broader corporate enterprise to forge future-ready technology strategies, operating models, and ways of working that enhance value and drive organizational growth.

Anjali Shaikh
anjalishaikh@deloitte.com

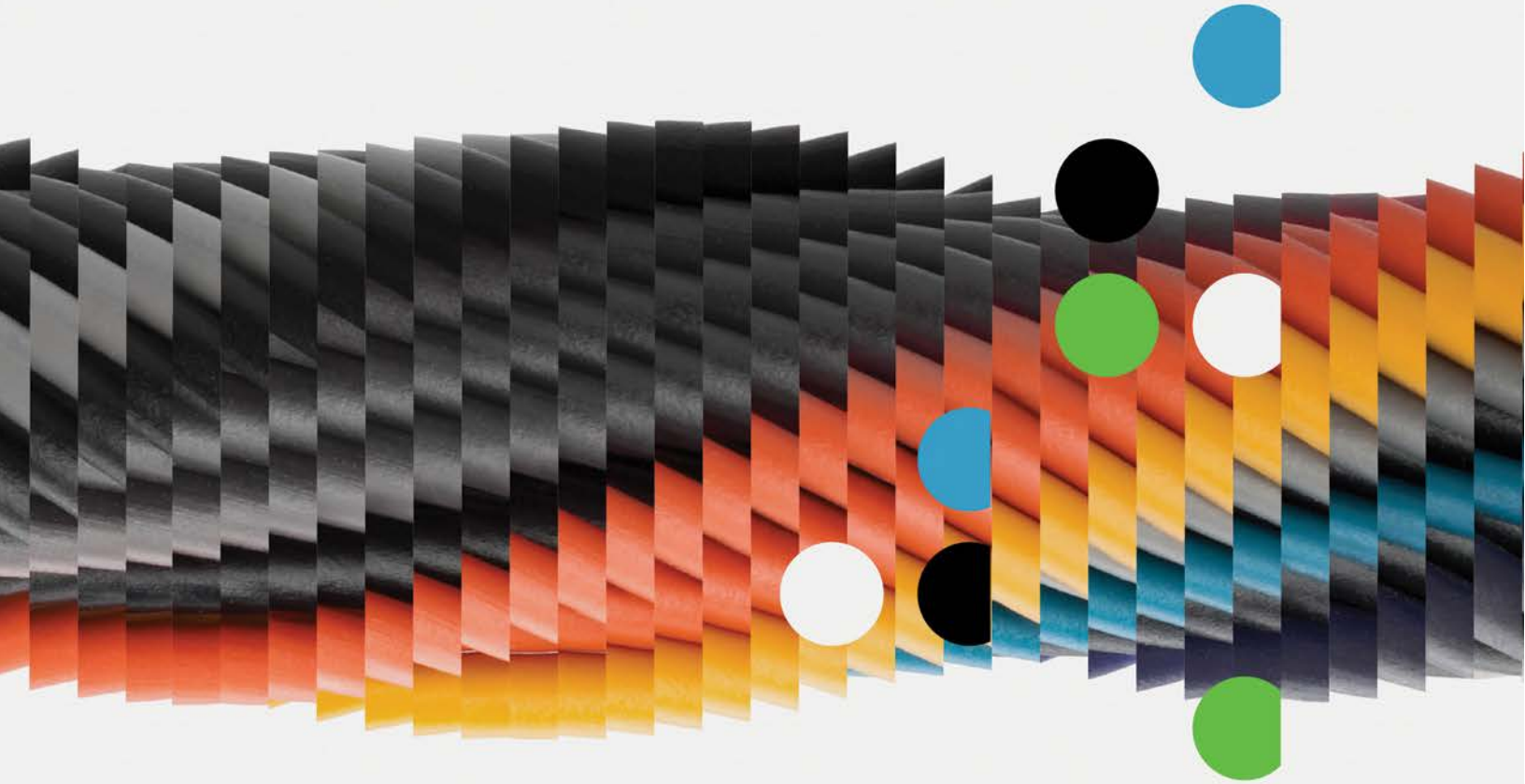
Anjali Shaikh leads Deloitte's technology executive programs, serving as an advisor to CIOs, CDAOs, and tech leaders and providing strategic direction for program development. Shaikh leads a team of skilled practitioners responsible for creating customized experiences and developing valuable insights that help executives navigate complex challenges; shape the tech agenda; build and lead effective teams; and excel in their careers.

Erika Maguire
ermaguire@deloitte.com

Erika Maguire is a researcher and editor focused on uncovering what's new and next in tech. As part of Deloitte's CIO program, she leads key thought leadership initiatives, including Deloitte's Global Technology Leadership Study, and provides clients with actionable insights to build better teams and better businesses. She previously spent nine years at *Forbes* creating data-driven content for today's biggest brands.

Acknowledgments

Much gratitude goes to the many subject matter leaders across Deloitte who contributed to our research for this chapter: **Ryan Casden**, **Nate Paynter**, **Tarun Sharma**, **Tim Smith**, and **Michael Wilson**.



The AI dilemma: Securing and leveraging AI for cyber defense

How can organizations navigate the cybersecurity paradox—getting ahead of the threats introduced by AI, while also recognizing and harnessing its powerful defense capabilities?

Sunny Aziz, Adnan Amjad, Naresh Persaud, Mark Nicholson, and Ed Burns

Organizations that deploy artificial intelligence at scale are discovering its paradox: The same AI capabilities that help them be more competitive can also introduce new security risks. And to add to the paradox, they're also recognizing that AI offers powerful new capabilities to counter the very vulnerabilities it creates.

