Tech Trends 2024

Executive summary

Three elevating forces (interaction, information, and computation) and three grounding forces (business of technology, core modernization, and cyber and trust) continue to be the bedrock upon which we build *Tech Trends*, Deloitte's annual exploration of the impact of emerging technologies. In *Tech Trends 2024*, our 15th annual foray, we highlight the stories of pioneering organizations that are ahead of the curve in using new technologies and approaches that stand to become the norm within 18 to 24 months. We also project where the trends could be headed during the coming decade.

Elevating forces

The history of IT has been defined by pioneering advances in interaction, information, and computation, which together serve as an enduring source of innovation.

Interaction

Interfaces in new places: Spatial computing and the industrial metaverse

Augmented and virtual reality for consumer applications have garnered a lot of attention, but these technologies are making their biggest impact in industrial settings. Companies are using the industrial metaverse to power things such as digital twins, spatial simulation, augmented work instructions, and collaborative digital spaces that make factories and businesses safer and more efficient. Factory workers, designers, and engineers are benefiting from immersive 3D interaction—through tested devices such as tablets and experimental ones such as smart glasses—in ways that traditional knowledge workers haven't yet. Accessible, high-fidelity 3D assets are paving the way to an operationalized spatial web, where a digital layer atop reality accelerates ways of working. Eventually, autonomous machines, advanced networking, and even simpler devices can lead to breakthrough spatial web applications, such as remote surgeries or entire factory floors being overseen by a single well-connected worker.

Information

Genie out of the bottle: Generative AI as growth catalyst

Philosophers have long debated whether machines are capable of thought, but generative AI makes the question moot. The underlying operation of these models shares much in common with earlier machine learning tools, but thanks to enhanced computing power, better training data, and clever coding, generative AI technology can imitate human cognition in a number of ways. Regardless of whether it possesses intelligence in the philosophical sense, it does in a practical sense, creating the opportunity for huge productivity and efficiency gains in enterprise settings. Now that machines can behave, comprehend, and narrate like humans, the question becomes how this will impact business and the world broadly.

Computation

Smarter, not harder: Beyond brute force compute

As technology has become a bigger differentiator for enterprises, businesses have built ever-more complex workloads. Typical cloud services still provide more than enough functionality for most business-as-usual operations, but for the cutting-edge use cases that drive competitive advantage, a new need for specialized hardware is emerging. Training AI models, performing complex simulations, and building digital twins of real-world environments require different kinds of computing power. Leading businesses today are finding new ways to get more out of their existing infrastructure and adding cutting-edge hardware to further speed up processes. Soon, some will be looking beyond traditional binary computing entirely.

Grounding forces

Existing systems and investments—represented by the business of technology, core modernization, and cyber and trust—will need to integrate well with pioneering innovations so that businesses can seamlessly operate while they grow.

Business of tech

From DevOps to DevEx: Empowering the engineering experience

As emerging technology is increasingly viewed as a differentiator and crucial part of the business, tech talent is becoming more important than ever. Yet, ways of working are far from efficient: In most companies, developers only spend 30 to 40 percent of their time on feature development. But now, a new focus is emerging for companies that are dedicated to attracting and retaining the best tech talent: developer experience, or DevEx, a developer-first mindset that aims to improve software engineers' day-to-day productivity and satisfaction by considering their every touchpoint with the organization. In the years to come, DevEx can lead to a future of integrated, intuitive tools that enable citizen developers across the business to drive tech value.

Cyber and trust

Defending reality: Truth in an age of synthetic media

With the proliferation of AI tools, it's now easier than ever for bad actors to impersonate and deceive their targets. We're seeing deepfakes being used to get around voice and facial recognition access controls. They're also

being used in phishing attempts. Security risks are multiplying with every new content-generation tool that hits the internet. However, leading organizations are responding through a mix of policies and technologies designed to identify harmful content and make their employees more aware of the risks.

Core modernization

Core workout: From technical debt to technical wellness

After years of investments in once-cutting-edge technologies, companies are grappling with an expanded set of core technologies, including mainframes, networks, and data centers, that are in dire need of modernization. Those that want to lead in the future need to forgo piecemeal approaches to technical debt for a new holistic frame of technical wellness. Preventative wellness assessments, rooted in business impact, can help teams prioritize which areas of the tech stack need treatment and which can continue serving IT's needs. In the years to come, companies are likely to develop a highly customized and integrated wellness plan across the tech stack, including investments in self-healing technologies that reduce tomorrow's modernization needs.

INTRODUCTION

Generative AI: Force multiplier for human ambitions

Last year, our team of futurists and researchers decided to use generative artificial intelligence (AI) to create the cover and chapter art in *Tech Trends 2023*. The result was nothing short of spectacular. Yet our exacting design standards required significant human collaboration and intervention in the generation process. On the heels of that successful experiment and the subsequent launch of ChatGPT and ensuing generative AI mania, we decided to explore the use of AI-generated text to help write the introduction of this year's *Tech Trends*. As with last year's artwork, substantial human intervention was required, supporting our point that in the era of artificially intelligent machines, humans are more important than ever.

As someone who's spent a quarter-century up to my eyeballs in all things newfangled, I want to provide some additional perspective on the current excitement around generative AI and frame this breakthrough technology within the context of our enduring macro technology forces.

Tech evolution, business revolution

First, while generative AI feels at once unprecedented and revolutionary, the technology itself is actually a surprisingly straightforward evolution of machine intelligence capabilities that we've been tracking and chronicling since *Tech Trends'* inception. Organizations have employed mechanical muscles (industrial robotics) for nearly 70 years, and mechanical minds (machine learning systems) for the last 25. That our inorganic colleagues can now paint a picture, write a product description, or sling Python is neither random nor unexpected—they're the next page in a book that future computer scientists might one day call *Cognitive Automation: The Early Years*. Indeed, the best companies have been engaged in this quest to reduce the cost of decision-making for at least the last 15 years (figure 1).

Technologically, generative AI is simply the next chapter in the ongoing history of information. But on the business side, the hyperbole is very much warranted. Make no mistake: The newfound opportunity to augment productive professionals with silicon-based intelligence is indeed a generational business opportunity. It's a full-on paradigm shift that is poised to unlock the doors to altogether-new business opportunities and fundamentally change how the enterprise itself organizes and operates.

You can't shrink your way to success

In my recent experience, far too many business leaders see generative AI as a mere weight loss pill—a quick and dirty means to simply reduce costs by automating and, in turn, eliminating jobs. Nipping and tucking at business cost centers is a short-term approach to pleasing shareholders, taxpayers, and other key constituents—but in the final calculus, you can't shrink your way to success. To be sure, B-school textbooks are rife with cautionary tales of once-great organizations that, seduced by the allure of automation and outsourcing, found themselves leaner, meaner, and, as a result, squarely in the crosshairs of competitors or acquirers.

Instead, generative AI should be considered rocket fuel for elevated ambitions. Virtually every C-level leader I meet tells me, in their own vivid way, how the intensity of their present demands precludes them from paying as much attention as they'd like to future ambitions. "Operations eats innovation for lunch," one chief technology officer (CTO) told me, in a spin on the famous Peter Drucker-ism "Culture eats strategy for breakfast." AI (traditional and generative alike) can free up precious human cycles from mundane operations and allow people to focus, finally, on higher-value work that better aligns with tomorrow's business imperatives—namely, new and improved products, services, experiences, and markets (in other words, the time-tested keys to profitable growth).

Wanted: Generative humans

Many worry that generative AI reduces the need for (or perhaps more accurately, diminishes the worth of) human creativity. I've observed the opposite is true: In an age of creative machines, creative humans matter more than ever.

For example, late last year, I gathered with a room full of C-suite executives to demonstrate a then-new generative AI tool that painted unique images based on text prompts. One of the attendees asked the tool, "Show me a sunset." The resulting picture was fine but unremarkable; the attendee shrugged and dismissed it as "just a sunset." Undeterred, another participant took her turn, prompting the tool, "Show me a war between

Figure 1

A brief history of information

TIME (years)	t–175	t–75	t–50	t–25	t–10	t	t+10	t+n	t=∞
	First computer design	First digital computer	Mid-20th century	Late 20th century	Early 21st century	*Today*	Horizon next	Furthest stars	Endgame
Information	Store	**Arithmetic calculation**	**Relational databases**	**Descriptive analytics**	**Predictive analytics**	**Cognitive automation**	**Exponential intelligence**	**General-purpose AI**	**Intelligence**

Source: Deloitte Technology Futures Report 2021.

pretzels and cheeseballs on Mars where the pretzels have nunchucks and the cheeseballs have squirt guns." The image generator produced an absurd, delightful image that made the room full of executives applaud and marvel. Most (understandably) celebrated the "miraculous machine" that rendered the image, but I couldn't help but quietly acknowledge the clever human with the magical mix of mind and moxie to even ask for such a thing. With generative AI as a force multiplier for imagination, the future belongs to those who ask better questions and have more exciting ideas to amplify.

As generative machines continue to find purchase in the many nooks and crannies of our professional lives, *people* will determine whether these tools scale with magic or mediocrity. With mindful and imaginative guidance, generative AI stands to unlock a world of magical new business possibilities. Without it, we run the risk of scaled mediocrity—or worse. As my friend and Deloitte's global CTO Bill Briggs likes to say, "Good does not come from making bad things faster."

Eyes to the skies, feet firmly on the ground

Finally—and this is a big one—none of this works without a solid technology foundation. We geeks (ahem, professional technologists) are typically well aware of the old trope "garbage in, garbage out." Our early forays into our shared AI future suggest that going forward, the experience will be more akin to "garbage in, garbage squared." Small biases in training data can beget cataclysmic biases in AI output—so get your enterprise data in order first.

And remember: *Information* is just one of the six macro technology forces that drive business (figure 2).

A creaky core in desperate need of modernization will buckle under tomorrow's AI-fueled workloads. An undifferentiated computation strategy will increasingly break the bank. Cumbersome interaction modalities will muddy your message, to say nothing of disengaged talent, or worse, cyberthreats. If you take anything from this year's report, it's this: Don't become so blinded by the buzz around generative AI that you neglect the five other fundamental forces.

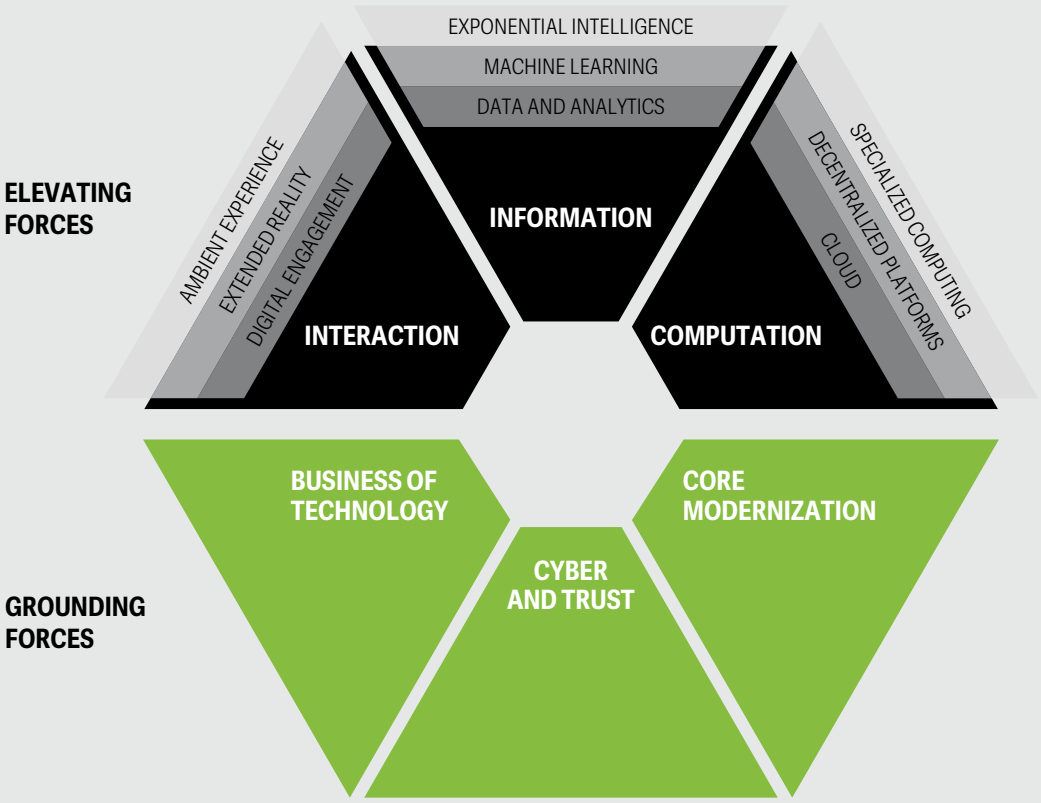
Indeed, AI matters more than ever, but this does not mean that everything else you've been working on suddenly *doesn't*.

Onward,
~!mb (with a little help from generative AI)

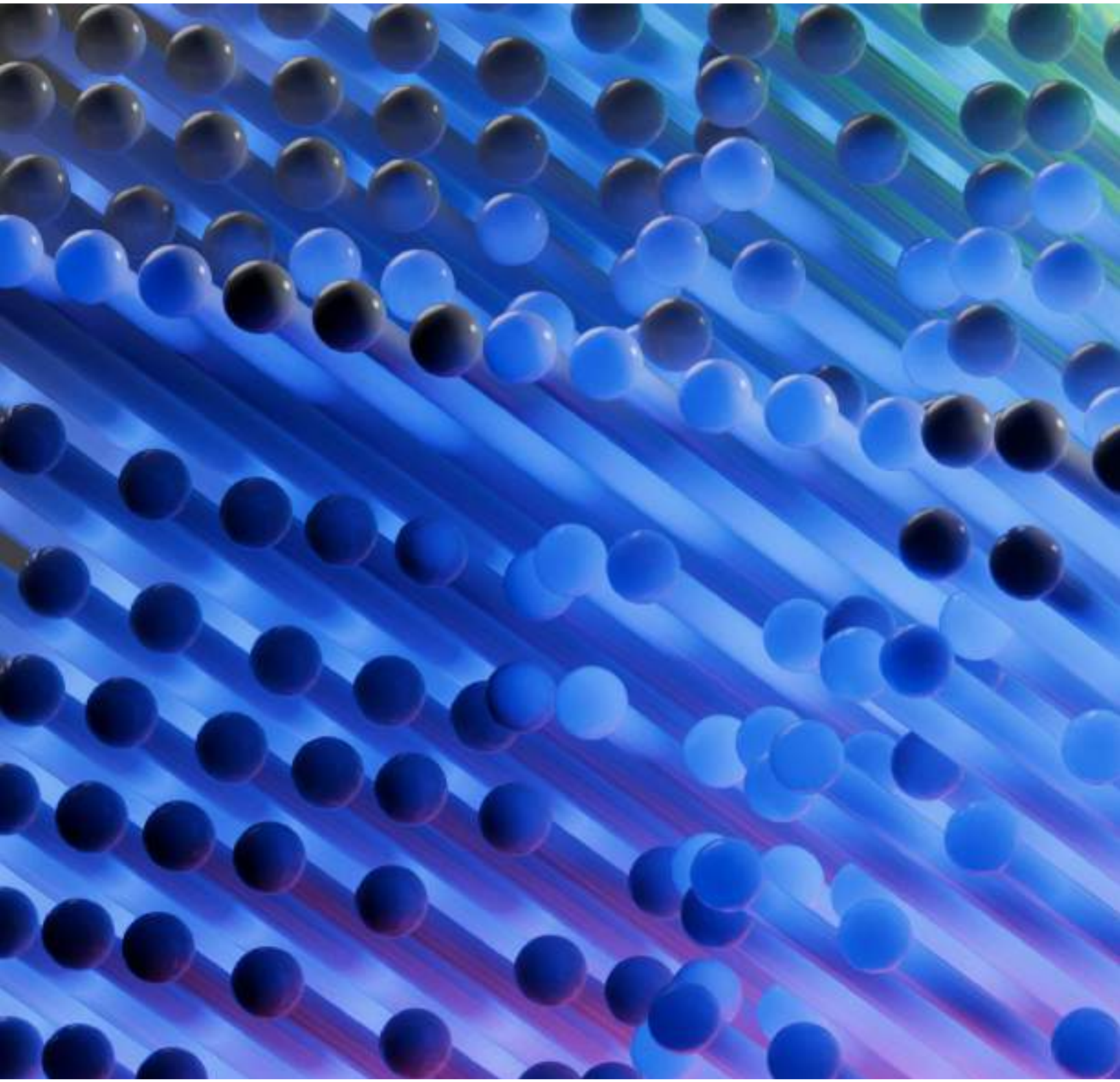
Mike Bechtel
Chief futurist, Deloitte Consulting LLP
mibechtel@deloitte.com

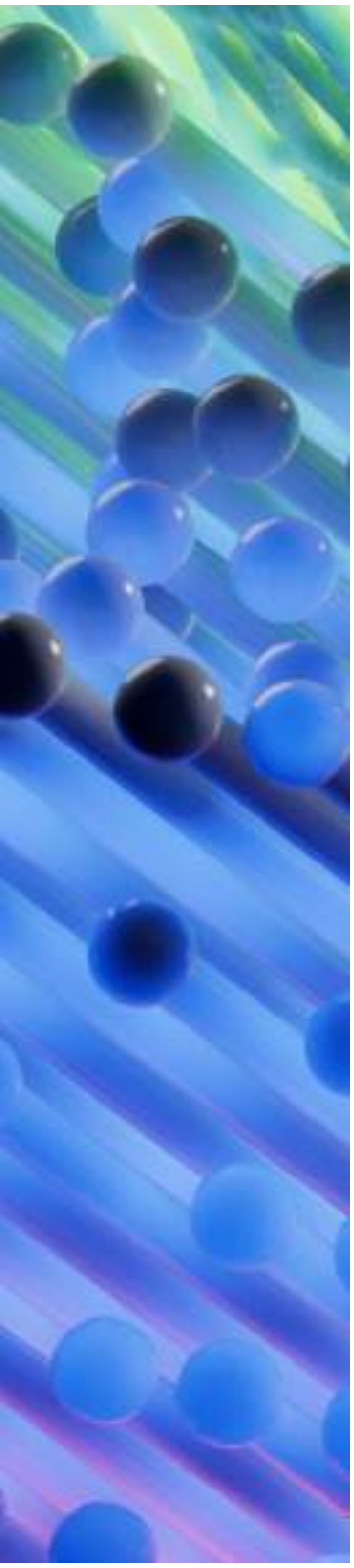
Figure 2

Six macro forces of information technology



Source: Deloitte analysis.





INTERACTION

Interfaces in new places: Spatial computing and the industrial metaverse

As the industrial metaverse transforms to enterprise tool, spatial technologies are taking hold in industrial applications, using data and AI to replicate real-life processes.

More often than not, novel technologies spark excitement with early adopters and consumers before quietly receding from the public eye, only to reappear months or years later as productive business tools. Some see this pattern as a movement along research firm Gartner's technology hype cycle,[1] while others understand it as a move from tech to toy to tool, as we discussed in *Tech Trends 2023*: In last year's trend "**Through the glass: Immersive internet for the enterprise**," we projected that the metaverse, or the immersive internet, would soon graduate to a full-blown enterprise tool as companies discover and build around new interaction capabilities such as augmented and virtual reality (AR/VR) simulations.[2]

This year, we've seen some of those metaverse capabilities progress in new directions, toward the broader realm of spatial computing. As they've turned the corner from consumer toy to enterprise tool, spatial technologies are especially taking hold in industrial applications, where companies are focused on **digital twins**,[3] **spatial simulation**,[4] augmented work instructions, and **collaborative digital spaces**[5] that make factories and businesses safer and more efficient. The opportunities are promising: Revenue driven by the industrial metaverse is projected to reach nearly US$100 billion by 2030, far outpacing the consumer (US$50 billion) and enterprise (US$30 billion) segments.[6]

Whether through time-tested devices such as tablets or experimental ones such as smart glasses, factory workers, designers, and engineers are benefiting from immersive 3D interaction in ways that traditional knowledge workers haven't yet experienced. The industrial metaverse is defined by real-world physics, using spatial data and artificial intelligence to render immersive visualizations that exactly replicate real-life processes. Imagine line workers using smart glasses to call an expert at a plant across the country, or engineers prototyping new equipment in physics-based, photorealistic digital twins. Where organizations have the opportunity to build new facilities, many are adopting a "simulation first" strategy before construction.

Improved and accessible high-fidelity 3D assets and hardware for extended reality (an umbrella term for immersive technologies such as AR, VR, and mixed reality) can pave the way to an operationalized spatial web, where a digital layer atop reality accelerates ways of working across industries. Eventually, this progress can lead to a simplified era of operations, where autonomous systems, instant 3D models, and

quantum computing are paired with optimized human involvement for applications such as remote surgeries. Or imagine an entire factory floor staffed by a single well-connected worker.

Now: Simulating the enterprise

Over the past few years, advancements in technology have been building the scaffolding for the industrial metaverse. Investments in digital twins, 5G enablement, cloud, edge, and AI have driven significant value and addressed long-standing pain points. That's why 92% of manufacturing executives surveyed in a recent Deloitte study said that their company is experimenting with or implementing at least one metaverse-related use case, and, on average, they are currently running more than six.[7] These executives already expect a 12% to 14% improvement in areas such as sales, throughput, and quality from investing in industrial metaverse use cases in the coming years.

The most common use cases highlighted by executives were process simulation and digital twins.[8] In industrial settings where operations are complex, pricey, and exact, robust simulations are a lifesaver. When connected to real-time data and models through the Internet of Things (IoT) and advanced networking, simulations can increase the chances of successfully building a new operation or optimizing an existing one. It's no surprise, then, that some analysts believe the global market for digital twins could grow from US$6.5 billion in 2021 to US$125.7 billion in 2030.[9]

The optimal way to interact with these full-scale digital twins is through AR, a medium that can overlay the physical world with a digital layer to create a shared, three-dimensional immersive internet. As a result, the global market for AR devices has been estimated at US$38.6 billion in 2022, with an annual growth rate of 36% through 2030 for related software and hardware.[10] While industrial and manufacturing applications currently make up the largest market share for AR, health care applications (such as training, surgical simulation, and vein visualization) are expected to grow by a compound annual growth rate of 44% through 2030. Consumer applications, catalyzed by the e-commerce boom of the pandemic, also abound, proving that the use cases for digital twins extend beyond just the enterprise.[11]

Spatial operations are just beginning, and enabling technologies continue to improve. Imagine powerful satellite networks combined with IoT sensors in a remote factory, processing real-time data on output and performance.[12] As technologies advance, a new era of digital twins is on the horizon, where simulations could be photorealistic, based on physics, and enabled by AI,[13] all while linked to company ecosystems, such as BMW's Omniverse platform.[14] This evolution is poised to affect multiple areas of the enterprise, from space planning to design to operations.

New: The spatial web is under construction

The impending spatial web (also known as Web 3.0) promises to eliminate the boundary between digital content and physical objects, effectively blending these two realities into one.[15] Through next-gen interfaces such as smart glasses, the spatial web can allow us to interact with real-time information prompted by our physical environment, through geolocation, computer vision, or biometric commands like voice and gestures. Given the possibilities, the market for spatial computing is poised to dwarf previous estimates for the metaverse, with some projections estimating upward of US$600 billion by 2032.[16]

While the true potential of the spatial web is still years away, innovators are building its infrastructure now. In the next 18 to 24 months, companies should pay attention to the value opportunities for adopting spatial operations and arming their employees with tech that supercharges their work.

US$600 billion by 2032

Given the possibilities, the market for spatial computing is poised to dwarf previous estimates for the metaverse, with some projections estimating upward of US$600 billion by 2032.

Augmented workforce

As workers in industrial settings continue to adopt AR/VR tools, companies are reaping the benefits of efficiency and effectiveness across a few key areas:

- **Increased monitoring.** As AR devices and spatial immersion allow employees to be in multiple "places" at once, fewer experts could monitor a greater number of facilities. For instance, Nokia's real-time eXtended Reality Multimedia provides 360-degree views, 3D audio, and live streaming to allow human operators to immerse themselves in a physical space many miles away.[17] This can bolster preemptive maintenance, security, and quality control.

- **Reduced onboarding time.** New employees can follow standard operating procedures that are built into simulations, along with visual cues that help them learn while in the flow of work, instead of having to separate learning from practice. For example, new employees at a global carmaker's manufacturing plants use AR devices to collaborate in real time with experts across the United States. Sharing the same vision and sound, the experienced line workers can instruct exactly where and how to strike a hammer on a door.[18]

- **Reduced safety risk.** As we discussed last year, companies can arm workers with AR/VR to better prepare them for risky settings. Stanford Medicine is piloting a VR system that combines images from MRIs and CT scans, among others, to create a 3D model of a patient's body prior to surgery. Surgeons can see and manipulate this anatomical digital twin, not only in training settings but in the operating room itself, as a more detailed guide to the body than 2D images. Doctors are already seeing benefits in improved accuracy and safety of some of the most complex procedures in medicine, such as brain surgeries.[19]

Product design, development, and sales

Use cases for spatial operations are not just limited to improving the bottom line; AR technologies can improve top-line revenue growth as well. For example, leading AR companies are enabling clothing retailers to integrate AR technology into their apps, websites, and physical locations to further differentiate their offerings. With generative AI, these retailers can soon use AR technology to create 3D models from 2D images, increasing the availability of digital assets for customer engagement in a spatial web.

Such AR technology can do much more than superimpose an image of clothing over a shopper. For example, it can simulate how fabric will fall on a customer or how different lines in the stitching create shadows. And the results are clear: Some retailers have seen an increase in revenue per visitor of more than 50% after building in AR technology.[20] As brands aim to stay relevant in spatial computing, AR companies are envisioning impact beyond retail, in sectors like education, entertainment, and travel.

Another way to take advantage of spatial operations is in design and testing of products under simulated conditions, which can lead to major improvements in agility, time to market, and even sustainability. For instance, instead of automakers subjecting their vehicles to hundreds of crash tests, they could use an initial set of data to simulate thousands of such tests and even consider events like natural disasters that can't be easily replicated in the real world. Pharmaceutical giant GSK applied these principles to employ simulations for vaccine production, enabling it to cut its time to run experiments from three weeks to a few minutes.[21] And in heavy asset industries such as mining, simulations can help fine-tune machine movements for efficiency and reduce emissions while preparing for the move to more renewable energy.

Space planning and simulation

The old adage of "measure twice, cut once" takes on new meaning in the age of spatial computing. Companies can employ spatial computing to visualize, simulate, and test layouts of facilities before undertaking costly investments: Measure 3,000 times, cut once. Architects can design an exact replica of a factory or hospital, replete with predictions of how many humans and machines will be present and how they'll interact and move. For instance, a busy hallway for triaging ER patients may need to be expanded after a hospital simulates its usual intake numbers. Or an auto manufacturer may want to predict how a planned factory will handle a surge in demand for electric vehicles in the years to come.

That's exactly what Hyundai Motor had in mind when partnering with Unity to build a pioneering full-scale factory simulation. The automaker plans to test the factory virtually to calculate an optimal method of operations and spacing, as well as one day enable plant managers to assess issues remotely.[22] Similarly, Siemens, a pioneer in the industrial metaverse field, has announced a new factory in Germany that will be entirely planned and simulated in the digital world first.[23] Only after adjusting its blueprints based on digital insights does the company plan to build the real-world campus.

Apart from the use cases for designing new spaces, spatial computing can also optimize a company's use of existing physical locations. For instance, the retail planning team at GUESS planned out in-store updates digitally and moved forward only after virtual testing, resulting in a 30% cost reduction and a lower carbon footprint from reducing travel to make in-store updates.[24]

Next: Let's get digital

The impending release of the Apple Vision Pro has made the term "spatial computing" more mainstream than ever.[25] While some may wonder if this latest trend may be a passing fad, we would not bet against simplicity. The history of technology has proven that simpler interaction modalities have reliably unlocked massive step changes in the accessibility, and in turn, use of technologies.[26] Spatial computing may be another such step change—where our natural gestures and ways of interacting with the physical world can be mapped onto the digital world, creating an ideal match between biology and technology.

As interaction technology continues to expand beyond computer science into the natural sciences (as we discuss in xTech dimensions[27]), brain-computer interfaces (BCIs) represent the furthest star of progress for simplicity. While today's BCI functionality is concentrated in *restoring* human capabilities (such as the ability to walk), future endeavors may *augment* human capabilities, enabling us to accomplish digital and physical tasks at a speed and scale that were previously unimaginable.

For that to take place, we'll need enabling technologies such as 6G networking and IoT. Through high-speed connectivity and massive machine-type communications, machines of the future may be able to coordinate with each other seamlessly.[28] And the World Economic Forum has already predicted that omnipresent IoT sensors can one day digitize physical human work, enabling a higher degree of automation.[29] Such advancements could pave the way for our interactions with machines to be much simpler as they become smarter at communicating about their environment and status.

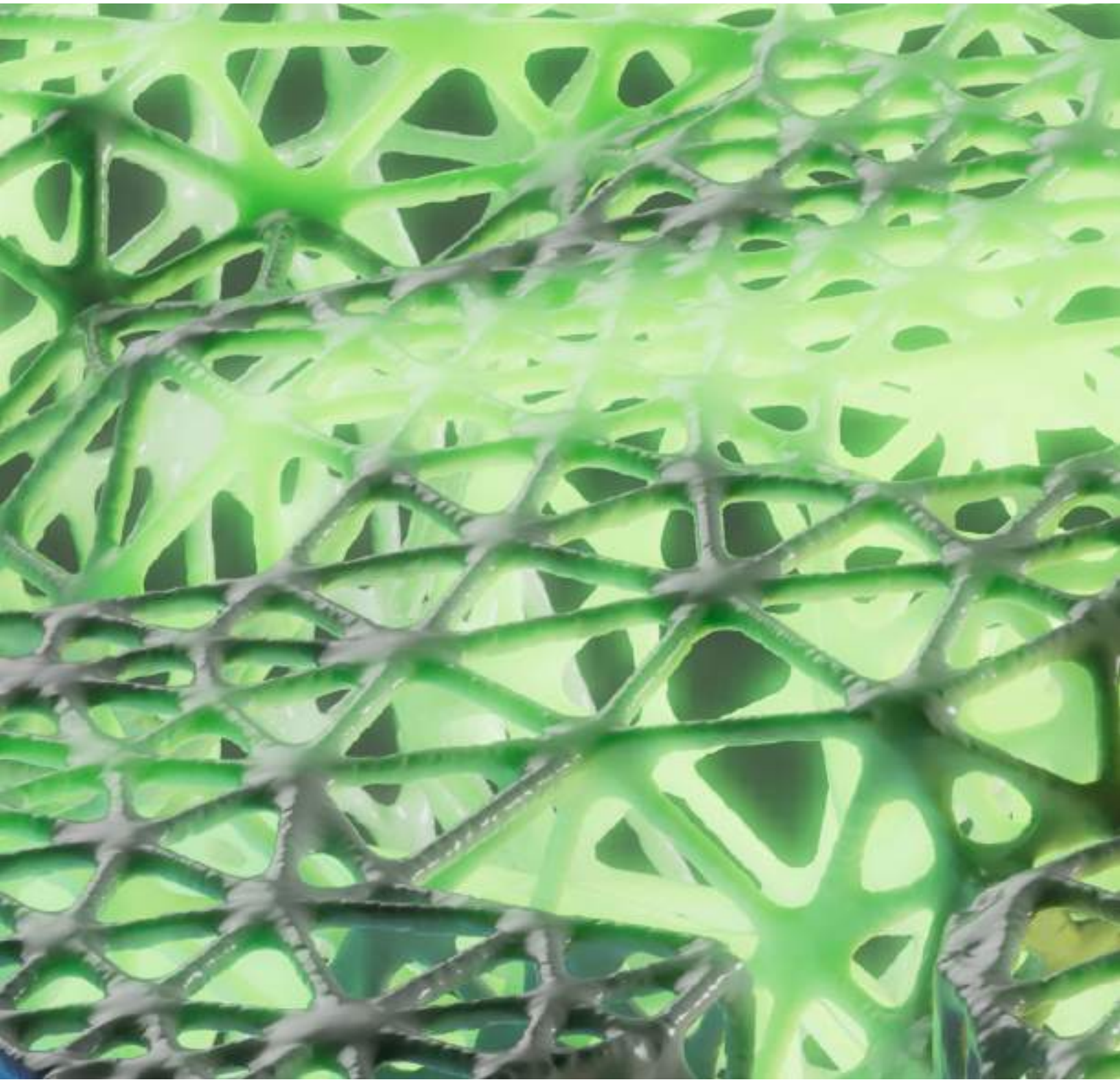
Imagine a future of interaction where BCIs enable us to start, monitor, and modify an interconnected series of machines on an assembly line. Industrial work could also become remote work, carried out from a desk. And language could feel like a bottleneck compared with the efficiency of human thought.

While the possibilities are exciting, companies are at a crossroads: They need to move beyond the buzzwords if they want to be early movers—or find themselves trying to catch up with innovators. Beyond hiring or training their engineers on computer vision, sensor tech, and spatial mapping algorithms, they should also get ahead of the potential risks. Opening up the physical world to digital manipulation comes with its fair share of privacy issues (as computer vision expands), cybersecurity issues (as the physical world becomes hackable), and data protection issues.[30] Fortunately, the progress of digital twin technologies and early 3D models offers valuable lessons for steps forward.

Once the initial benefits of spatial operations are underway in industrial settings, enterprises should be prepared: The natural evolution of spatial computing may radically change the way we interact with consumer and enterprise applications in the years to come.

Endnotes

1. Gartner, "Gartner Hype Cycle," accessed October 2023.
2. Deloitte Insights, *Through the glass: Immersive internet for the enterprise*, December 6, 2022.
3. Aaron Parrott, Lane Warshaw, and Brian Umbenhauer, *Digital twins: Bridging the physical and digital*, Deloitte Insights, January 15, 2020.
4. Deloitte, "Unlimited Reality for operations," accessed October 2023.
5. Deloitte, "Unlimited Reality for the workforce," accessed October 2023.
6. ABI Research, *Evaluation of the enterprise metaverse opportunity*, September 20, 2022; Transparency Market Research, *Industrial metaverse market outlook 2031*, June 2023.
7. Paul Wellener et al., "Exploring the industrial metaverse," Deloitte and Manufacturing Leadership Council, accessed October 2023.
8. Ibid.
9. J. Pankaj, M. Neha, and V. Vitika, *Digital twin market size, share and trends analysis by 2030*, Allied Market Research, July 2022.
10. Grand View Research, *Augmented reality market size and share report*, 2023.
11. Ibid; Markets and Markets, *Augmented reality market report*, August 2021.
12. Deloitte, *xTech Futures: SpaceTech*, 2023.
13. MIT Technology Review Insights and Siemens, *The emergent industrial metaverse*, March 29, 2023.
14. Deloitte, "Connect and extend: NVIDIA's vision for modernizing legacy applications," Deloitte Insights, November 9, 2022.
15. Allan V. Cook, Siri Anderson, Mike Bechtel, David R Novak, Nicole Nodi, and Jay Parekh, *The spatial web and Web 3.0*, Deloitte Insights, July 21, 2020.
16. Market.us, *Global spatial computing market report*, August 2023.
17. Nokia, "Real-time eXtended Reality Multimedia," accessed October 2023.
18. Jack Siegel, "HoloLens 2 brings new immersive collaboration tools to industrial metaverse customers," Microsoft, December 20, 2022.
19. Mandy Erickson, "Virtual reality system helps surgeons, reassures patients," Stanford Medicine News Center, July 11, 2017.
20. Deloitte interviews.
21. Deloitte, "Unlimited Reality for operations."
22. Hyundai Motor Company, "Hyundai Motor and Unity partner to build Meta-Factory accelerating intelligent manufacturing innovation," press release, January 6, 2022.
23. Siemens, "Siemens to invest €1 billion in Germany and create blueprint for industrial metaverse in Nuremberg metropolitan region," press release, July 13, 2023.
24. Deloitte, "Unlimited Reality for operations."
25. *Tech Trends* is an independent publication and has not been authorized, sponsored, or otherwise approved by Apple Inc.
26. Deloitte, *Tech Trends 2023 Prologue: A brief history of the future*, Deloitte Insights, December 6, 2022.
27. Deloitte, *Tech Trends 2023 epilogue*, Deloitte Insights, December 6, 2022.
28. Charles McLellan, "What is the state of 6G, and when will it arrive? Here's what to look out for," ZDNET, February 17, 2023.
29. Francisco Betti, Thomas Bohné, and Cathy Li, "The industrial metaverse and its future paths," World Economic Forum, January 19, 2023.
30. Wellener et al., "Exploring the industrial metaverse."