Enterprises face multiple threats related to artificial intelligence, including shadow AI deployments, AI-accelerated attacks, and the intrinsic risks of AI systems.[1] Yet even as AI drives new threat vectors, traditional cybersecurity principles remain constant and should be applied to autonomous systems that learn, adapt, and operate at machine speed. Many of these techniques will require significant adaptation because most cyber organizations were not designed to lean on digital intelligence.

The window for reactive security approaches is closing. Last year, many organizations focused on mobilizing AI and exploring its possibilities. Now, as they realize the risks of unchecked adoption, they're cataloging emerging threats and implementing targeted governance frameworks that help balance innovation speed with security.

The enemy within

External threats persist. Deepfakes, synthetic personas, and AI-powered social engineering continue to evolve, as we discussed in Tech Trends 2024. But many of today's most pressing AI-related risks originate inside the organization. Two such risks are shadow AI and inadequate controls on agentic AI governance.

Shadow AI, the unsanctioned AI deployment implemented by individual teams across enterprises, creates governance blind spots and introduces autonomous decision-making systems that can access sensitive data, make consequential choices, and interact with other systems.[2] Each deployment represents a potential source of data leakage, model manipulation, model drift, or unauthorized access. Many of the approaches enterprises have developed over the last several years to respond to shadow IT deployments include monitoring the network to discover all applications and developing policies to ensure new deployments comply with privacy and security standards.[3]

Over the past year, enterprises have focused on effectively integrating AI into business process workflows. Now, as they scale AI use cases across operations, they're discovering that AI adoption creates a new set of risks that have corresponding mitigation strategies.

From risk identification to risk mitigation

AI security risks manifest across four domains: data, AI models, applications, and infrastructure. While organizations continue to discover the full scope of threats, the window for reactive approaches is closing. Many existing security practices can be adapted to address these AI-specific risks.

Data security risks: Large language models (LLMs) and other AI systems concentrate vast amounts of

information in single locations and therefore require additional protection. Data security concerns encompass information handled by AI models during training, testing, validation, and inference after deployment (figure 1).

AI model security risks: Model security encompasses the model's architecture and unique training parameters and training, testing, and validation processes (figure 2). Transparency requirements often differ by model type, creating important regulatory considerations.

Figure 1

AI-related data security risks and associated mitigation strategies

Risks	Mitigation
Confidentiality and data privacy: AI tools can inadvertently expose sensitive data. **Training data poisoning:** Adversaries can tamper with training data to compromise the accuracy and utility of outputs. **Model skewing:** This includes deliberate manipulation of training data to create backdoors or systematic biases.	**Secure data management practices:** Catalog AI data sources, maintain high-quality human-generated training data, and carefully manage synthetic data. **Data integrity monitoring:** Continuously monitor data to help ensure it's not manipulated and to detect anomalies. **Robust access controls:** Provide only authorized personnel with access to training data or supporting services, implementing the principle of least privilege.

Source: Deloitte analysis.

Figure 2

AI model security risks and associated mitigation strategies

Risks	Mitigation
Model collapse: AI models trained on synthetic data degrade gradually over time. **Model stealing:** Unauthorized access to proprietary models enables adversaries to identify vulnerabilities and replicate capabilities. **Model inversion:** Model outputs can be used to reconstruct and expose sensitive training data. **Excessive agency abuse:** Generative AI applications gain and use excessive authority to perform unintended actions.	**Model isolation:** Implement clear separation of data and environments across training and deployment. **Privileged access management:** Control and monitor access to trained models through comprehensive identity and access management.

Source: Deloitte analysis.

Application security risks: These risks are related to the external layer hosting the model and sitting on infrastructure, acting as the user interface through which users and systems interact with AI capabilities (figure 3).

Infrastructure security risks: Infrastructure security encompasses the hardware and networking components used for developing and hosting AI systems, representing the foundational layer on which AI capabilities operate (figure 4).

Figure 3

AI application security risks and associated mitigation strategies

Risks	Mitigation
Ethical use concerns: Models mirror human behavior, making AI decisions susceptible to inaccuracies or bias. **Input injection:** Malicious inputs override controls or alter model behavior. **Unauthorized access:** Unauthorized users access AI applications or data.	**Network and user access management:** Large language models require secure enclaves accessible only to authorized users. **Comprehensive access controls:** Control and monitor access to training data, trained models, or supporting services. **Third-party service provider evaluation:** Identify potential partner risks and extend security requirements through the vendor ecosystem.

Source: Deloitte analysis.

Figure 4

AI-related infrastructure security risks and associated mitigation strategies

Risks	Mitigation
Insecure interfaces and APIs: Vulnerabilities can be exploited to attack models or gain information about systems and data. **Model denial of service:** Carefully crafted inputs trigger resource-intensive operations, rendering systems unavailable or increasing costs. **Supply chain vulnerabilities:** Use of third-party data sets, pretrained models, and frameworks can introduce risks that propagate throughout AI systems. **Deployment misconfigurations:** Misconfigurations in hosting environments can lead to unauthorized access, data exposure, or system compromise. **Lateral movement attacks:** Attacks use lateral services or compromised accounts to gain access to AI systems.	**Secure deployment in virtual networks:** Secure sandboxes isolate AI workloads from production environments during testing. **Perimeter and workload hardening:** Reduce breach risk with strict controls like firewalls, network segmentation, and traffic inspection. **Secure machine learning operations integration:** Integrate security into machine learning operations.

Source: Deloitte analysis.