INFORMATION

Genie out of the bottle: Generative AI as growth catalyst

Since gen AI technology exploded on the scene, many enterprises have been scrambling to determine how their businesses might benefit. The answer might be simpler than they think.

Starting around 2015, people began referring to almost any application of machine learning as artificial intelligence. Some pundits and industry experts pushed back. These applications were pattern matchers, they said.[1] Given an input, they return an output. The models didn't think, but rather computed probabilities, so how could they be intelligent?

Generative AI makes moot the question of whether machines can be intelligent. The underlying operation of these models shares much in common with earlier machine learning tools, but thanks to accelerated computing power, better training data, and clever applications of neural networks and deep learning, generative AI technology can imitate human cognition in a number of ways. More and more often, machines that possess intelligence in at least a functional, practical sense create the opportunity for huge productivity and efficiency gains in enterprise settings, as well as the opportunity to bring innovative new products and services to new markets.

In plenty of instances, AI tools perform at least as well as, if not better than, human counterparts in tests of cognitive capabilities. ChatGPT recently scored a 5—"extremely well qualified" on the notoriously challenging Advanced Placement biology test.[2] The Dall-E 2 image generator was able to solve Raven's Matrices, a test given to measure a subject's visual IQ.[3] Anthropic's Claude 2 chatbot scored above the 90th percentile in the verbal and writing sections of the GRE test, which is used by many graduate schools in the United States and Canada as part of their admissions standards.[4] In fact, AI tools now consistently outperform humans on measures of handwriting, speech, and image recognition; reading comprehension; and language understanding.[5]

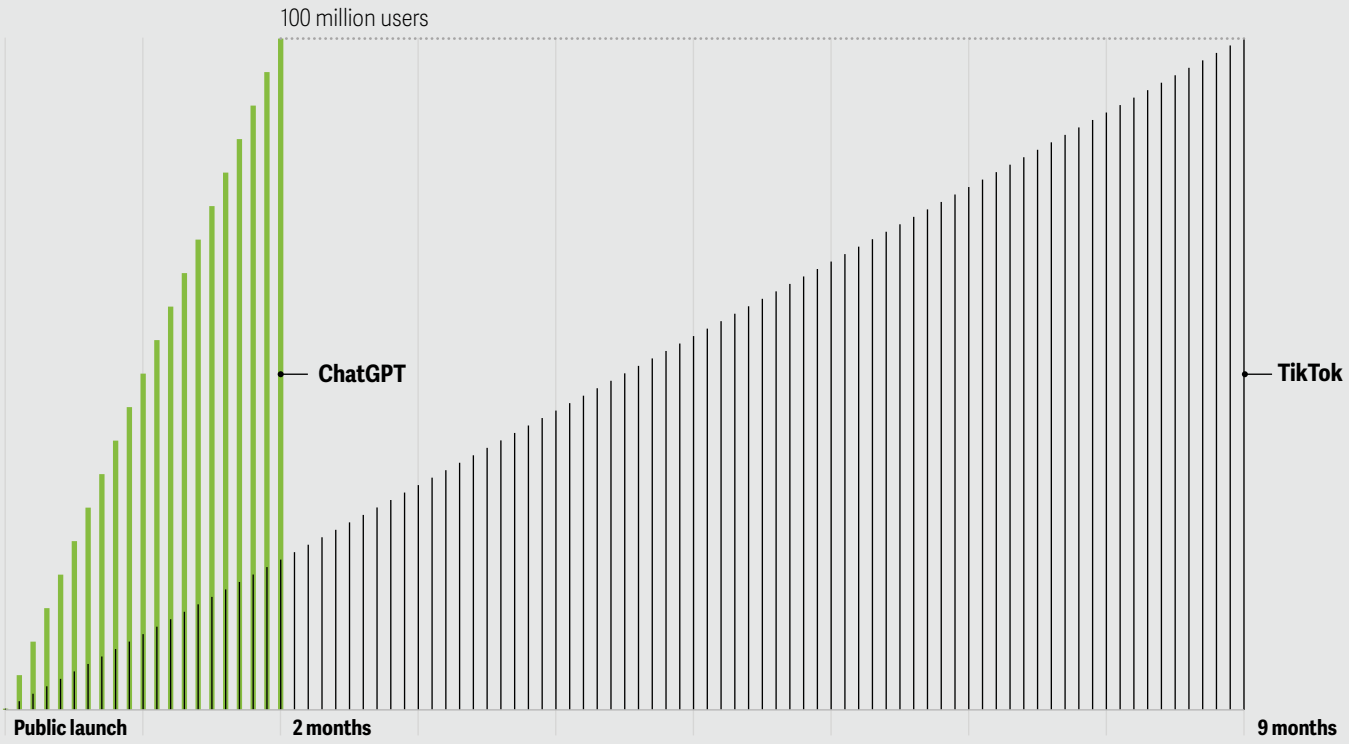
The question is no longer whether AI tools are intelligent. Today the question is more about how to deploy these cognitive tools in ways that provide real business impact.

Now: Generative AI interest and adoption soar, promising disruption

Generative AI captured the public's imagination when it burst onto the scene in the second half of 2022 and first few months of 2023. Few technologies have ever debuted to such fanfare. Adoption and use of generative AI have been sudden and rapid among the public. OpenAI reported reaching 100 million users within 60 days of releasing ChatGPT to the public; in comparison, it took TikTok nine months to reach that milestone (figure 1).[6] Midjourney's image generator has around 16 million users.[7] There are 1.5 million daily users of Dall-E 2.[8] Google's Bard chatbot had 10 million page views in July.[9] Growth in the use of generative AI in enterprise settings has been no less impressive, according to Deloitte's 2023 CEO Priorities Survey (figure 2).[10]

Figure 1

Generative AI interest and adoption soar



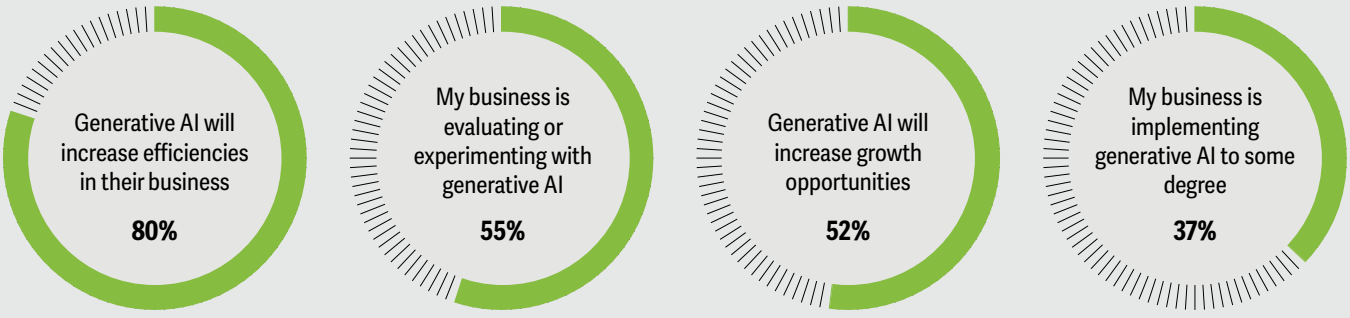
Note: Graphs are figurative and do not depict actual rates of growth.

Source: Krystal Hu, "ChatGPT sets record for fastest-growing user base–analyst note," Reuters, February 2, 2023.

Figure 2

Business leaders are increasingly using generative AI in enterprises

Percentage of business leaders who agreed with the following statements



Source: Fortune/Deloitte CEO Survey Insights, summer 2023.

What's made generative AI so impactful is a convergence of factors. First, advanced hardware—primarily specialized AI chips used in training models—have helped produce more advanced models such as large language models (LLMs). These tools have gone mainstream due to a seamless user experience, enabling even nontechnologists to engage with very advanced models.

All this attention has kicked off a gold rush among investors (figure 3). Investors are pouring money into startups that have generative AI technology at their core, betting that we're witnessing the dawn of a new paradigm for business technology, one where insights are surfaced automatically, contracts review themselves, and a never-ending stream of content is generated to keep brands in front of their audiences.

While there's been plenty of talk about how AI may threaten jobs, there's no real indication that business leaders are planning on using it to automate knowledge jobs at any kind of scale. In a survey of leaders, improving content quality, driving competitive advantage, and scaling employee expertise were the most common reasons for deploying generative AI. Reducing headcount was one of the lowest priorities.[11] It looks more likely that AI will liberate workers from rote, repetitive tasks and free them to focus on more creative aspects of their jobs.

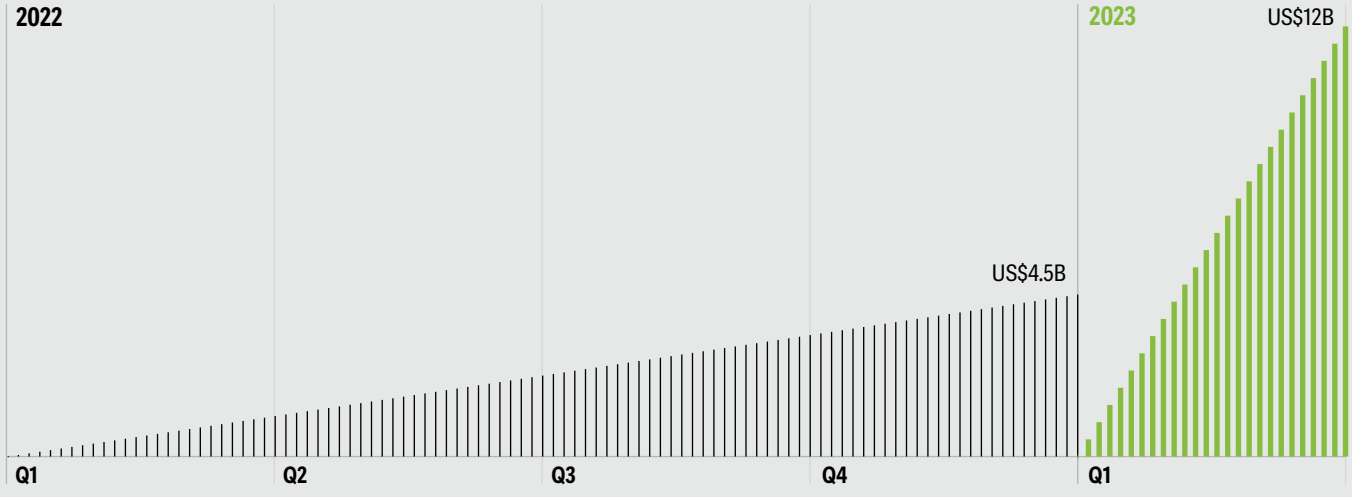
The picture that's emerging is that AI is coming, and for some, it's already here. But, as the saying goes, leading businesses know they can't **shrink their way to growth**—that is, minimize risks or costs as a path to growth.[12] This means the most productive uses of generative AI won't be about replacing people but instead will focus on arming employees with tools that help them advance and enhance their productivity, knowledge, and creativity—which, in turn, will help drive innovation in the enterprise.

Executives are increasingly under pressure to speed this transition and stay ahead of their competitors. According to one survey, 64% of CEOs say they're facing significant pressure from investors, creditors, and lenders to speed the adoption of generative AI.[13] But just as leaders know they can't shrink their way to growth, they also know the importance of **leading with need**.[14] Shoehorning generative AI into any and all processes just because it's a shiny new thing is unlikely to deliver meaningful gains. Instead, businesses may benefit from a more strategic

Figure 3

Investment in generative AI has exploded

Annual cumulative venture capital money invested in generative AI technologies



Note: Graphs are figurative and do not depict actual rates of growth.

Source: Jacob Robbins, "The most active investors in generative AI," News & Analysis, June 15, 2023.

approach to implementation that focuses on leveraging generative AI's unique capabilities to solve existing problems and help businesses differentiate themselves from competitors. That's the approach innovative enterprises are taking today.

New: Enterprises aim for scalability and domain expertise

The true value of generative AI is likely to be unlocked when organizations can use it to transform business functions; reduce costs; disrupt product, service, and innovation cycles; and create previously unachievable process efficiencies. To get there, business leaders may consider more of an evolutionary approach to their enterprise data and technology strategy.

Becoming an AI-fueled organization takes careful discipline and a focus on maintaining systems and algorithms.[15] Just as a rocket needs a launch pad and flight controls to reach its destination, generative AI tools need infrastructure and control systems to succeed in enterprise settings. The good news is a lot of the muscle memory businesses have been developing over the past several years while building up data analytics and machine learning capabilities also applies to generative AI, though some practices may require subtle retooling.

Generative AI typically requires terabytes of data on graphics processing unit–enabled high-performance computing clusters. Since few businesses have this infrastructure, most will access it as a service. Via application programming interfaces, engineers can weave generative AI capabilities into their existing software without needing to build out new infrastructure.[16] While AI vendors are prioritizing ease of use in their products, it's still important for enterprises to keep these engineering requirements in mind.

Additionally, it's important to pick your use cases wisely. AI can be used to reduce costs, speed up processes, reduce complexity, transform customer engagement, fuel innovation, and build trust.[17] The specific application of generative AI will vary from business to business, but looking for projects that drive improvements in one area is a good place to start.

Here are some additional considerations from businesses that have already adopted the technology.

Data is the fuel that drives the generative AI engine

Businesses need to ensure their data is architected properly and accessible to AI applications to enable model training as well as next-generation use cases.

This was one of the learnings for Enbridge, the largest natural gas utility in North America. Several years ago, when it began an ambitious cloud migration journey, it didn't set out to pioneer new generative AI uses. The primary goals were to modernize its infrastructure and eliminate technical debt by reducing the size of its on-premises data centers. Along the way, it built a centralized data repository that collects data from across the enterprise, including regulatory, marketplace, HR, and other data. This centralized data marketplace replaced what used to be hundreds of silos.

Once generative AI arrived on the scene, Enbridge's leadership knew this centralized data marketplace was the perfect engine to drive new AI-fueled efficiencies. The technology team rolled out a generative-AI-based copilot tool that helps developers quickly and more efficiently build out code. It also supplied the company's office staff with a copilot tool to help them navigate productivity applications.

The goal, says Joseph Gollapalli, director of cloud, IT ops, and data at Enbridge, is to "accelerate our delivery and drive innovation and efficiency. These AI solutions have the potential to enhance our operations, improve safety, elevate the customer experience, and enhance our environmental performance."[18]

Governance is more important than ever

Without effective governance guardrails, AI can't scale. A governance framework should define the business's vision, identify potential risks and gaps in capabilities, and validate performance.[19] These types of considerations not only safeguard the business but can also help scale projects beyond the proof-of-concept stage.

At CarMax, the largest used car retailer in the United States, effective use of generative AI is predicated on a systematic, enterprise-wide approach that embraces

the power of the technology while also putting in place guardrails to ensure employees are using it effectively. One of CarMax's most prominent applications is a tool that adds AI-generated content to research pages for vehicles. These pages summarize information from thousands of actual customer reviews to let shoppers quickly see what other buyers had to say.

Shamim Mohammad, executive vice president and chief information and technology officer at CarMax, says these kinds of use cases deliver the most business value when they are done in a controlled manner.[20] CarMax has prioritized governance, which may not feel like the most exciting aspect of generative AI but is key to scaling it. The company has created an AI governance team dedicated to ensuring teams across the organization are using AI appropriately. The key is that this team is not charged with simply saying no to new use cases. Part of its mission to help scale impactful applications across the enterprise by standardizing how models are trained and used. The goal is that generative AI is used beyond just technology or product teams.

"We've done a lot of cool things through machine learning and AI," says Mohammad. "What I'm focused on now is ensuring we're using it in a responsible manner and making sure that, as a company, whatever we deploy, it's being done in ways that are consistent with our core values."

Make sure you have the (copy)right

Generative AI has altered the copyright landscape. Now anyone can create images, video, text, and audio with a few clicks. However, some models have been trained on content that comes from third parties. One US court recently ruled that this makes AI-generated content ineligible for copyright protection.[21] Additionally, training models on copyrighted material scraped from the web may present legal risks, including intellectual property infringement.[22]

However, these don't have to be problems. The content provider Shutterstock, for one, has shown that it is possible to use generative AI in ways that both respect the rights of the original copyright holder and ensure that AI-generated content can be used for commercial purposes.

Shutterstock recently unveiled an image-generating tool that creates visuals based on users' prompts. Like other image generators, the tool was trained on images created by third-party artists. However, unlike other image generators, every artist whose work was used in training the model agreed ahead of time to participate. Participating artists are also paid when their work is used to train a model and when a user licenses an image generated on the platform. Shutterstock licenses its content as data, which allows it to offer added legal protections to end users.

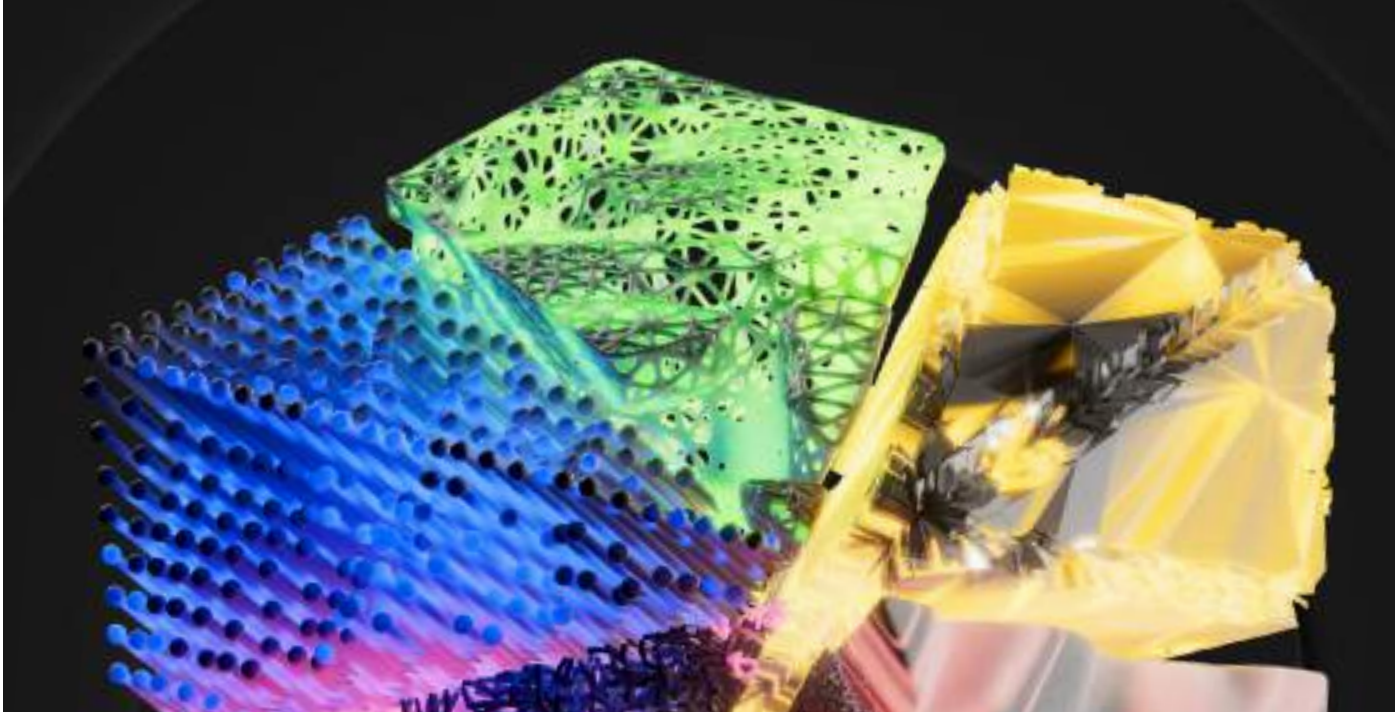
"Everyone is creating content, from CEOs to folks who work in retail," says Michael Francello, director of innovation at Shutterstock. "The need to create content was absolutely exploding. We saw an early opportunity to look at our content as data that could train generative AI models. It's about protecting the core of our business, but also respecting the core, which is the artists and the contributors."[23]

Crawl, walk, run, fly

This approach has for years been an effective way for enterprises to scale up their use of service offerings.[24] Generative AI is no different. In the crawl stage, applications may be ad hoc and require lots of manual effort. These eventually graduate to the walk stage, in which processes become more defined at the foundational level and automated. In the run stage, use cases get standardized and become pervasive at the enterprise level. When it's time to fly, the organization leverages the work it has already done to embrace next-generation capabilities.

That approach helped chemical company Eastman begin developing generative AI-based internal services. The company has a long track record of using data and analytics in an industry that isn't typically known for it. For example, it has an advanced intelligence service (with proprietary thermal stability measures) that will predict when a heat transfer fluid used in its customers' industrial processes is likely to degrade, allowing engineers to maintain optimal fluid quality, forecast predictive maintenance needs, and avoid costly downtime on manufacturing lines.

Building on this experience, the company is now experimenting with how generative AI can enhance its sales processes. It built an AI-enabled tool that can read



through natural language text files. Still in the development stages, the tool is being tested on extracting insights from notes from sales calls. These documents are generated by sales teams after every call but rarely get read by anyone, even though they hold significant intelligence. Now, with the help of generative AI, the company is starting to unlock those insights.

"It lets us, a chemical company, bring a digital service layer to the table to differentiate ourselves in the market and create a competitive advantage," says Aldo Noseda, chief information officer at Eastman.[25]

Given the pace with which generative AI is progressing, it may be wise to apply this kind of framework to new enterprise use cases. Let proof-of-concept projects lead to standardized practices that become standard operating procedures across the enterprise. Once a business has achieved this kind of maturity, the sky is the limit.

In the near future, it may become even easier for businesses to reap the benefits of generative AI within their industries thanks to the emergence of models that are trained on more specific data. Today, most enterprises that are using generative AI are using tools built on foundational models that were trained on general-purpose data. That tools with such a general knowledge base can be used in very specific subject-matter areas shows the power of LLMs. But the next generation of LLMs is likely to be more hyper-focused and tailored to businesses' specific needs.[26]

This is a trend that's already begun to emerge. NVIDIA has introduced a tool called BioNeMo, an LLM aimed at the biotech sector.[27] Google's Contact Center AI is a tool trained to handle customer service interactions.[28] BloombergGPT is designed to answer finance industry-related questions.[29] ClimateBERT is a model trained on climate change research and can advise businesses on their climate-related risks.[30]

As businesses realize the benefits of models trained specifically for their sector, we're likely to see more demand for these types of services. More than one-third of enterprises are already planning to train and customize LLMs for their business needs in the future.[31] Private LLMs are likely where the true potential of generative AI lies for businesses. They are developed and maintained by organizations that keep underlying code proprietary and closed to the public. These LLMs are purpose-specific, hosted securely, and trained on private data, and they can offer tremendous competitive advantage to organizations. This is likely the next wave in the generative AI journey.

Next: Imaginative executives wanted

The motivational poster has gone from mere corporate cliché to its own category of meme, but one overused aphorism may reclaim its stature as an enterprise imperative: We're only limited by our imagination.

While you might have heard the saying before, teams and organizations have always been bound by limiting factors. They don't have enough data or the right data. Leadership is skeptical. Or, most dreaded of all: "That won't move the needle."

But in a generative AI world, imagination truly is the only limit. It's now possible to create constant streams of content, identify new operational efficiencies, or scan regulatory filings or customer reviews in minutes. Now the only question is, what do you want to know?

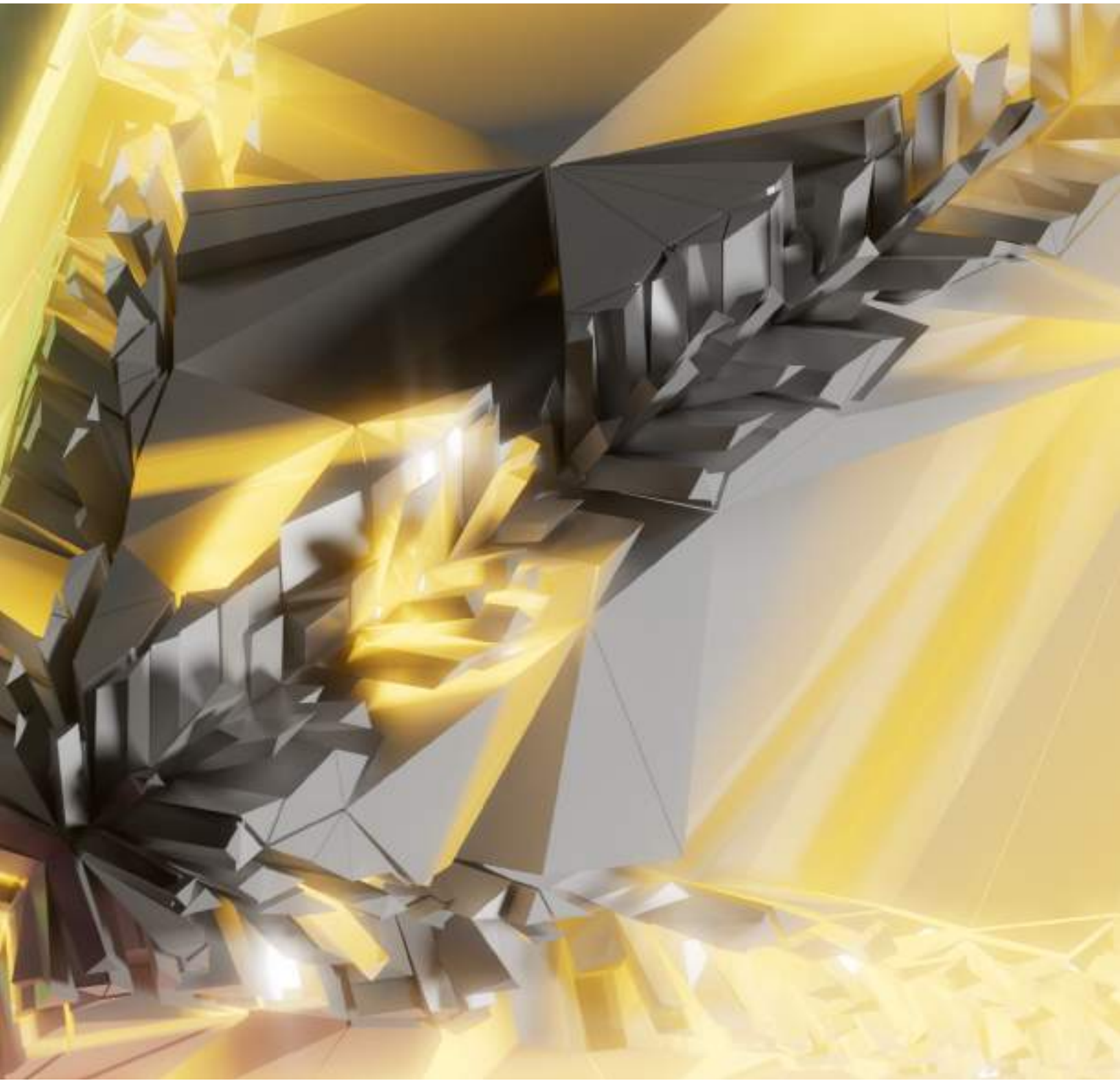
Asking better questions will become a crucial skillset in enterprises that have adopted generative AI. This trend may create demand for a new type of leader, one that is driven more by creativity than we've seen in the past. The past 20 years or so have seen leaders rewarded for steering their organizations based on data and insights, rather than gut and instinct. But the next few years could see more imaginative leaders leap ahead. Give an image generator a boring prompt, and it will produce a boring picture. The same is true of generative AI applications at the enterprise level. Unimaginative use cases produce limited impact. As more businesses attempt to differentiate themselves from their competition, leaders who can find creative new applications for generative AI may separate themselves from their peers who are busy just following data.

This isn't to say that data-driven decision-making will become passé. In fact, it will be as important as ever, if not more so. But the definition of what it means to be data-driven may change because the range of data that leaders can access will increase, thanks to generative AI. So much of an enterprise's data is buried in natural language text files, machine logs, and, increasingly, intelligent products.[32] Generative AI gives organizations the ability to make sense of this digital exhaust. The creative leader will understand what these oft-overlooked data sources have to say about their business and will use generative AI to ask intelligent questions of the data sources. And they will ask these questions at the speed of thought, rather than waiting for their weekly report.

But all that barely scratches the surface of generative AI's full range of likely impacts. We're pretty sure it's going to be seismic. We just don't know exactly where the ground will shift the most.

Endnotes

1. Michael I. Jordan, "Artificial intelligence—the revolution hasn't happened yet," *Harvard Data Science Review,* July 1, 2019.
2. Tom Huddleston Jr., "Bill Gates watched ChatGPT ace an AP Bio exam and went into 'a state of shock'," CNBC, August 11, 2023.
3. Saliha Malik, "How will the Open AI products DALL.E and DALL.E 2 change the face of augmented reality?," *Medium*, March 1, 2023.
4. Anthropic, "Claude 2," July 11, 2023.
5. Douwe Kiela et al., "Dynabench: Rethinking benchmarking in NLP," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 6–11, 2021, pp. 4110–24.
6. Krystal Hu, "ChatGPT sets record for fastest-growing user base – analyst note," Reuters, February 2, 2023.
7. Rob Krivec, "Midjourney statistics (How many people are using Midjourney?)," *Colorlib*, September 14, 2023.
8. Leigh McGowran, "OpenAI unveils Dall-E 3 art generator with ChatGPT support," *Silicon Republic*, September 21, 2023.
9. Similarweb, Google Bard overview, accessed October 2023.
10. Deloitte, "Summer 2023 Fortune/Deloitte CEO Survey insights," accessed October 2023.
11. IBM Institute for Business Value, "Enterprise generative AI," July 2023.
12. Mike Bechtel, Khalid Kark, and Nishita Henry, "Innovation Study 2021: Beyond the buzzword," Deloitte Insights, September 30, 2021.
13. IBM Institute for Business Value, "Enterprise generative AI."
14. Bechtel, Kark, and Henry, "Innovation Study 2021: Beyond the buzzword."
15. Nitin Mittal, Irfan Saif, and Beena Ammanath, *State of AI in the Enterprise, 5th edition report*, Deloitte, October 2022.
16. Deloitte AI Institute, "Generative AI is all the rage," 2023.
17. Deloitte AI Institute, *The AI Dossier—expanded*, 2022.
18. Interview with Joseph Gollapalli, director of cloud and IT operations at Enbridge, June 13, 2023.
19. Beena Ammanath et al., "Trustworthy AI in practice," Deloitte AI Institute, 2022.
20. Interview with Shamim Mohammad, executive vice president and chief information and technology officer at CarMax, August 3, 2023.
21. Trishla Ostwal, "Judge rules GenAI content does not have copyright protection," *Adweek*, August 22, 2023.
22. Gil Appel, Juliana Neelbauer, and David A. Schweidel, "Generative AI has an intellectual property problem," *Harvard Business Review*, April 7, 2023.
23. Interview with Michael Francello, director of innovation at Shutterstock, May 12, 2023.
24. Jagjeet Gill, Deepak Sharma, and Anne Kwan, "Scaling up XaaS," Deloitte, September 29, 2019.
25. Interview with Aldo Noseda, chief information officer, Eastman Chemical Company, October 11, 2023.
26. Deloitte AI Institute, *A new frontier in artificial intelligence*, 2023.
27. Ibid.
28. Google Cloud, "Contact Center AI," accessed October 2023.
29. Bloomberg, "Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance," press release, March 30, 2023.
30. ChatClimate, "ClimateBert," accessed October 2023.
31. expert.ai, "Nearly 40% of enterprises surveyed by expert.ai are planning to build customized enterprise language models," press release, May 11, 2023.
32. Jagjeet Gill et al., "Analytics operating models," Deloitte, 2020.



COMPUTATION

Smarter, not harder: Beyond brute force compute

Businesses are getting more out of their existing infrastructure and adding cutting-edge hardware to speed up processes. Soon, some will look beyond binary computing entirely.

As technology has become a bigger differentiator for enterprises, businesses have built ever-more computationally complex workloads. Training artificial intelligence models, performing complex simulations, and building digital twins of real-world environments requires major computing resources, and these types of advanced workloads are beginning to strain organizations' existing infrastructure. Typical cloud services still provide more than enough functionality for most business-as-usual operations, but for the cutting-edge use cases that drive competitive advantage, organizations now require highly optimized and specialized computing environments.[1]

Optimizing code bases for the hardware they run on is likely the first step toward speeding up business applications. An area that's long been overlooked, this optimization can provide significant performance gains. Beyond that, emerging hardware geared specifically for training AI and other advanced processes is becoming an enterprise mainstay. Graphics processing units (GPUs), AI chips, and, one day, quantum and neuromorphic computers are beginning to define the next era of computing.[1]

Most advances in computing performance have focused on how to get more zeros and ones through a circuit faster. That's still a fertile field, but, as we're starting to see, it may not be for much longer. This is leading researchers and tech companies to look for innovative ways to navigate around—rather than through—constraints on computing performance. In the process, they could be laying the groundwork for a new paradigm in enterprise computation in which central processing units (CPUs) work hand in hand with specialized hardware, some based in silicon, others, potentially not.

Now: Past performance not indicative of future returns

The last 50 or so years of computing—and economic—progress have been shaped by Moore's Law, the idea that the number of transistors on computer chips, and therefore performance, roughly doubles every two years.[2]

However, chipmakers are increasingly running into physical constraints. At a certain point, there are only so many transistors a piece of silicon can hold. Some observers believe Moore's Law is already no longer valid.[3] This is contested, but at the very least, the end of the runway may be coming into view. Chips are getting both more power-hungry and harder to cool, which hampers performance,[4] so even as chip manufacturers add more transistors, performance doesn't necessarily improve.