SECURITY FOR AI: SANMI KOYEJO ON THE FUNDAMENTAL SECURITY CHALLENGES OF AI SYSTEMS

Sanmi Koyejo is an assistant professor at Stanford University and the co-founder of Virtue AI, which develops enterprise solutions for AI safety and security. His research on AI evaluation, adversarial robustness, and safety assessment has been implemented in production systems at major technology companies.

Q: Compared with traditional computing systems, why are AI systems difficult to protect?

A: AI systems have very different behaviors, use cases, and scope than much of the computing infrastructure we've seen in the past.

The biggest difference is how much more flexible and contextual AI is compared with classic computing systems. This means many of the tool sets that had been developed and perfected for traditional security are not nearly as effective when applied to AI systems—and even less so when applied to emerging frameworks like agentic systems.

Also, in classic computing, data and compute were siloed, so you could use traditional cybersecurity techniques to separate what's attacking data from what's attacking infrastructure. But in AI systems, the data and the compute are combined, so attacking one often means attacking both.

The emerging AI use cases and the complexity of the threat surface require rethinking what it means to secure computing systems. AI systems have become so good at merging in with standard traffic and looking like human information that many traditional detection strategies fail.

There's a lot of excitement about going beyond language to vision, audio, and other multimodal systems. People engage with audio or video much more than [they do with] text, so they believe things more readily. The risk to stakeholders and to computing system infrastructure grows because of the broader modality set and this ability to engage with different modes.

Q: In terms of security for AI, what approaches are emerging?

A: On the ecosystem front, something very interesting is happening. We've seen two main buckets of companies. They're [taking] highly complementary but different approaches.

First, there are classic cybersecurity companies adding on AI—both in the security-for-AI space and the AI-for-security space. They've been investing in AI to help solve classic security issues, but most interestingly, they're adding security-for-AI capabilities like data sanitation, guardrails, and AI firewalls against prompt injections and other agentic deployment issues.

The other bucket is native AI folks tackling security for AI. It's a fascinating contrast. The approaches look different when you're starting from security and asking how to scaffold to handle AI threats versus starting from native AI infrastructure, where you've built these systems, understand them deeply, understand their vulnerabilities, and exploit that knowledge to think about security infrastructure.

I believe AI-native approaches are likely much more effective for security-for-AI applications. Native AI companies have a special sauce because they understand the systems much better and can be much more targeted than traditional security companies.

Q: Looking two to four years out, what kinds of other attacks or attack vectors can we expect? Or is it impossible to predict because of how fast this is moving?

A: My experience and broad frame of reference is that risks and capabilities tend to go along with each other quite closely. The more the system can do, the more we give it access and agency, and the more we have new kinds of security surfaces to cover. So, if you want to see where risks are going, look at where capability is going—what people are trying to do with it, what use cases people are excited about, and where future investment is going.

With AI, the excitement about [the] technology leads us to ignore security and safety issues, focus on a capability, and then realize we left a gap. We should take another look and figure out what security risks might arise, and treat security and safety as model capabilities as well.[4]

What's old is new again

Most of the practices needed to secure AI deployments aren't new; they're just being updated to address AI risks. Rich Baich, senior vice president and chief information security officer at AT&T, says his approach to mitigating AI risks leans heavily on existing cybersecurity leading practices. In particular, he focuses on enforcing strong software development life cycle approaches. Regardless of whether a tool is homegrown or vendor-supported, each should be tested, red-teamed (see the section "Advanced AI-native defense strategies"), meet architectural requirements, and have access controls in place. Baich says this approach allows his team to bring in the AI tools they need to innovate and push operations forward while ensuring they aren't creating new problems.

"What we're experiencing today is not much different than what we've experienced in the past," Baich says.

"The only difference with AI is speed and impact."[5]

Accelerated attack timelines, shadow AI, and the complexity of managing autonomous agents mean that basic security table stakes—from data cataloging to agent monitoring—have become urgent requirements. But as we'll demonstrate in the following section, AI is also playing a bigger role within cyber, risk, and compliance teams, helping them tackle some of these emerging challenges.

Escalating the AI arms race

AI introduces new vulnerabilities, but it also provides powerful defensive capabilities. Leading organizations are exploring how AI can help them operate at machine speed and adapt to evolving threats in real time. AI-powered cybersecurity solutions help identify patterns humans miss, monitor the entire landscape, speed up threat response, anticipate attacker moves, and automate repetitive tasks. These capabilities are changing how organizations approach cyber risk management.

Advanced AI-native defense strategies

One area where cyber teams are taking advantage of AI is red teaming. This involves rigorous stress testing and challenging of AI systems by simulating adversarial attacks to identify vulnerabilities and weaknesses before adversaries can exploit them. This proactive approach helps organizations understand their AI systems' failure modes and security boundaries.

Brazilian financial services firm Itau Unibanco has recruited agents for its red-teaming exercises. It employs a sophisticated approach in which human experts and AI test agents are deployed across the company. These "red agents" use an iterative process to identify and mitigate risks such as ethics, bias, and inappropriate content.

"Being a regulated industry, trust is our No. 1 concern," says Roberto Frossard, head of emerging technologies at Itau Unibanco. "So that's one of the things we spent a lot of time on—testing, retesting, and trying to simulate different ways to break the models."[6]

AI is also playing a role in adversarial training. This machine learning technique trains models on adversarial examples—inputs designed to fool or attack the model—helping them recognize and resist manipulation attempts and making the systems more robust against attacks.