All this comes at a bad time: Businesses are increasingly moving toward computationally intensive workloads. Industrial automation is ramping up, with many

companies developing digital twins of their real-world processes. They're also increasingly deploying connected devices and the Internet of Things, both of which create huge amounts of data and push up processing requirements. Machine learning, especially generative AI, demands complex algorithms that crunch terabytes of data during training. Each of these endeavors stands to become major competitive differentiators for enterprises, but it's not feasible to run them on standard on-premises infrastructure. Cloud services, meanwhile, can help bring much-needed scale, but may become cost-prohibitive.[5]

The slowing pace of CPU performance progress won't just impact businesses' bottom lines. NVIDIA CEO Jensen Huang said in his GTC conference keynote address that every business and government is trying to get to net-zero carbon emissions today, but doing so will be difficult while increasing demand for traditional computation: "Without Moore's Law, as computing surges, data center power use is skyrocketing."[6]

After a certain point, growing your data center or increasing your cloud spend to get better performance stops making economic sense. Traditional cloud services are still the best option for enabling and standardizing back-office processes such as customer relationship management, enterprise resource planning (ERP), enterprise asset management, and human capital management. But running use cases that drive growth, such as AI and smart facilities, in traditional cloud resources could eventually eat entire enterprise IT budgets. New approaches, including specialized high-performance computing, are necessary.[7]

New: Making hardware and software work smarter, not harder

Just because advances in traditional computing performance may be slowing down doesn't mean leaders have to pump the brakes on their plans. Emerging approaches that speed up processing could play an important role in driving the business forward.

Simple

When CPU performance increased reliably and predictably every year or two, it wasn't the end of the world if code was written inefficiently and got a little bloated. Now, however, as performance improvements slow down, it's more important for engineers to be efficient with their code. It may be possible for enterprises to see substantial performance improvements through leaner code, even while the hardware executing this code stays the same.[8]

A good time to take on this task is typically during a cloud migration. But directly migrating older code, such as COBOL on a mainframe, can result in bloated and inefficient code.[9] Refactoring applications to a more contemporary code such as Java can enable enterprises to take advantage of the modern features of the cloud and help eliminate this problem.

The State of Utah's Office of Recovery Services recently completely a cloud migration of its primary case management and accounting system. It used an automated refactoring tool to transform its code from COBOL to Java and has since seen performance improvements.

"It's been much faster for our application," says Bart Mason, technology lead at the Office of Recovery Services. "We were able to take the functionality that was on the mainframe, convert the code to Java, and today it's much faster than the mainframe."[10]

Situated

Using the right resources for the compute task has helped Belgian retailer, Colruyt Group, embark on an ambitious innovation journey that involves automating the warehouses where it stores merchandise, using computer vision to track and manage inventory levels, and developing autonomous vehicles that will one day deliver merchandise to customers.

One way to manage the compute workload is to leverage whatever resources are available. Brechtel Dero, division manager at Colruyt Group, says thanks to the proliferation in smart devices, the company had plenty of computation resources available.[11] However, many of these resources were in operational technologies and weren't tied to the company's more traditional digital infrastructure. Developing that connective tissue was initially a challenge. But Dero says Colruyt benefitted from a supportive CEO who pushed for innovation. On the technical side, the company operates a

flexible ERP environment that allows for integration of data from a variety of sources. This served as the backbone for the integration between information and operations technology.

"It's about closing the gap between IT and OT, because machines are getting much smarter," Dero says. "If you can have a seamless integration between your IT environment, ERP environment, and machines, and do it so that the loads and compute happen in the right place with the right interactions, we can make the extra step in improving our efficiency."[12]

Specialized

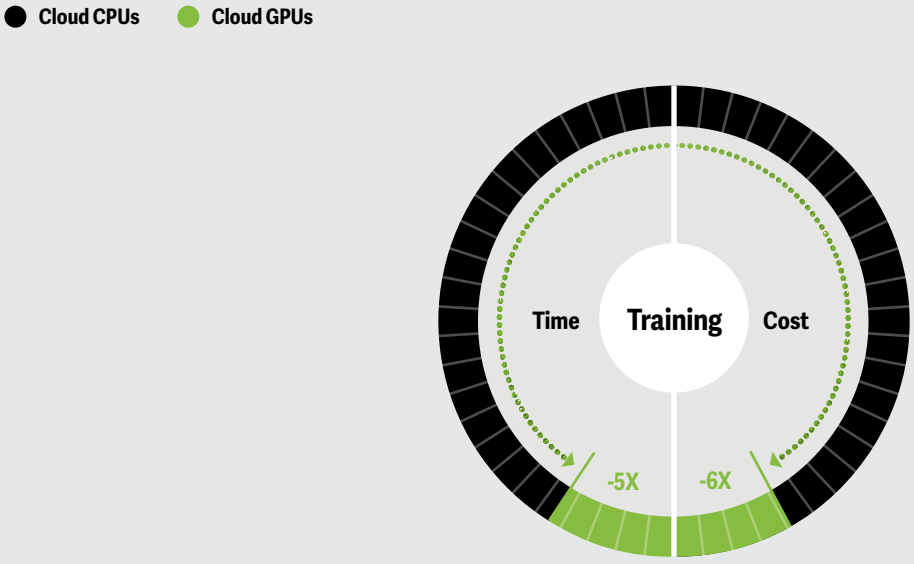
Smarter coding and better use of existing compute resources could help enterprises speed up many of their processes, but for a certain class of problems, businesses are increasingly turning to specialized hardware. GPUs have become the go-to resource for training AI models, a technology that is set to drive huge advances in operational efficiency and enterprise innovation.

As the name suggests, GPUs were originally engineered to make graphics run more smoothly. But along the way, developers realized that the GPUs' parallel data-processing properties could streamline AI model training, which involves feeding terabytes of data through algorithms, representing one of the most computationally intensive workloads organizations face today. GPUs break problems down into small parts and process them at once; CPUs process data sequentially. When you're training an AI algorithm on millions of data points, parallel processing is essential.[13] Since generative AI has gone mainstream, the ability to train and run models quickly has become a business imperative.

Large tech and social media companies as well as leading research, telecom, and marketing companies are deploying their own GPUs on their premises.[14] For more typical enterprises, however, using GPUs on the cloud is likely to be the most common approach. Research shows cloud GPUs reduce AI model training costs by six times and training time by five times compared with training models on traditional CPUs on the cloud (figure 1).[15] Most leading chip manufacturers are offering GPU products and services today, including AMD, Intel, and NVIDIA.

Figure 1

GPUs can reduce AI model training time and costs



Source: Deloitte analysis.

However, GPUs aren't the only specialized hardware for training AI models. Amazon offers a chip called Inferentia, which it says aims to train generative AI, including large language models. These chips are built to handle large volumes of data while using less power than traditional processing units.[16]

Google also is in the AI chip game. It offers a product it calls Tensor Processing Units, or TPUs, which it makes available through the Google Cloud service. These processors fall under the category of application-specific integrated circuits, optimized to handle matrix operations, which underlie most machine learning models.[17]

Specialized AI chips are likely to continue to gain prominence in enterprise settings in the coming months as businesses realize the value of generative AI. Increased adoption of AI may strain most organizations' existing data center infrastructure, and the higher performance of custom chips compared with general-purpose resources could become a major competitive differentiator.

This doesn't mean enterprises will reap these benefits overnight. Historically, there's always been a lag between the wide availability of specialized hardware and the development of standards and ecosystems necessary for using hardware to its fullest. It could be years before enterprises move at pace to adopt these innovations. Enterprises can develop ecosystem partnerships to prepare for emerging technologies and have ready the skills needed to take advantage of these innovations as soon as the business case is ripe.

Next: Beyond binary

The beauty of the CPU has always been its flexibility. It can power everything from spreadsheets to graphic design software. For decades, enterprises could run just about any application on commodity hardware without having to think twice.

But researchers and tech companies are developing new approaches to processing data and building entirely new worlds of possibilities in the process. One of the most promising new paradigms may be quantum computing—a technology that's been discussed for years and whose impact is becoming clearer.

Quantum annealing is likely to be one of the first enterprise-ready applications of quantum computing, promising a new route to solving optimization tasks such as the traveling salesperson problem.[18] These types of problems have traditionally been attacked using machine learning. But due to the complexity of optimization problems, the underlying math, and therefore computation, gets incredibly intricate, while still delivering less-than-perfect answers.

But quantum annealing uses the physical attributes of quantum bits to find an optimal solution, enabling quantum computers to find solutions to notoriously complex problems that involve a high number of variables—such as space launch scheduling, financial modeling, and route optimization.[19] Quantum annealing can find solutions faster while demanding less data and consuming less energy than traditional approaches.

Quantum annealing may be the first widely available application of quantum computers, but it's not likely to be the last. The technology is maturing rapidly and could soon be applied to a range of problems to which classical computers are poorly suited today. Quantum computers process information in fundamentally different ways than classical computers, which allows them to explore challenges from a different perspective. Problems involving large amounts of data over long periods of time are potentially a good fit. For example, IBM recently worked with Boeing to explore how quantum computing could be applied to engineer stronger, lighter materials and find new ways to prevent corrosion.

"It is time to look at quantum computers as tools for scientific discovery," says Katie Pizzolato, director of theory and quantum computational science at IBM Quantum.[20] "In the history of the development of classical computers, as they got bigger, we found amazing things to do with them. That's where quantum is today. The systems are getting to a size where they're competitive with classical computers, and now we need to find the problems where they provide utility."

Quantum computers represent an entirely new way of performing calculations on data compared with our current state of binary computation, but it's not the only new approach. Another promising field is neuromorphic computing. This approach takes its inspiration from the neuron-synapse connections of the human brain. Rather

than a series of transistors processing data in sequence, transistors are networked, much like brain neurons, and computing power increases with the number of connections, not just transistors. The major benefit is the potential for increased performance without increased power.[21]

Better AI applications are the most likely use case for neuromorphic computing. While it's still early days for this computing approach, it's easy to see how a computer that is modeled on the human brain could give a boost to cognitive applications. Natural language understanding, sensing, robotics, and brain-computer interfaces are all promising use cases for neuromorphic computing. The field is still relatively new, but it has the backing of computing heavyweights such as IBM, which is developing a neuromorphic chip called TrueNorth,[22] and Intel, which just introduced the second generation of its research-grade chip, Loihi.[23]

Optical computing is another promising approach. Here, processors use light waves to move and store data, rather than electrons crawling across circuit boards. The advantage is that data is literally moving at the speed of light. This field is less developed than quantum and neuromorphic computing, but research is underway at major technology companies, such as IBM and Microsoft.[24]

The common advantage to all these paradigms is using lower power than CPUs or GPUs while achieving similar, and potentially better, performance. This is likely to become even more important in the years ahead as businesses and nations as a whole push toward net-zero carbon emissions. Demand for faster and more pervasive computing is only going to increase, but simply spinning up more traditional cloud instances isn't going to be an option if businesses are serious about hitting their targets.

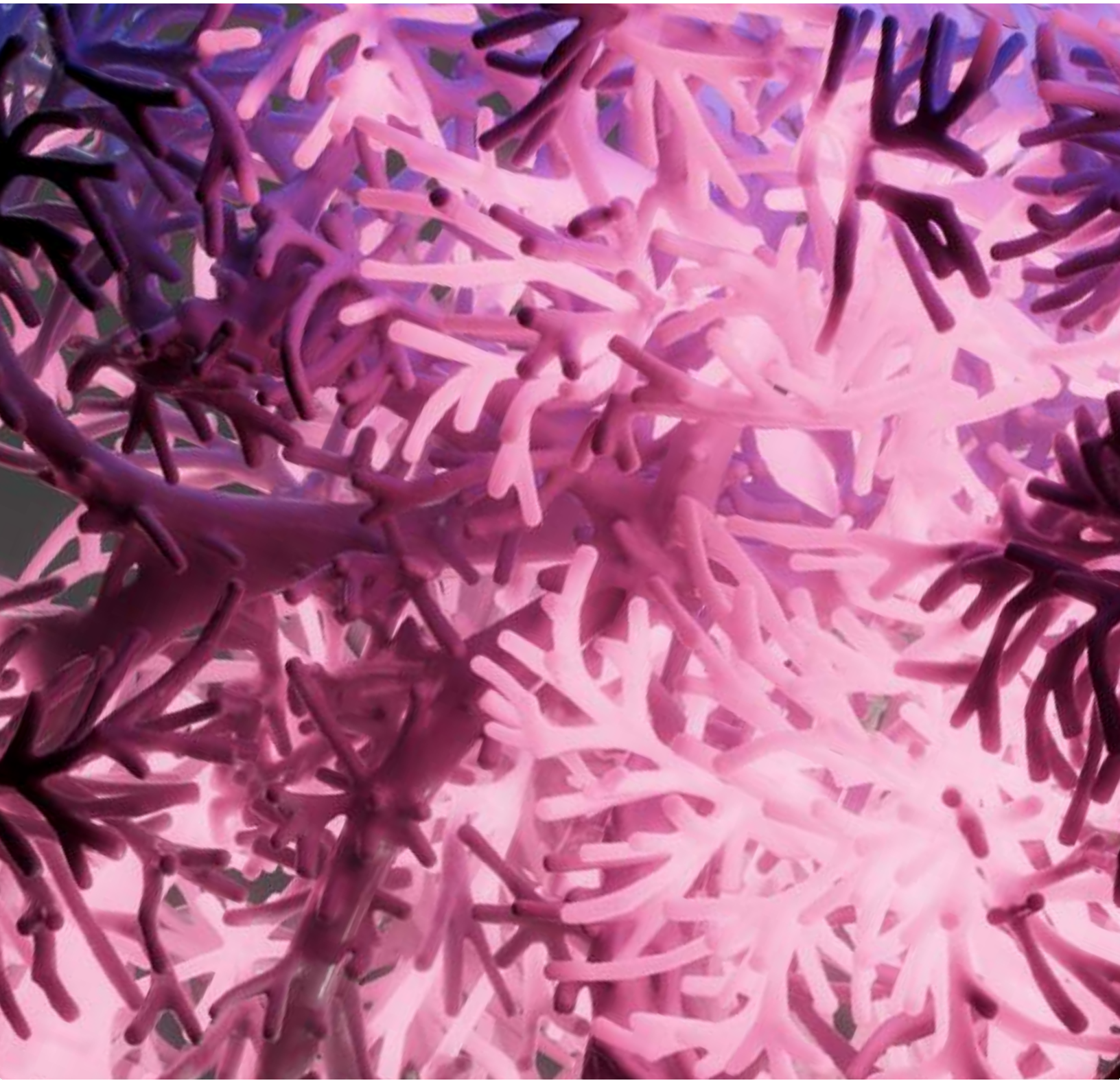
This doesn't mean these technologies are going to be a panacea for tech-related climate worries. There are still concerns around cooling and water use for quantum, and, as with any form of computing, bulky code could drive up energy requirements for technologies such as neuromorphic computing. The need for simplified code will persist, even as new computing options unfold.

These innovations aren't likely to replace CPUs at any point. Traditional computing resources remain the most useful and trustworthy tools for the vast majority of enterprise workloads, and that's not likely to change. But businesses may be able to advance their most innovative programs by incorporating some of these technologies into their infrastructure in the future. And just as we're seeing cloud services that combine CPUs and GPUs in one product today, future offerings from hyperscalers may add quantum, neuromorphic, or optical capabilities to their products, so engineers may not need to even think about what kind of hardware is running their workloads.

Our informational worlds today are defined by zeros and ones, and, without a doubt, this model has taken us far. But the future looks ready to lean into the near-limitless possibilities of not-only-digital computing, and this could drive a new era of innovation whose outlines we're only just beginning to see.

Endnotes

1. Shankar Chandrasekaran and Tanuj Agarwal, *The secret to rapid and insightful AI-GPU-accelerated computing*, Deloitte, 2022.
2. Brittanica, "Moore's law: Computer science," accessed October 31, 2023.
3. David Rotman, "We're not prepared for the end of Moore's Law," *MIT Technology Review*, February 24, 2020.
4. A16Z podcast, "AI hardware, explained," podcast, July 27, 2023.
5. Ranjit Bawa, Brian Campbell, Mike Kavis, Nicholas Merizzi, *Cloud goes vertical*, Deloitte Insights, December 7, 2021.
6. Jensen Huang, "NVIDIA GTC 2024 keynote," speech, NVIDIA, accessed October 31, 2023.
7. Christine Ahn, Brandon Cox, Goutham Balliappa, and Tanuj Agarwal, *The economics of high-performance computing*, Deloitte, 2023.
8. A16Z podcast, "AI hardware, explained."
9. Stephanie Glen, "COBOL programming skills gap thwarts modernization to Java," TechTarget, August 10, 2022.
10. Interview, Bart Mason, technology lead, Utah Office of Recover Services, July 28, 2023.
11. Interview with Brechtel Dero, division manager, Colruyt Group, August 18, 2023.
12. Ibid.
13. Ahn, Cox, Balliappa, and Agarwal, *The economics of high-performance computing*.
14. NVIDIA, "NVIDIA hopper GPUs expand reach as demand for AI grows," press release, March 21, 2023.
15. Ahn, Cox, Balliappa, and Agarwal, The economics of high-performance computing.
16. Amazon Web Services, "AWS inferentia," accessed October 31, 2023.
17. Google Cloud, "Introduction to cloud TPU," accessed October 31, 2023.
18. Cem Dilmegani, "Quantum annealing in 2023: Practical quantum computing," AIMultiple, December 22, 2022.
19. Deloitte, "Quantum annealing unleashed: Optimize your business operations," video webinar, August 3, 2023.
20. Interview, Katie Pizzolato, director of theory and quantum computational science, IBM Quantum, October 16, 2023.
21. Victoria Corless and Jan Rieck, "What are neuromorphic computers?" *Advanced Science News*, March 13, 2023.
22. Filipp Akopyan et al., *TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip*, IBM, October 1, 2023.
23. Intel Labs, "Neuromorphic computing and engineering, next wave of AI capabilities," accessed October 31, 2023.
24. Bert Jan Offrein, "Silicon photonics," IBM, accessed October 31, 2023; Microsoft, "AIM (Analog Iterative Machine)," accessed October 31, 2023.



BUSINESS OF TECHNOLOGY

From DevOps to DevEx: Empowering the engineering experience

A new focus is emerging for companies that are dedicated to attracting and retaining the best tech talent: developer experience.