Governance, risk, and compliance evolution

Enterprises using AI face new compliance requirements, particularly in health care and financial services, where they often need to explain the decision-making process.[7] While this process is typically difficult to decipher, certain strategies can help ensure that AI deployments are compliant.

Some organizations are reassessing who oversees AI deployment. While boards of directors traditionally manage this area, there's a growing trend to assign responsibility to the audit committee, which is well-positioned to continually review and assess AI-related activities.[8]

Governing cross-border AI implementations will remain important. The situation may call for data sovereignty efforts to ensure that data is handled locally in accordance with appropriate rules, as discussed in "The AI infrastructure reckoning."

Advanced agent governance

Agents operate with a high degree of autonomy by design. With agents proliferating across the organization, businesses will need sophisticated agent monitoring to analyze, in real time, agents' decision-making patterns and communication between agents, and to automatically detect unusual agent behavior beyond basic activity logging. This monitoring enables security teams to identify compromised or misbehaving agents before they cause significant damage.

Dynamic privilege management is one aspect of agent governance. This approach allows teams to manage hundreds or even thousands of agents per user while maintaining security boundaries. Privilege management policies should balance agent autonomy with security requirements, adjusting privileges based on context and behavior.

Governance policies should incorporate life cycle management that controls agent creation, modification, deactivation, and succession planning—analogous to HR management for human employees but adapted for digital workers, as covered in "The agentic reality check."

This can help limit the problem of orphaned agents, bots that retain access to key systems even after they've been offboarded.

As AI agents become empowered to spin up their own agents, governance will grow more pressing for enterprises. This capability raises significant questions about managing privacy and security, as agents could become major targets for attackers, particularly if enterprises lack visibility into what these agents are doing and which systems they can access.

The force multiplier effect

Many cyber organizations are using AI as a force multiplier to overcome complex threats. AI models can be layered on top of current security efforts as enhanced defense mechanisms.

AI can assist with risk scoring and prioritization, third-party risk management, automated policy review and orchestration, cybersecurity maturity assessments, and regulatory compliance support. When deployed in these areas, AI capabilities enable security teams to make faster, more informed decisions about resource allocation.

AI is also playing a role in controls testing and automation, secure code generation, vulnerability scanning capabilities, systems design optimization, and model code review processes. This accelerates the identification and remediation of security weaknesses.

The need for AI blueprints

Cybersecurity team operations weren't designed for AI, but business efforts to implement AI throughout the organization create an opportunity to rethink current cyber practices. As businesses roll out AI (and agents in particular) across their operations, many are choosing to completely reshape the workforce, operating model, governance model, and technology architecture. While rearchitecting operations to take advantage of AI agents, organizations should build security considerations into foundational design rather than treating them as an afterthought. This proactive approach to heading off emerging cyber risks can prepare enterprises for today's threats and position them well against dangers that are likely to hit two to five years down the road, which is the subject of the following section.

AI cyber risks will progress, but so will solutions

As we look ahead, emerging trends may challenge fundamental assumptions about cybersecurity, physical security, and even geopolitical stability. While some scenarios remain speculative, understanding potential futures enables organizations to prepare architectures and governance frameworks that can adapt as threats evolve.

When everything is a weapon: AI–physical reality convergence

As AI proliferates across every physical system—power grids, water treatment facilities, transportation networks, supply chains, health care delivery systems—and as AI capabilities improve, physical risks increase exponentially. The convergence of AI and physical infrastructure creates attack surfaces that could lead to unprecedented disruption.

Future threats may involve single attacks corrupting AI systems simultaneously across multiple sectors, including transportation, health care, and utilities. An adversary gaining access to interconnected AI systems could orchestrate cascading failures that compound across critical infrastructure sectors.

Sophisticated attacks could employ "boiling frog" tactics, where AI systems subtly degrade system performance over months, making detection difficult until significant damage has accumulated.

Organizations can prepare for AI–physical convergence risks through several approaches.

- **Automated supply chain vulnerability detection:** Organizations should deploy monitoring tools that constantly check supply chain risks, implementing early warning systems for compromise indicators.
- **Physical system resilience:** Organizations should build backup manual controls for critical physical systems, ensuring that human operators can override automated decisions when necessary.
- **Cascade prevention architecture:** Organizations should create barriers that stop problems from spreading across connected systems, implementing isolation boundaries that contain failures.

Autonomous cyber warfare

The evolution toward autonomous cyber warfare—AI-versus-AI combat with fully automated attack and defense systems operating at machine speed without human intervention—represents a paradigm shift in cybersecurity.

Future attack capabilities may include:

- **Swarm attack coordination:** AI systems could overwhelm defensive systems through coordinated actions that adapt in real time to defensive responses.
- **Adaptive persistent threats:** Attacks could evolve based on defensive measures, learning from each defensive action to identify weaknesses and continuously adjust tactics.
- **Geopolitical dimensions:** The weaponization of AI for cyber warfare creates new geopolitical risks, including the manipulation of public opinion through altered or outright fabricated media, as described in Tech Trends 2024.
- **Economic warfare risks:** Stock markets increasingly rely on AI for trading, risk assessment, and market analysis. Some experts suggest the next financial crisis could be driven by AI rather than by traditional economic factors.[9]

Emerging frontiers: Space and quantum security

As AI security evolves, two emerging frontiers warrant urgent attention despite their nascent development: space-based infrastructure and quantum computing.

Space infrastructure vulnerability: The commercial space industry has opened new attack surfaces; every satellite is essentially a computer vulnerable to exploitation. As adversaries develop capabilities to infiltrate satellites, the potential for disruption extends to GPS, communications, weather monitoring, and a nation's security systems.