With emerging technologies dominating the news, tech talent remains as important as ever to businesses. The worldwide developer population is projected to reach nearly 29 million worldwide in 2024,[1] outpacing the entire population of Australia—yet barely keeping up with the pace of demand, as we discussed in *Tech Trends 2023*.[2] Despite this growth, developer productivity is far from optimized at most organizations: Developers typically spend only 30% to 40% of their time on feature development.[3]

Shifts to Agile, DevSecOps, and cloud engineering have all become mainstream in recent years because they enhance speed, quality, and cross-functional collaboration. Now, a new focus is emerging for companies that are dedicated to attracting and retaining the best tech talent: developer experience, or DevEx, a developer-first mindset that considers each touchpoint software engineers have with the organization to improve day-to-day productivity and satisfaction.[4]

Leaders agree that a good developer experience results in better end-user and customer experiences, increasingly shifting focus from measuring speed and quantity to providing the proper tools, platforms, and feedback mechanisms and ultimately creating a culture that works for *developers*. Metrics around speed—such as lines of code or story points per developer—are giving way to more holistic measures such as time-to-first pull request (how long it takes for a developer to publish their first major batch of code), backlog changes, and defect ratios.[5] Decentralized teams and fragmented tool sets are giving way to pod structures that formalize collaboration across engineering, user experience, cyber, risk, quality management, and product teams, in addition to tailored performance management and streamlined architecture and tooling. The upside of all these changes? Eighty-one percent of companies have realized a moderate or significant impact on profitability from their investments in developer experience.[6]

Improving engineering experience can lead to a future state in which newly hired software engineers are productive from day one on the job, and a company's internal technology landscape is thoroughly integrated with its business strategy. Looking forward, companies may look to the benefits of integrated intuitive tools and realize that the investments made for developer experience may enable other aspects of the business to drive tech value.

Now: Engineers are in high demand but hindered

Digital transformation was kicked into high gear by the recent COVID-19 pandemic. Eighty-five percent of global CEOs agree that their organizations have significantly accelerated transformation after 2020,[7] and global spending on digital transformation is expected

to reach US$2.51 trillion in 2024, nearly double the amount spent in 2020.[8] This increased investment has led to an elevated role for tech leaders and employees, as discussed in our *2023 Global Technology Leadership Study*.[9] Organizations across industries—not just tech—are adding software to their core offerings and operational infrastructure. Think of, for example, auto manufacturers' autonomous driving algorithms and vehicle connectivity platforms that enable new mobility services; industrial manufacturers' use of connected equipment like turbines and generators to collect performance data, identify issues before failures, and optimize maintenance scheduling; and consumer brands' virtual try-on apps that use augmented reality to let shoppers digitally try on clothes. Software engineering excellence, and the developers who bring those capabilities, is critical for companies to capitalize on these transformation opportunities.

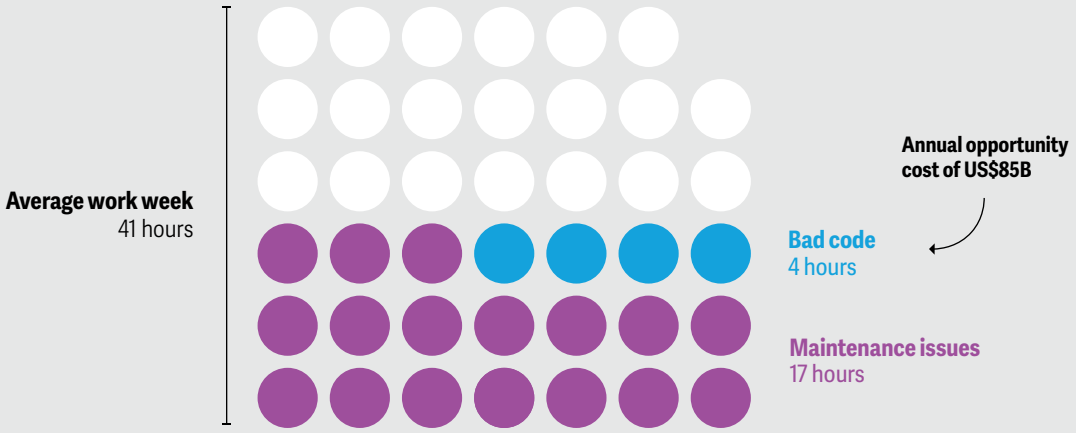
As a result, the demand for developers has skyrocketed. Jobs in software development are expected to grow by 25% within the next decade, compared with an 8% average growth rate for other professions.[10] And that's not just within the tech industry. In fact, only 10% of new software developer roles are concentrated in tech companies, as the need for digital goods and services across industries is only poised to grow in the years to come.[11]

Companies across government and commercial sectors are following the developer experience practices established by tech-forward companies to attract and retain developers. For example, the implementation of continuous integration and continuous deployment pipelines drives more frequent code releases; the notion of "shifting left" calls for adopting automation and testing earlier in the software development process;[12] and full-stack engineering exposes tech talent to the full spectrum of development on a product (front end, back end, web, and so on) through simulations, apprenticeship models, and rotations.[13]

Yet, despite the demand for software developers, many companies have not cleared the roadblocks to developer

Figure 1

Software developers suffer from productivity challenges



Source: Stripe, *The developer coefficient*, September 2018.

productivity and satisfaction (figure 1).[14] Time spent on configuration, tool integration, and debugging takes away from time spent building new features and applications that can grow revenue.[15]

Moreover, developers typically deal with a notoriously homogenous and noninclusive culture that hinders job satisfaction.[16] And on top of this, the proliferation of low-code and no-code platforms such as Appian, Outsystems, and Zoho Creator has lowered the barriers to software development, enabling the "citizen developer" movement. While this brings new opportunities, such as enabling faster innovation by decentralizing software development across the business, it could also pose potential risks around governance, security, and tech debt accumulation.

The problems that engineering leaders face when designing leading developer experiences are multifaceted. Instead of making one-off changes, a holistic change in engineering experience can help attract and retain the best talent by arming them with the tools, performance measures, and processes to succeed.

New: The DevEx difference

While the shift to DevOps focused on productivity tools and frameworks, developer experience consists of a range of mutually reinforcing capabilities that an organization provides to maximize developer productivity and satisfaction, which operate in a virtuous cycle. Developers empowered with the right tools, processes, and culture typically perform better. In fact, according to the *Harvard Business Review*, employees are 230% more engaged and 85% more likely to stay beyond three years in their jobs if they feel they have the technology that supports them at work.[17] In turn, developers who are satisfied are able to move fast, deploy code frequently, and collaborate in ways that enable efficiency.

To enable this virtuous cycle, organizations need a new, thorough framework that considers all aspects of impact to developers, not just tooling or talent (figure 2). Shifts in DevEx could then manifest in improvements in product performance and customer experience.

Figure 2

Organizations can improve engineers' effectiveness and experience through standardized capabilities

Platforms and tools

- **Architecture**
  Maximizing polyglot modularity and maintainability to promote scalability, reusability, and reliability
- **Measurement tools**
  Garnering data on platform health, product usage, and developer efficiency
- **Enablement tools**
  Creating tools to ensure collaboration and knowledge-sharing among engineers

Ways of working and flow

- **Development accelerators**
  Enhancing efficiency and reducing friction in day-to-day development activities
- **Service ownership**
  Owning and integrating responsibilities across the life cycle to reduce risk
- **Workflow management and DevSecOps**
  Increasing organization and coordination of activities to drive consistent outcomes

Talent experience

- **Community and culture**
  Promoting a fun, productive, and diverse workplace environment
- **Continuous learning**
  Developing education pathways for engineers throughout their careers
- **Career progression and development**
  Mobilizing advancement opportunities for engineers at all stages of their careers

Source: Deloitte analysis.

Platforms and tools

One aspect of establishing an effective developer experience is providing standardized platforms and tools. Though the notion may seem simple, it is far from simple in action. Developers today often wrestle with an average of more than 250 software as a service applications and other technical environments that are poorly integrated and cause fragmentation of knowledge across teams.[18] Companies can address this inefficiency through three key capabilities:

- **Architecture**—Maximizing polygot modularity and maintainability to promote scalability, reusability, and reliability.
- **Measurement tools**—Garnering data on platform health, product usage, and developer efficiency.
- **Enablement tools**—Creating tools to ensure collaboration and knowledge-sharing among engineers.

Leading organizations are acting on this trend by creating a one-stop platform for developers, where they can access a source code repository, onboarding information, documentation, tools, software development kits, and more. Only 37% of developers have access to such a portal today,[19] but Gartner estimates that by 2025, 75% of organizations with platform teams will provide self-service developer portals to improve developer experience and accelerate innovation.[20]

Ways of working and flow

Once the ideal technologies (that is, platforms and tools) are in place, the second aspect of DevEx is building clear, continuous processes for developers so they can accomplish tasks in a flow, without facing friction from disconnected systems or poor governance. Organizations can focus on three capabilities here:

- **Development accelerators**—Enhancing efficiency and reducing friction in day-to-day development activities.
- **Service ownership**—Owning and integrating responsibilities across the life cycle to reduce risk.
- **Workflow management and DevSecOps**—Increasing organization and coordination of activities to drive consistent outcomes.

The ideal developer experience would likely entail a single process and pipeline across the organization for code validation and testing, performance measurements, and safe rollbacks of code without causing outages. While cutting-edge tech organizations are working to approach this scenario, enterprises across industries are making progress on maturing their developer experience across the capabilities outlined above.

For instance, CarMax, the largest used car retailer in the United States, has found clear success in modernizing development processes.[21] The technology team replaced a project-based operating model with a product-based model made up of cross-functional teams. Instead of measuring developers on projects completed, CarMax began setting transparent quarterly objectives for more frequent delivery. It also placed a huge focus on rapid testing of products with associates and customers so that it could gather feedback and iterate before rolling out a new feature. In a similar vein, after Etsy invested 20% of its engineering budget in developer experience, it was able to scale its organization from 250 people to almost 1,000.[22]

Talent experience

Finally, for process and technology changes to be accepted, the culture must be conducive to a more modern engineering experience. Developers in many companies still specialize in traditional mainframe languages and ways of working, but others are eager to spend their time on innovation toward a purpose that resonates with them. Companies looking to attract and retain such talent can build out these capabilities:

- **Community and culture**—Promoting a fun, productive, and diverse workplace environment (much needed in most technology divisions).[23]

- **Continuous learning**—Developing education pathways for engineers throughout their careers. With more tech talent learning skills from a wide variety of resources and methodologies (including blogs, online coursework, books, and formal education), it is more important than ever for organizations to standardize onboarding and training.[24]

- **Career progression and development**—Mobilizing advancement opportunities for engineers at all stages of their careers, as discussed in last year's *Tech Trends*.[25] For example, Citibank has defined career paths for engineers who want to keep building their technical skills, allowing them to stay current with coding trends. By prioritizing technical expertise, the organization facilitates enduring careers in deep technical roles, providing technologists with diverse and compelling avenues for progression.[26]

Most importantly, a shift in culture can help companies realize that developers shouldn't be measured in the same way as other employees. Because developers are often asked to build new features and work in an experimental capacity, standards of velocity and quality won't always be accurate measures of learning or growth. Rather, tech talent needs an avenue to collectively brainstorm, learn from others, and feel connected to end goals.

CarMax paid close attention to talent experience, not just process, when undergoing its own transformation. On top of physically moving employees to sit in cross-functional teams so IT wouldn't be isolated, it organized product showcases. Every two weeks, engineers would present on technology capabilities in development, along with outcomes and lessons learned, to increase transparency and hear feedback from senior leadership. To further signify the elevated role of the technology team, the IT department was also formally renamed CarMax Technology, with a focus on business outcomes over traditional IT requirements and deadlines.

Next: Every employee is a tech employee

Companies often hope to hire "10x" engineers, those who are 10 times as productive as the average developer. But searching for unicorns in the talent market is rarely a winning strategy. Instead, with the right platforms, process, and culture in place, 10x engineers could become much less rare. Especially as generative AI continues to bolster developer productivity and opens up a future of increased workplace automation, many of today's hindrances may not be relevant in the next five to 10 years. As we mentioned in last year's trend on "serial specialists," engineers who are interested in challenging themselves can use productivity enhancements to free up their time and work on new and interesting projects and technologies over the course of their career.[27]

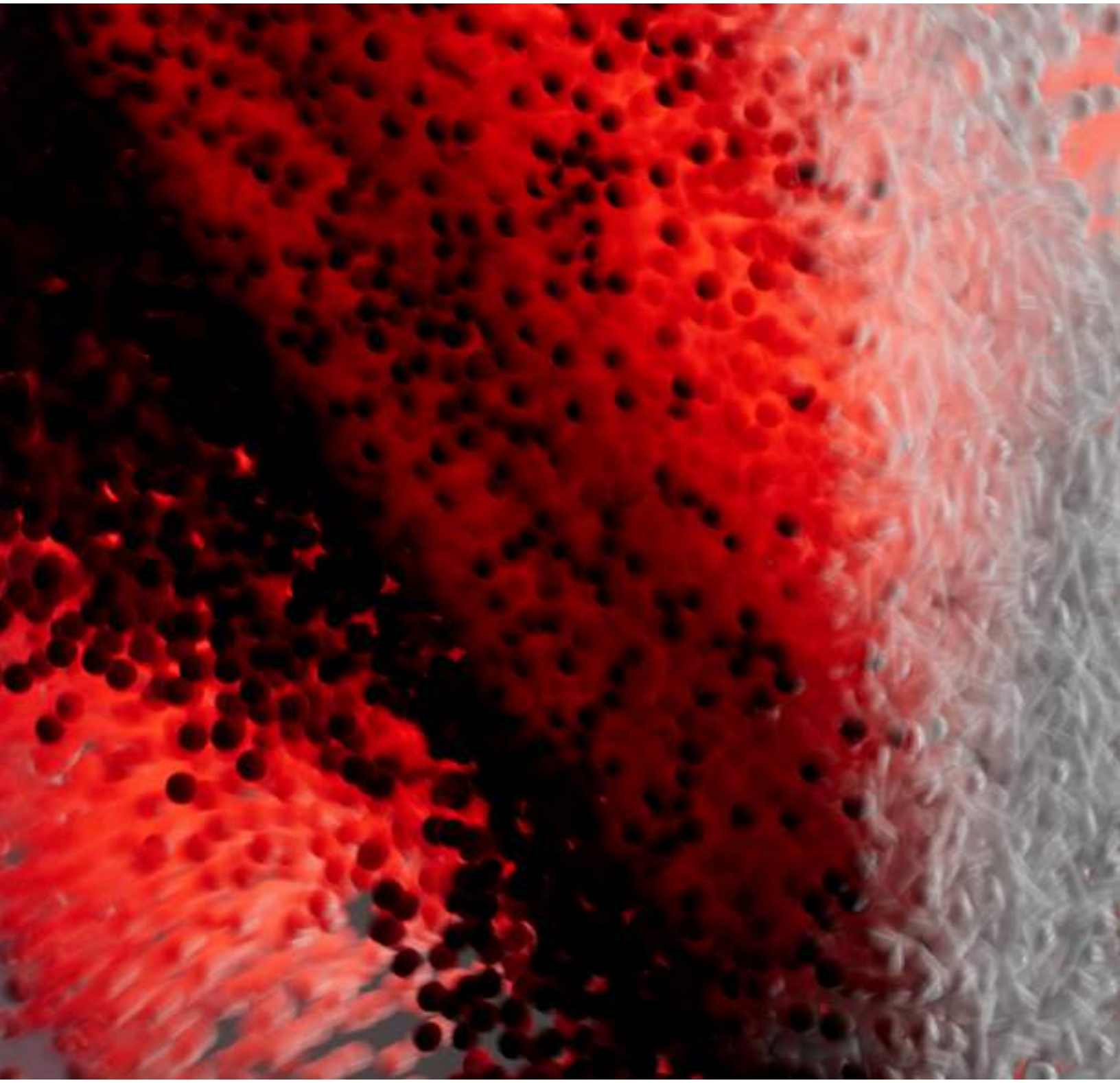
Crucially, the work organizations do in the next few years to set up a new developer experience won't be contained to the technology division. As technology itself continues to become more and more central to the business, technology tasks and required talent will likely become central as well. Standardized tools and platforms, like the ones discussed above, as well as advanced low- or no-code tech, may one day enable *all* employees of a business to become low-level engineers.

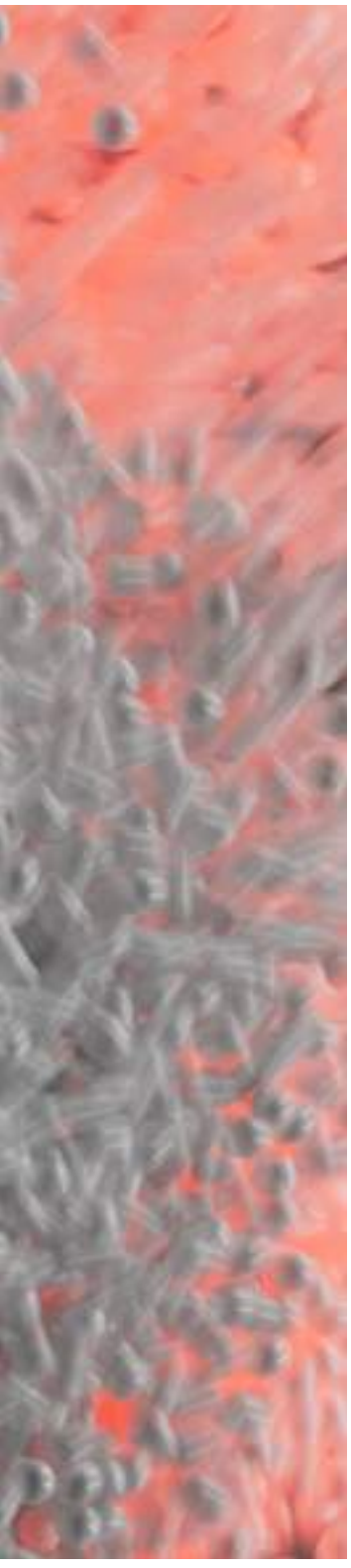
Instead of transforming from a 1x to a 10x engineer, employees outside the tech division could be going from zero to one. These citizen developers are likely to be empowered by a future where the most common programming language is not Python or Java but English, or whatever natural language they choose.

Depending on how quickly automation improves, more employees should carry out basic technology tasks in the years to come or simply oversee automated digital processes. Expanding the pool of technologists then allows experienced engineers to focus on the highly complex tasks and novel builds on which they're excited to work. By all accounts, the opportunity to focus on cutting-edge innovations and challenging problems is likely to bolster both productivity and satisfaction for the next generation of developers.

Endnotes

1. Statista, "Number of software developers worldwide in 2018 to 2024 (in millions)," 2023.
2. Deloitte Insights, *Flexibility, the best ability: Reimagining the tech workforce*, *Tech Trends 2023*, December 6, 2022.
3. Jacob Bo Tiedemann and Tanja Bach, "Why should you invest in good developer experience today," Thoughtworks, May 10, 2021.
4. Deloitte, "Accelerating developer experience (DevEx)," accessed October 2023.
5. Nolan Wright, "Three engineering performance metrics the business can understand," *Forbes*, August 5, 2019.
6. Carrie Tang, "Forrester snapshot: Platform engineering is key to reducing time to market," Humanitec Blog, March 17, 2023.
7. Deloitte Insights, *How digital transformation—and a challenging environment—are building agility and resilience*, April 29, 2021.
8. Statista, "Spending on digital transformation technologies and services worldwide from 2017 to 2026 (in trillion US dollars)," October 2022.
9. Deloitte Insights, "Global CIO and technology leadership survey collection," accessed October 2023.
10. Bureau of Labor Statistics, US Department of Labor, *Occupational Outlook Handbook*, accessed October 2023.
11. Will Markow, Jonathan Coutinho, and Andrew Bundy, *Beyond tech: The rising demand for IT skills in non-tech industries*, Burning Glass Technologies and Oracle Academy, September 2019; Steve Rogers, Kasey Lobaugh, and Anthony Waelter, *The rise of digital goods: Opportunity over threat*, Deloitte Insights, January 23, 2023.
12. Mike Kavis, "DevOps—shift everything left," Deloitte, February 28, 2018.
13. Deloitte, *Technology Skills Insights report*, accessed October 2023.
14. Stripe, "The developer coefficient," September 2018.
15. VMware Tanzu, "Developer experience: Optimizing DevOps UX," accessed October 2023.
16. Wiley Edge, *Diversity in tech: 2021 US report*, accessed October 2023.
17. Brad Anderson and Seth Patton, "In a hybrid world, your tech defines employee experience," *Harvard Business Review*, February 18, 2022.
18. Deloitte, "Accelerating developer experience (DevEx)."
19. Stack Overflow, "Developer experience: Processes, tools, and programs within an organization," accessed October 2023.
20. Gartner, "Gartner identifies the top 10 strategic technology trends for 2023," press release, October 17, 2022.
21. Deloitte Insights, *Technology transformation revs up CarMax's business*, accessed October 2023.
22. DX, "Inside Etsy's multiyear DevEx initiative | Mike Fisher (Etsy, PayPal)," podcast, April 19, 2023.
23. Deloitte, "Accelerating developer experience (DevEx)."
24. Statista, "How did you learn to code?," June 2023.
25. Deloitte Insights, *Flexibility, the best ability*.
26. Interview with Colin Heilman, global functions CTO at Citibank, October 11, 2023.
27. Ibid.





CYBER AND TRUST

Defending reality: Truth in an age of synthetic media

With the proliferation of AI tools, it's now easier than ever to impersonate and deceive, but leading organizations are responding through a mix of policies and technologies.

You may have recently seen an ad with Tom Hanks pitching a dental plan. The actor himself didn't participate in the shoot. Someone simply used his likeness, together with deepfake technology, to make it appear as though he had.[1]

Take it as a sign of the times, when anyone can be made to look as though they said or did anything. Artificially generated content, driven by rapid advances in generative AI, has reached a point where it's almost impossible for people to separate what's real from what was conjured from the depths of computers.

It's not just celebrities in the crosshairs. With the proliferation of artificial intelligence tools, it's now easier than ever for bad actors to impersonate others and deceive their targets. Many are using deepfakes to get around voice and facial recognition access controls, as well as in phishing attempts. AI applications themselves, which demand huge amounts of data, are rich targets for hackers. The security risks are multiplying with every new content-generation tool that hits the internet.

But leading organizations are responding through a mix of policies and technologies designed to identify harmful content and make their employees more aware of the risks. The same generative AI tools used by bad actors to exploit organizations can be used to identify and predict attacks, allowing enterprises to get ahead of them.

Now: The next generation of social engineering hacks

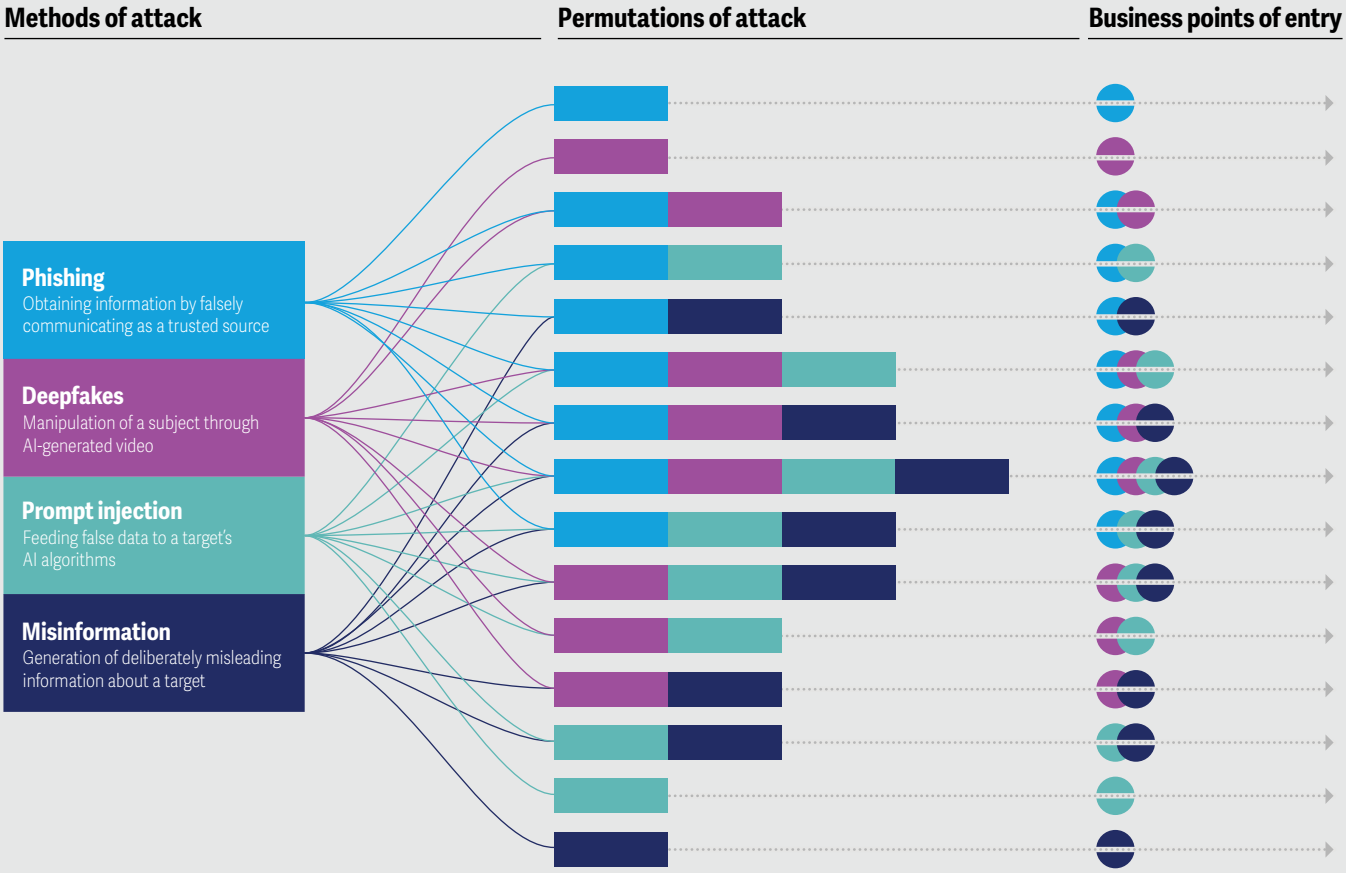
Social engineering hacks have always relied on convincing a person to hand over data or access to systems for illegitimate purposes. Though the strategy can be very effective, it also requires a lot of personal interaction between the bad actor and the victim. Artificially generated content enables attackers to create that personal touch with a much lower time investment. A wave of artificially generated content is now targeting enterprises, exploiting vulnerabilities by impersonating trusted sources. The problem is accelerating rapidly.[2]

Currently, there's a large gap between AI's ability to create realistic-sounding content and people's ability to recognize it. A majority of people say they can tell the difference between AI- and human-generated content, but another 20% aren't sure.[3] However, the first group is likely being overconfident. Few people can reliably distinguish between the two precisely because AI content generators are trained on human-created content and developed to replicate it as closely as possible.[4] People may expect artificially generated content to look or sound robotic in some way, but more than ever, it feels human.

Bad actors are likely to use artificially generated content to attack businesses in several ways (figure 1).

Figure 1

Bad actors are likely to use artificially generated content to attack businesses in several ways



Source: Deloitte analysis.

Improved phishing: Phishing is the most common type of cyberattack, with 3.4 billion spam emails sent every day. In 2021, cybercriminals stole an estimated US$44.2 million through phishing attacks.[5] Phishing attacks typically succeed not because they're high quality but because they're sent out in massive volume—out of billions of emails, eventually a few will achieve their goal. Most recipients are generally able to identify phishing attempts because of the use of poor grammar and spelling, or because the sender clearly doesn't know the recipient. But generative AI tools allow fraudsters to craft convincing, error-free messages quickly and easily and to provide relevant context, which enables them to tailor messages to each recipient, making the messages harder to ignore. The problem is likely to get worse as the quality of publicly available models improves.[6]

Deepfakes: Deepfakes have been around for years, but until fairly recently, they haven't been convincing enough to be used in cybercrimes. Now, we're starting to see them used to attack businesses. For example, the CEO of a UK-based energy firm was conned out of US$243,000 by scammers using deepfake AI voice technology to impersonate the head of the firm's parent company.[7] Deepfake

tools have advanced significantly since this incident and are likely to continue improving rapidly, making it harder for people to know with confidence with whom they are dealing.

Prompt injection: Web browsers and email clients with virtual assistants could be leveraged by bad actors who leave malicious prompts in webpages or emails that instruct the assistant to forward data such as contact lists, banking information, and health data.[8] Most types of social engineering hacks have historically worked by tricking people into handing over data or access to systems. But with prompt injection, hackers don't even need to bother with this step. The prompts execute automatically, without the victim's knowledge.

Misinformation: Social media campaigns against businesses are nothing new, but artificial content is adding fuel to the fire. AI tools can be used to create massive amounts of content quickly. Bad actors can use the tools to target enterprises, causing reputational harm or even threatening stock prices.[9] In the past, attackers had to personally craft messages, but content-generating tools now give them the ability to churn out misinformation at scale, allowing them to experiment and test out messages with the public until they find one that resonates.

The wide availability of generative AI and the pace with which content-generating models are improving are likely to supercharge these problems. For little to no cost and with virtually no technical skill, anyone will be able to create convincing media to separate businesses from their money and data.

New: Arming the enterprise against an emerging threat

None of this means enterprises are powerless against the tidal wave of artificially generated content coming their way. Leading enterprises are taking proactive steps to make sure they don't become victims.

Social engineering is nothing new, and while synthetic media may give hackers a new tool in their toolbox, many of the tried and true methods for preventing this type of attack are still applicable today. Being suspicious of online communications, verifying the identity of people with whom you're communicating, and requiring multifactor authentication to access sensitive assets are all ways enterprises can guard against this new attack vector.

As with most types of social engineering threats, tackling the problem of synthetic content starts with awareness. "While AI is exciting and there's a lot of cool things happening, it's also giving a lot of capability to cybersecurity bad actors," says Shamim Mohammed, chief information and technology officer at CarMax. "A big focus for me is making sure that we're staying current and [even] ahead so we can protect and defend our company."[10]

One way he does that is by working with a set of ecosystem partners. Mohammed says CarMax partners with both leading tech companies and AI-focused cybersecurity startups to get smart on the threat landscape and access the latest tools for preventing attacks.

"We have a very strong technology ecosystem," Mohammed says. "We work with big players who are on top of the AI revolution as well as a lot of startups that are focusing on AI. So we have the best tools available to protect our information from this emerging trend."

Effective tools are emerging to help enterprises identify potentially harmful content. Just as AI can create content, it can also assess images, video, and text for authenticity. These tools may soon be able to predict the sorts of attacks enterprises are likely to face.

When it comes to both creating and detecting artificial content, scale, diversity, and freshness of training data are paramount. When generative AI models first became publicly available, bad actors had an advantage because these models were trained by huge tech companies with access to the most powerful hardware and largest sets of training data. The first generation of detectors pushed out by large tech companies didn't match that scale while training tools to identify synthetic content.[11]

That's changing. Reality Defender, for example, trains its synthetic media detection platform on a petabyte-scale database of text, images, and audio, some of it artificially generated. When training on such a large corpus, subtle tells begin to emerge that indicate something was created by an AI tool. For example, AI-generated images often have specific deformations or pixelations. Text has a

measurable degree of predictability. These things may not be obvious to the naked eye, but an AI model trained on sufficient data can learn to reliably pick them out.

Ben Colman, CEO at Reality Defender, says being able to identify harmful content and respond to it is critical for enterprises, particularly when it comes to misinformation and disinformation campaigns that may seek to harm the business's reputation or that of its leadership. "Once something has gone viral, it's too late," he says. "If a brand is harmed in the court of public opinion, it doesn't matter if it comes out a week or two later that the content was untrue."[12]

Other tools exist to detect AI-generated content based on specific signifiers.[13] Soon, synthetic media detectors will become even more finely tuned. Intel recently introduced a deepfake detection tool that looks beyond data and analyzes videos for signs of blood flow in the faces of people in the videos. When a person's heart pumps blood through their veins, the veins change color slightly. This is something that can be measured in authentic videos but is very hard for AI models to mimic.[14]

Expect more efforts like this. According to some estimates, as much as 90% of online content will be synthetically generated by 2025.[15] Much of it will be for legitimate purposes such as marketing and customer engagement, but cybercriminals will likely use generative tools for their own advantage. It has never been more important for enterprises to be able to identify the veracity of the content their employees interact with.

Next: The cat and mouse game continues

Many organizations were quick to **add AI reinforcements to their arsenals** a couple years ago,[16] but generative AI has given bad actors a new weapon of their own. Enterprises are now catching up. Expect this process to continue in the future as new paradigms such as quantum computing mature and deepen AI's capabilities.

Quantum computing is still a few years away from being broadly available, but it is rapidly maturing, and it may well become the next tool of choice for both hackers and enterprises. One of the most promising use cases for the technology looks to be quantum machine learning. Like any tool, what matters is how you use it. It has the potential to supercharge the problem of artificially generated content but also could be a boon to enterprises' cyber defenses.

Quantum machine learning has shown the potential to generate more accurate predictive models on less training data.[17] Classical computing data exists as a binary: Data is either a 0 or a 1. But quantum data can take on more than one state at a time, allowing quantum bytes to contain richer information. When applied to machine learning, this allows for the development of much more complex models than are possible today with even the most advanced graphic processing unit hardware.[18]

This could result in hackers creating better-targeted content without needing to gather more data about their intended victims. Instead of a model requiring hundreds of hours of video training data to create a convincing deepfake of a person, a few snippets could suffice in a quantum machine learning world.