Quantum communication channels: Quantum communication promises theoretically unbreakable encryption but also threatens to render current encryption methods obsolete. As discussed in Tech Trends 2025, organizations should prepare for this transition while securing quantum communication infrastructure against adversaries seeking to compromise or control these capabilities.

The imperative for balanced innovation

Organizations should simultaneously pursue both innovation and security through strategic frameworks that embed security into AI initiatives from inception.

Businesses can start by implementing fundamental security controls: data security, access management, model protection, and infrastructure hardening. Skipping these fundamentals in pursuit of rapid AI deployment can create vulnerabilities that may eventually compromise their competitive position.

From there, enterprises may consider investing in advanced AI-powered defense capabilities. Fighting AI threats requires AI-powered security systems that can operate at machine speed, identify subtle attack patterns, and adapt to evolving adversary tactics. The organizations that treat AI security as a force multiplier rather than a cost center are likely to build lasting defensive advantages.

Finally, preparing architectures and governance frameworks for emerging threats will become more important beyond the next few years. While autonomous cyber warfare and AI-physical convergence may seem distant, building adaptable security architectures today helps build organizational resilience tomorrow as the threat landscape evolves.

The AI dilemma is ultimately not a dilemma at all; it's a call to action. Organizations that approach AI security strategically, implementing multiple defense layers while innovating rapidly, can better protect their assets and may be able to establish competitive differentiation through leading risk management capabilities. The future belongs to enterprises that master this balance, treating security as an enabler of AI adoption, not a constraint.

Endnotes

1. Gartner, "Gartner survey reveals gen AI attacks are on the rise," press release, Sept. 22, 2025.
2. CybSafe, "STUDY: Almost 40% of workers share sensitive information with AI tools, without employer's knowledge," press release, Sept. 26, 2024.
3. Dana Raveh, "What is shadow IT?" CrowdStrike, July 10, 2024.
4. Sanmi Koyejo (assistant professor, Stanford University), interview with Deloitte, Sept. 26, 2025.
5. "A no-nonsense approach to secure AI enablement at AT&T," *Deloitte Insights*, Nov. 21, 2025.
6. Roberto Frossard (head of emerging technologies, Itau Unibanco), interview with Deloitte, Sept. 17, 2025.
7. Pat Niemann, "Cyber and AI oversight disclosures: What companies shared in 2025," Harvard Law School Forum on Corporate Governance, Oct. 28, 2025.
8. Deloitte US, "Artificial intelligence: An emerging oversight responsibility for audit committees?" accessed Nov. 11, 2025.
9. John Divine, "How AI could spark the next financial crisis," *US News & World Report*, June 30, 2023.

About the authors

Sunny Aziz
saziz@deloitte.com

Sunny Aziz is a principal in Deloitte's Cyber and Strategic Risk services with over 25 years of experience in helping clients manage, implement, and operate complex cyber programs. He advises clients on cyber strategies and executing large cyber transformation initiatives. Aziz also serves as Deloitte's Financial Services Industry Insurance sector lead for Cyber, specializing in managed security services, cyber strategy and assessments, identity and access management, and more.

Adnan Amjad
aamjad@deloitte.com

Adnan Amjad serves as the US Cyber Leader at Deloitte, overseeing the growth and strategy of Deloitte's Cyber offerings, including Cyber Defense & Resilience, Cyber Operate, Cyber Strategy & Transformation, Digital Trust & Privacy, and Enterprise Security. In this role, Amjad advises clients in navigating the evolving threat landscape through powerful cyber solutions and managed services that aim to simplify complexity and protect and enable businesses to succeed, build resilience, and supercharge transformation—helping them to secure the enterprise of the future.

Naresh Persaud
napersaud@deloitte.com

Naresh Persaud is a principal in Deloitte Risk and Financial Advisory focused on cyber risk across industries. He has over 20 years of experience in identity and access management through multiple roles. Persaud has strong domain knowledge in both identity management and relational database security and experience leading large security implementations and operations across sectors.

Mark Nicholson
manicholson@deloitte.com

Mark Nicholson is a principal in Deloitte Advisory with 25 years of experience in cybersecurity. Prior to its acquisition by Deloitte, Nicholson was the cofounder of the cybersecurity firm Vigilant, Inc. He is currently the AI leader for Deloitte Cyber.

Ed Burns
edburns@deloitte.com

Ed Burns leads the client stories initiative within the Office of the CTO known as Trend Lines. This project serves as a key research input to Tech Trends and other eminence. Prior to his current role, he led a tech news publication that covered all things AI, analytics, and data management.

Acknowledgments

Much gratitude goes to the many subject matter leaders across Deloitte who contributed to our research for this chapter: **Giri Saravanan Chandramohan**, **Edward Guerrero**, **Kieran Norton**, and **Abhishek Sekhri**.

Cutting through the noise: Tech signals worth tracking as AI advances

What are the smaller technology trends—the tremors ahead of seismic shifts? From neuromorphic computing to edge AI, these are areas worth keeping an eye on.

Kelly Raskovich, Bill Briggs, and Caroline Brown

In communications theory, a signal is information that cuts through noise: a pattern that reveals something meaningful about the system transmitting it. In technology, signals are early indicators of directional change—the tremors ahead of seismic shifts. Signals aren't predictions. They're observations of forces already in motion, patterns emerging from the compounding effects described earlier in this report.

The preceding Tech Trends chapters explored five emerging technology trends that are reshaping organizations over the next 18 to 24 months: physical artificial intelligence and robotics, agentic AI, AI infrastructure, tech function transformation, and cybersecurity in the AI age. But the emerging tech landscape has more than five trends. Deciding which to include in our flagship report is an art as much as it is a science, and not without a little intuition.

The signals that follow—some directly connected to our core trends, others operating in parallel— are emerging developments that technology leaders should track. They didn't make the cut for full chapters, not because they lack significance but because they're adjacent to our core themes or still developing. All are worth watching.

Most of them are present-tense phenomena, not speculative futures. Some are already reshaping industries, while others are just beginning to show measurable impact. In a landscape where the distance between "emerging" and

"mainstream" is collapsing, leaders need to know where to direct attention and resources—which developments warrant investment now, which require monitoring, and which dependencies might create risk if ignored.

Are foundation models reaching a plateau? Foundation models—large AI systems trained on massive data sets—face a critical question: Will they keep improving exponentially, or will their capabilities plateau? New models are still improving, but some metrics show they've not delivered the dramatic performance leaps seen in earlier generations.[1] Plus, bigger models drive up energy consumption and computing costs. New scaling approaches, like techniques that allow models more time to process complex problems,[2] could shift us away from "bigger model = better performance." This means current models can improve by optimizing prompting and implementation strategies. How well businesses deploy, fine-tune, and integrate AI into redesigned processes will likely matter more than having the latest foundation model.

New data > synthetic data > old data. As foundation models train on similar publicly available data sets, data itself stops being a competitive advantage. Old data loses value as the world changes. Synthetic data—AI-generated content used to train other AI—helps fill gaps, with predictions that 80% of data used by AI tools will be synthetic by 2028, up from 20% in 2024.[3] However, it has a performance ceiling, typically achieving 90% to 95% of real data quality.[4] Worse, AI trained primarily on AI-generated content can create model collapse, a degenerative process where models lose information about rare patterns, confuse concepts, and eventually produce bland, repetitive outputs.[5] Those with access to fresh information—real-time user interactions, proprietary business data, breaking research—have the advantage. Translation: Companies controlling the interaction layer (search engines, social platforms, AI assistants, smart devices) win.

Neuromorphic chips supercharge computing. Neuromorphic chips are brain-inspired processors that are more energy-efficient than traditional graphics processing units (GPUs) for certain AI tasks. GPUs have separate areas for memory and processing, while neuromorphic chips combine both in the same place. Neuromorphic chips are event-driven—they only process information when something happens—while GPUs run constantly at full speed. This means neuromorphic chips can use 80 to 100 times less energy for AI tasks involving sporadic signals, like analyzing sensor data or processing information in autonomous vehicles, though GPUs remain superior for continuous, high-throughput computation.[6] As AI moves from data centers to billions of edge devices (see next signal), the energy efficiency advantage becomes critical. Widespread adoption of neuromorphic computing is expected by 2030.[7]

The rise of edge AI and on-device processing. Instead of sending data to distant cloud servers, edge AI runs directly on devices—your phone, smartwatch, security camera, or industrial robot. Why it matters: latency (autonomous vehicles can't wait for server responses), privacy (data never leaves your device), exploding costs (cloud bills reaching tens of millions monthly), and internet dependency. Edge AI's potential is reflected in the market for generative AI–capable smartphones, which grew nearly 364% year over year in 2024 to 234.2 million units sold annually, heading toward 912 million by 2028.[8] Real-world applications include smart cameras doing real-time recognition locally, industrial sensors predicting equipment failures, and health wearables monitoring vitals without broadcasting medical data. This is a fundamental shift already in motion.

Will AI-native personal devices and wearables become mainstream? Companies are experimenting with AI-native wearable devices beyond smartphones—pendants that record and transcribe conversations, smart glasses with real-time translation, and screenless pins with voice interaction. The global wearable technology market is projected to reach US$265.4 billion by 2026, and tech giants are investing heavily in next-generation form factors.[9] However, market adoption remains deeply uncertain, and the landscape is littered with failed glasses, pins, and other wearable or pocket-sized form factors.[10] Questions persist about whether consumers actually want separate AI devices or prefer AI integrated into the phones and earbuds they already use. The winning form factor—if one emerges at all—remains undetermined, with success dependent on addressing privacy concerns and delivering functionality that justifies carrying an additional device.

Biometric authentication as next-level cybersecurity. Because AI can replicate voices, forge documents, and mimic behavioral patterns, biometric authentication is

becoming critical for verifying physical presence and identity. As deepfakes and AI-powered fraud grow more sophisticated, organizations are rapidly adopting biometric systems: In one study of chief information security officers, 92% of those surveyed said they have already implemented, are implementing, or plan to implement passwordless authentication.[11] However, biometrics won't be the sole solution. Compromised biometric data cannot be changed like a password, and privacy concerns remain significant. The future points toward hybrid approaches where biometrics serve as the primary but not exclusive verification method.

The AI agent privacy trade-off. Truly capable personal AI assistants require unprecedented access to personal data—and that access is already being granted. To book restaurants, manage schedules, or filter emails effectively, personal AI agents need years of message history, calendar entries, browsing data, stored passwords, credit card information, and intimate personal preferences.[12] But the trade-off is stark: Once personal data is incorporated into AI models, the right to erasure becomes nearly impossible.[13] And of course, security concerns are significant. The public response is already splitting: Some eagerly grant permissions for capability while others resist. But the consent paradox remains. Users must grant extensive permissions to make AI assistants useful, but most don't fully understand the scope or permanence of what they're agreeing to share.

GEO overtakes SEO. Users are increasingly turning to AI chatbots over traditional search engines. The race is on to appear in AI-generated answers—a shift from search engine optimization (SEO) to generative engine optimization (GEO). AI-generated answers already dominate search results across major search engines, reducing click-through rates to conventional websites by more than a third.[14] AI platforms now drive 6.5% of organic traffic, projected to hit 14.5% within a year.[15] GEO differs fundamentally from SEO, prioritizing semantic richness over keywords, author expertise over backlinks, and being cited in AI responses over page views.[16] Just as paid search defined the 2000s and social media advertising dominated the 2010s, AI-generated responses are becoming the most critical marketing channel of the 2020s.

Some of these signals may mature into dominant forces. Others may fade. But all of them reflect the same underlying reality: The pace of technological change has fundamentally shifted. But the speed of adaptation matters more than the certainty of prediction. The organizations that thrive won't be those that predict which signals become trends; they'll be those that build the capacity to sense, evaluate, and respond quickly to what emerges. Those that wait for clarity will find themselves adapting to changes their competitors are already leveraging.

Endnotes

1. Casey Newton, "AI companies hit a scaling wall," Platformer, Nov. 14, 2024.
2. Matthias Bastian, "AI progress in 2025 will be "even more dramatic," says Anthropic co-founder," The Decoder, Dec. 25, 2024.
3. Grant Gross, "Synthetic data takes aim at AI training challenges," *CIO Magazine*, Feb. 19, 2025.
4. Emmett Fear, "Synthetic data generation: Creating high-quality training datasets for AI model development," RunPod Inc., July 31, 2025.
5. IBM, "Examining synthetic data: The promise, risks and realities," accessed Nov. 11, 2025.
6. TokenRing AI, "Neuromorphic computing: The brain-inspired revolution reshaping next-gen AI hardware," WRAL News, Oct. 7, 2025.
7. Research and Markets, "Growth opportunities in neuromorphic computing 2025-2030: Neuromorphic technology poised for hyper-growth as market surges over 45x by 2030," press release, *GlobeNewswire*, April 18, 2025.
8. IDC Research, "Worldwide generative AI smartphone shipments forecast to reach 70% of the market by 2028 with more than 360% growth in 2024, according to IDC," press release, July 30, 2024.
9. *PR Newswire*, "AI-powered wearables transform how consumers interact with everyday technology," Sept. 15, 2025.
10. Amanda Yeo, "Three Products that Flopped in 2024," Mashable, November 28, 2024.
11. Janna Bureson, "Passwordless hits the tipping point in enterprise security," Portnox, Oct. 20, 2025.
12. Mark McCarthy, "The privacy challenges of emerging personalized AI services," Tech Policy Press, May 28, 2025.
13. Zack Whittaker, "For privacy and security, think twice before granting AI access to your personal data," TechCrunch, July 19, 2025.
14. Ryan Law and Xibeijia Guan, "AI overviews reduce clicks by 34.5%," Ahrefs, April 17, 2025.
15. Jake Stainer, "Generative engine optimization (GEO): Complete 2025 guide," Skale, Sept. 30, 2025.
16. Leigh McKenzie, "Generative engine optimization (GEO): How to win in AI search," Backlinko, Oct. 23, 2025.

About the authors

Kelly Raskovich
kraskovich@deloitte.com

Kelly Raskovich is a senior manager and lead within Deloitte's Office of the CTO, and serves as the executive editor of Tech Trends, Deloitte's flagship report on emerging technologies. Her mission is to educate clients, shape the future of Deloitte's technology brand and offerings, cultivate talent, and enable businesses to achieve future growth. She is responsible for technology eminence, client engagement, and marketing/PR efforts.

Bill Briggs
wbriggs@deloitte.com

As the chief technology officer, Bill Briggs helps clients anticipate the impact that emerging technologies may have on their business in the future and how to get there from the realities of today. He is responsible for research, eminence, and incubation of emerging technologies affecting clients' businesses and shaping the future of Deloitte Consulting LLP's technology-related services and offerings. Briggs also serves as executive sponsor of Deloitte's CIO Program.

Caroline Brown
carolbrown@deloitte.com

Caroline Brown is a senior manager within Deloitte's Office of the CTO. She leads a cross-functional editorial and design production team in developing thought leadership. She serves as the editor of Tech Trends, Deloitte's flagship technology report. A writer and researcher, Brown earned undergraduate and graduate degrees in English and journalism from the University of North Carolina at Chapel Hill.

Acknowledgments

Special thanks

Ed Burns, **Preetha Devan**, **Makarand Kukade**, **Erika Maguire**, **Heidi Morrow**, and **Sarah Mortier** for being the engine powering Tech Trends. Ed, your continued dedication to editorial excellence and ability to deftly weave research and insights into compelling narratives have truly elevated our work. Erika, what a great first-time effort—we're so grateful for your research and writing skills, business instinct, sense of humor, and ability to roll with the punches. Heidi, your leadership in design and creative vision have set a standard for visual excellence, bringing our ideas to life in ways that captivate and inspire. Makarand, thank you for jumping in during your first year on Tech Trends and bringing fresh perspective to our supplemental assets and visual materials. Sarah, your leadership in managing the production process has been instrumental in keeping us on track. Your organizational prowess, attention to detail, and collaborative spirit have navigated us through challenges and kept the editorial moving forward. Preetha, we appreciate you bringing your publishing expertise to Tech Trends this year and helping us refine our processes and workflows. We're lucky and thankful that the six of you are part of the team.

Caroline Brown, for leading the Tech Trends editorial and production with steady guidance, strategic insight, and unwavering support. Your leadership has been essential in navigating the complexities of this year's report, and we're grateful for your partnership in bringing Tech Trends to life.

Catarina Pires and **Haley Gove Lamb** for championing Tech Trends and delivering exceptional client experiences. Your dedication to bringing Tech Trends to life for our clients and creating meaningful engagements ensures the report reaches and resonates with the audiences who need it most. Thank you for being such effective ambassadors for our work.

Katarina Alaupovic, **Alison Cizowski**, **Deanna Gorecki**, **Ben Hebbe**, **Bri Henley**, **Abria Perry**, **Mikaeli Robinson**, and **Madelyn Scott** for your tireless dedication and innovative strategies in promoting Tech Trends. Your creativity in marketing, communications, and outreach significantly amplifies our reach and impact year after year. Thank you for your passion and commitment to spreading the value of Tech Trends far and wide.

Amanpreet Arora and **Nidhi John** for the breath of fresh air you brought to the Tech Trends process by pitching in with research, data, and insights. We appreciate your enthusiastic and cheerful willingness to tackle whatever came your way throughout the entire life cycle of the report, from identifying trends to bringing the numbers that back our work.

Raquel Buscaino and **Mark Osis** for being our collaborators as we identified trends and signals and for helping us hone our research craft. Thank you for generously sharing your knowledge and expertise with us.

Diana Kearns-Manolatos and **Duncan Stewart** for your expertise and willingness to share knowledge across teams. Your collaboration has enriched our work and strengthened the connections between our research efforts. Thank you for your generosity and partnership.

Hannah Bachman, **Aditi Rao**, **Elisabeth Sullivan**, and the entire **Deloitte Insights team** for your continued partnership and support as we evolve Tech Trends together. We appreciate your flexibility, strategic guidance, and commitment to excellence as our collaboration deepens and our practice continues to grow.

Sylvia Chang, **Jim Slatton**, **Manya Kuzemchenko**, **Melissa O'Brien**, **Molly Piersol**, **Natalie Pffaf**, **Harry Wedel**, **Jaime Austin**, **Govindh Raj**, **Megha Priya**, and **Naveen Bhusare** for your creativity and dedication in developing the visual assets that bring Tech Trends to life. Your artistic vision and attention to detail create the captivating imagery and graphics that make our report not just informative but truly engaging. We're grateful for your commitment to collaboration and creative excellence.

Continue the conversation

Our insights can help you take advantage of emerging trends. If you're looking for fresh ideas to address your challenges, let's talk.

The Office of the CTO

The Deloitte US Office of the CTO is a team centered on engineering technology futures. We identify, research, and incubate emerging technology solutions to shape demand for future markets, cultivate talent, and enable businesses for future growth.

If you'd like to connect and discuss more, please feel free to contact us at OCTO@deloitte.com.

Executive editor

Kelly Raskovich
Client & Marketing Lead, Office of the CTO
Deloitte Consulting LLP
kraskovich@deloitte.com

Kelly Raskovich is a senior manager and lead within Deloitte's Office of the CTO (OCTO), and serves as the executive editor of Tech Trends, Deloitte's flagship report on emerging technologies. Her mission is to educate clients, shape the future of Deloitte's technology brand and offerings, cultivate talent, and enable businesses to achieve future growth. She is responsible for technology eminence, client engagement, and marketing/PR efforts. Prior to her leadership role, she led several data and analytics projects for global Fortune 500 organizations across the oil and gas industry.

Executive sponsor

Bill Briggs
Global chief technology officer
Deloitte Consulting LLP
wbriggs@deloitte.com

As chief technology officer, Bill Briggs helps clients anticipate the impact that emerging technologies may have on their business in the future and how to get there from the realities of today. He is responsible for research, eminence, and incubation of emerging technologies affecting clients' businesses and shaping the future of Deloitte Consulting LLP's technology-related services and offerings. Briggs also serves as executive sponsor of Deloitte's CIO Program, offering CIOs and other technology executives insights and experiences to navigate the complex challenges they face in business and technology.

Bill earned his undergraduate degree in computer engineering from the University of Notre Dame, and his MBA from the Kellogg School of Management at Northwestern University. He proudly serves on the board of directors for the Kids In Need Foundation, partnering with teachers and students in under-resourced schools and providing the support needed for teachers to teach and learners to learn.

Deloitte.
Insights

Sign up for Deloitte Insights updates at **www.deloitte.com/insights**

Deloitte Insights contributors

Editorial: Aditi Rao, Hannah Bachman, Pubali Dey, Anu Augustine, Cintia Cheong, and Arpan Kumar Saha
Creative: Jaime Austin, Sylvia Chang, Natalie Pfaff, Molly Piersol, and Harry Wedel
Deployment: Atira Anderson
Cover artwork: Jim Slatton

About Deloitte Insights

Deloitte Insights publishes original articles, reports and periodicals that provide insights for businesses, the public sector and NGOs. Our goal is to draw upon research and experience from throughout our professional services organization, and that of coauthors in academia and business, to advance the conversation on a broad spectrum of topics of interest to executives and government leaders.

Deloitte Insights is an imprint of Deloitte Development LLC.

About this publication

This publication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms, or its and their affiliates are, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your finances or your business. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser. None of Deloitte Touche Tohmatsu Limited, its member firms, or its and their respective affiliates shall be responsible for any loss whatsoever sustained by any person who relies on this publication.

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.

Copyright © 2025 Deloitte Development LLC. All rights reserved.
Member of Deloitte Touche Tohmatsu Limited