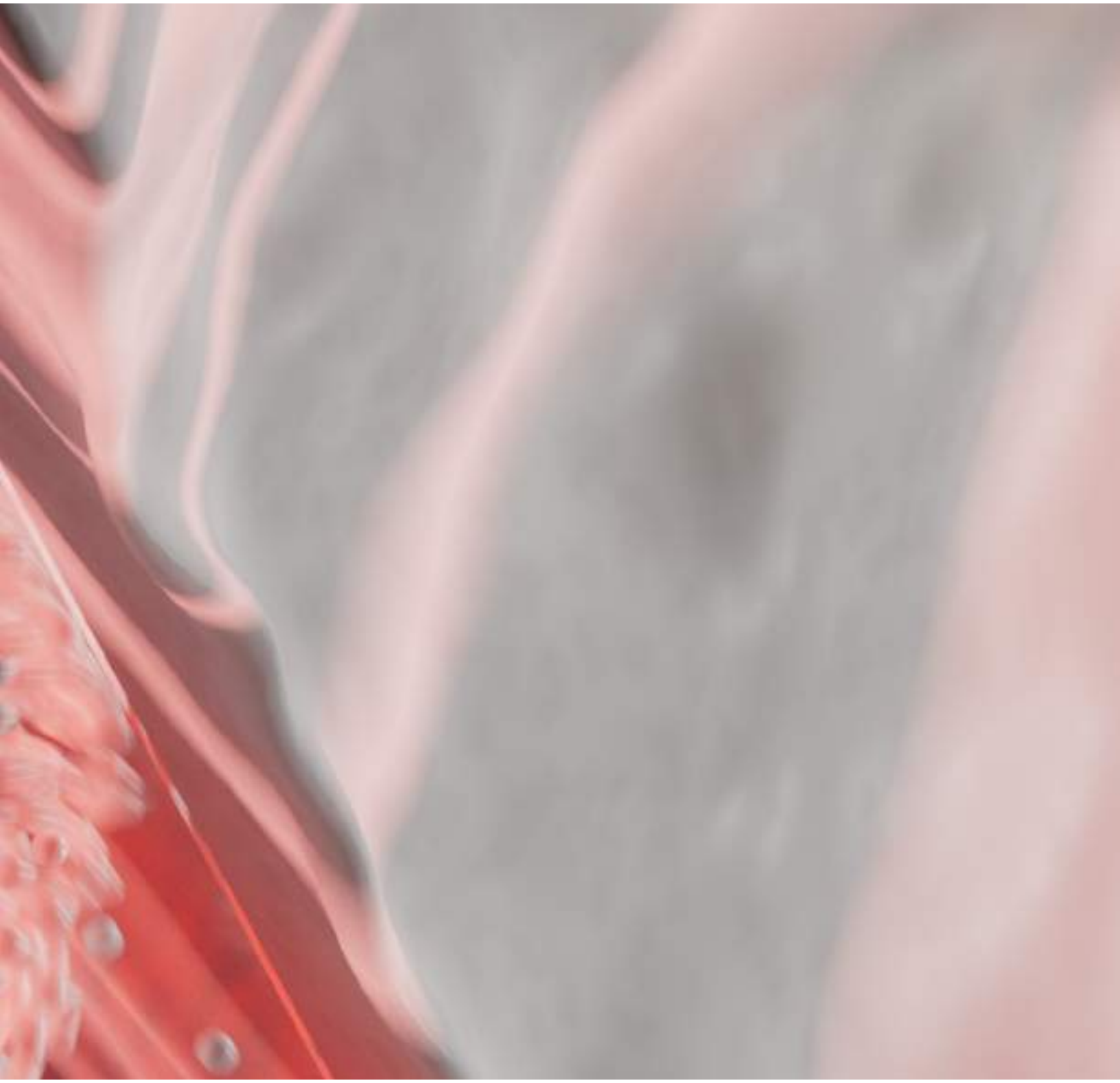
However, for enterprises looking to improve their cybersecurity, quantum machine learning could also significantly improve synthetic media detectors. Rather than requiring billions of data points before they learn to recognize artificially generated media, detectors may learn to spot fakes after seeing a handful of examples.

Quantum computers may even enable enterprises to better predict the types of attacks they're likely to face. Quantum machine learning excels at predictions, potentially exceeding classical machine learning. This is because quantum algorithms can explore the likelihood of various predictions being wrong and return an answer that is less likely to miss its mark.[19] It may seem as though predicting the source of attacks is impossible today because they can come from almost anywhere, but the maturation of quantum machine learning could help make the problem more manageable. This could put businesses in the position of preventing attacks rather than responding to them.

Enterprises need to prepare for this reality now because bad actors aren't sitting still. Getting ahead of the problem now will help keep them from being swept under by the tidal wave of artificial content that is set to come their way.

Endnotes

1. Issy Ronald and Jack Guy, "Tom Hanks says dental plan video uses 'AI version of me' without permission," *CNN Entertainment*, October 2, 2023.
2. IBM, "When it comes to cybersecurity, fight fire with fire," accessed November 6, 2023.
3. Kathy Haan, "Over 75% of consumers are concerned about misinformation from artificial intelligence," *Forbes*, July 20, 2023.
4. Pavel Korshunov and Sebastien Marcel, *Deepfake detection: Humans vs. machines*, arXiv:2009, September 7, 2020; David Ramel, "Researchers: Tools to detect AI-generated content just don't work," Virtualization & Cloud Review, July 10, 2023.
5. Charles Griffiths, "The latest 2023 phishing statistics," AAG IT, October 2, 2023.
6. Ralph Stobwasser and Nicki Koller, "On high alert: The darker side of generative AI," Deloitte, accessed November 6, 2023.
7. Catherine Stupp, "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case," *Wall Street Journal*, August 30, 2019.
8. Melissa Heikkilä, "We are hurtling toward a glitchy, spammy, scammy, AI-powered internet," *MIT Technology Review*, April 4, 2023.
9. Stobwasser and Koller, "On high alert."
10. Interview with Shamim Mohammad, executive vice president and chief information and technology officer at CarMax, August 3, 2023.
11. College of Computer, Mathematical, and Natural Sciences, "Is AI-generated content actually detectable?," University of Maryland, May 30, 2023.
12. Interview with Ben Colman, cofounder and CEO, Reality Defender, August 2023.
13. GPTZero, "Homepage," accessed November 6, 2023; Jan Hendrik Kirchner, Lama Ahmad, Scott Aaronson, and Jan Leike, "New AI classifier for indicating AI-written text," OpenAI blog, January 31, 2023.
14. Intel, "Intel introduces real-time deepfake detector," November 14, 2022.
15. Publications Office of the European Union, *Facing reality? Law enforcement and the challenge of deepfakes*, Europol Innovation Lab, 2022.
16. Ed Bowen, Wendy Frank, Deborah Golden, Michael Morris, and Kieran Norton, *Cyber AI: Real defense*, Deloitte Insights, December 7, 2021.
17. Los Alamos National Laboratory, "Simple data gets the most out of quantum machine learning," July 5, 2023.
18. Tariq M. Khan and Antonio Robless-Kelly, "Machine learning: Quantum vs. classical," *Institute of Electrical and Electronic Engineers Access* 8, 2020: pp. 219275–219294.
19. Surya Remanan, "Beginner's guide to quantum machine learning," Paperspace, 2020.



CORE MODERNIZATION

Core workout: From technical debt to technical wellness

Businesses that want to lead in the future need to forgo piecemeal approaches to technical debt in favor of a new holistic frame of technical wellness.

Over the years, *Tech Trends* has chronicled the progression of once-cutting-edge technologies as they become legacy systems in dire need of a modernization salvo. Just last year, in *Tech Trends 2023*, we discussed how mainframes can be viewed as "rusty but trusty," ripe for connecting to emerging technologies through innovative middleware.[1] In prior years, we focused on app modernization or cloud migration of aging databases[2]—in each case, championing a different aspect of the organization's core tech stack.

This year, we take a step back to broaden the concept of core systems that need to be modernized. Businesses need to deal with aging networks that can't keep up with 5G and Wi-Fi 6. Their data centers are still being shifted to cloud, but data management needs to be cleaned up for generative AI's primetime. Enterprise resource planning (ERP) vendors are pushing out new versions that require significant upgrades. Even relatively recent software-as-a-service (SaaS) implementations that were supposed to be a remedy for years of legacy core woes aren't aging well. On top of all this, companies are dealing with a mix of contractors, captives, and traditional workers who aren't ready to pivot to modern engineering.

Until now, companies have been monitoring their technical debt—the implied cost of not modernizing systems and working with suboptimal performance—through piecemeal assessments. But those that aim to lead in the future will need a more continuous and complete view of their core. Instead of watching their technical debt accumulate and overwhelm their systems, this view can provide guidance on which technology upgrades matter and when and why to make them. Going forward, the best proxy for understanding your legacy technology realities may not be debt at all, but *health*.

From a health and wellness mindset, organizations could treat disparate technology systems (cyber, data, infrastructure) like parts of a body, subject to thorough annual checkups, similar to those conducted by integrated medical care providers. Instead of fixing aging systems that break or bottleneck IT's progress one at a time, teams can use preventative health assessments to identify and prioritize the areas of the tech stack that need treatment. These assessments can be rooted in real business problems: increasing costs and risks and stifled innovation and growth. For example, some aspects of core systems, such as the mainframes we discussed last year, may be in good health, only needing connectors to keep doing what they do best. Others may be due for a thorough upgrade or replacement.

Moving away from previously siloed modernization efforts, businesses are likely to have a highly customized and integrated wellness plan across their tech stack in the coming years. After all, today's white-hot innovations will likely continue to become tomorrow's legacy systems, in need of checkups, especially at the current pace of technology innovation.

Now: Yesterday's innovations are aging quickly

Whether they are ERP systems or data centers, technologies that once revolutionized business are now prone to slowing it down. Up to 70% of technology leaders view technical debt as a hindrance to their organization's ability to innovate and the No. 1 cause of productivity loss.[3] Perhaps the population that suffers from this most directly is software developers, who spend an estimated 33% of their time dealing with technical debt maintenance.[4] This time spent can also have an outsized impact on the productivity and satisfaction of developers, as discussed in this year's trend on developer experience. As many as 78% of developers felt that spending too much time on legacy systems had a negative impact on morale; other impacts cited were employee and customer churn along with lost deals.[5]

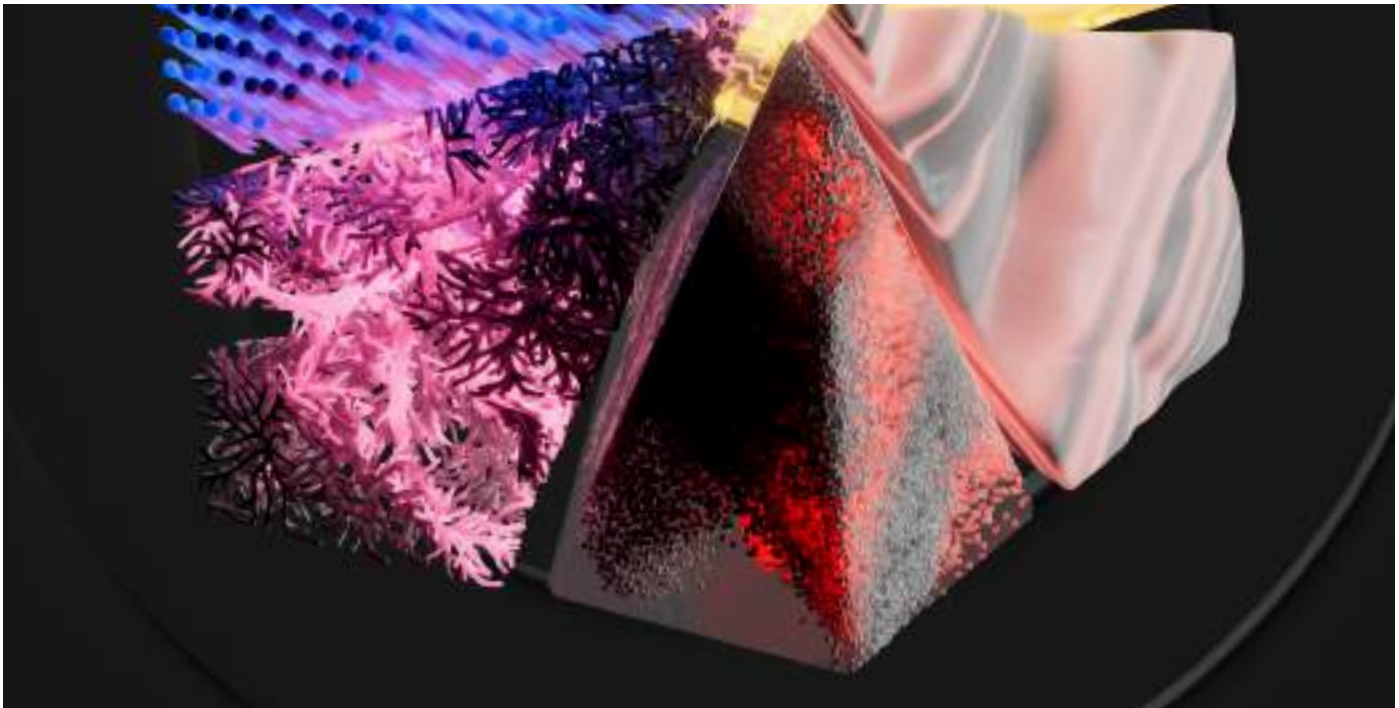
As technology's rapid advancement increases (as evidenced by generative AI), businesses and government agencies tend to struggle with two competing truths. On one hand, they deeply believe their future business models, existing products and services, and internal operations will be fundamentally transformed (or disrupted) by technology. On the other hand, they often struggle to make the necessary investments in their infrastructure, data, applications, cyber, and workforce capabilities in a way that could adapt to that future. This often leads to technical debt sprawl: In 2022, the estimated cost of technical debt in just the United States had grown to US$1.5 trillion, despite chief information officers spending 10% to 20% of their budgets on resolving issues related to outdated systems.[6]

Many companies have experienced subpar transformation programs that amounted to massive bets on a single dimension of their core systems, which ultimately failed to deliver the benefits promised. Instead of undertaking random acts of innovation or "low-hanging" transformation investments, technology leaders may need to face a hard truth: Their technology house is ailing. And they need new ideas for where to focus time and effort so they can begin to heal.

To move forward, leaders can look for a smarter way to invest in modernization: a systematic assessment of their company's needs, strengths, and budget across the key areas of modernization spending. A holistic view of their organization's systems, grounded in real business context, can help them eschew ever-growing technical debt for a more long-term view of technical health, one that improves over time and provides more confidence to business and technology teams alike.

New: Core health gets a checkup

Because organizations are unclear which sources of technical debt are causing the most drag, issues are often underprioritized or poorly managed.[7] In reality, out of hundreds of applications or systems in a company's core, just a dozen applications that have a handful of issues may be driving the bulk of the impact of outdated systems. Instead of applying funding every year that seems to go nowhere, companies could benefit from taking a step back to tackle key issues.



THE COSTS AND RISKS OF TECHNICAL DEBT

The investment drivers of a more holistic core modernization strategy span a variety of barriers, costs, and potential risks that companies may face by keeping legacy technology in place:

- **Direct costs:** The capital and operating expenses directly associated with maintaining legacy hardware and software, such as expiring support licenses or contract terms, as well as the workforce (both full-time employees and third parties) that are directly tied to legacy system upkeep
- **Indirect costs:** Operating expenses incurred due to inefficiencies in the legacy technology environment, such as (usually untracked) time and effort spent manipulating data and analytics between legacy systems and modern applications
- **Time-to-market impact:** Delays to or dilution of business initiatives caused by complexity or inefficiency in the legacy technology stack
- **Barriers to innovation:** Constraints to growth initiatives due to deficiencies of legacy technology, such as the inability to process business-to-consumer orders, or language or currency restrictions
- **Operational risk:** Potential future impediments to business scalability, reliability, and performance due to technical limitations
- **Security risk:** Susceptibility to cyber threats, since legacy systems may not have the same security capabilities as modern technologies
- **Talent risk:** Festering technical debt and antiquated tools that make it difficult to attract and retain top technology and business talent

A core modernization framework built around technical wellness can be more comprehensive than the traditional framing of debt. In such a framework, the focus is placed on preventive care—in this case, using enhanced tracking, measurement, and predictions to address suboptimal legacy technologies before they become larger issues for the business. Rather than accumulating debt and paying it off periodically with large but perhaps ineffective investments, a wellness framing would encourage businesses to iteratively pinpoint their technical ailments and predict when investments would be most effective, based on cost, operational risk, and innovation readiness.

The focus of such a wellness diagnostic, or core checkup, would be on the five areas of statistically largest spend and biggest opportunity. Each has a current target for modernization, which is likely to change over time as emerging technologies either become more sophisticated or new ones become popular.

Infrastructure

Infrastructure is the broadest category and often the most difficult and expensive area. Fortunately, as evidenced by the State of Utah, entire mainframe systems can be migrated to the cloud within 18 months if an organization is aligned on its transformation goals.[8] Within this category, technical wellness translates to mainframes, servers, and end-user devices (such as virtual desktops) being migrated to the cloud across technical environments (sandbox, quality assurance, production). In addition, aging fiber, LAN, and WAN networks across facilities (such as data centers and corporate offices) are on a journey to modernize to 5G, Wi-Fi 6+, low-energy Bluetooth, and satellite communications. These upgrades enable companies to take advantage of private networks, advances in software-defined networking, and other advanced connectivity offerings.

Data

Data life cycles (including cleansing, manipulation, and management) and data storage constitute this category. Businesses need to streamline data cleansing and manipulation through automation so they can spend less time on data management and more time on analyzing insights. Reporting on data usage and cleanliness, especially from a trust perspective, is also key, since AI models are only as good as the data they ingest.

Storage across data centers, offices, and remote assets can be modernized to cloud storage systems and data can even be streamed in real time. Amazon Web Services recently established streaming data pipelines that move data from connected devices across multiple sources to centralized repositories where it can be better leveraged, ensuring that users are working with the freshest information possible.[9] This data can then be used in

applications such as predictive maintenance, environmental monitoring, and smart city management. This application opportunity rests on a vast landscape of data stores, lakes, spindles, and drives—each carrying complexity and very real costs.

"Anything outside of using real-time data becomes very frustrating for the end consumer and feels unnatural now," says Mindy Ferguson, vice president, messaging and streaming, at Amazon Web Services. "Having real-time data always available is becoming an expectation for customers. It's the world we're living in."[10]

Applications

This broad category includes legacy custom applications that organizations have been modernizing over time through one or more of the "five Rs": replatforming, revitalizing, remediating, replacing, and retrenching. It also includes package applications, such as ERP and SaaS applications, which require a clear upgrade strategy as vendors continue to improve their offerings, while dealing with the inevitable litany of customizations that complicate upgrade paths and integration.

Operations technology applications and product technology stacks, such as embedded products and digital offerings for customers, would also be considered for checkups in this category of modernization.

Workforce

Many companies are struggling with a workforce that is a mix of internal and third-party contractors who may or may not be ready to bring about the modernization efforts described above. To improve talent acquisition and retention of tech teams, leaders need to prioritize the modern engineering experience, bolstered by investments in tooling (across the software development life cycle), processes, and culture, as discussed in our trend "*From DevOps to DevEx*."

Cyber risk and trust

Finally, companies need to consider their relative health in cybersecurity across multiple areas: security and privacy, regulatory compliance, and ethics and morality. The first two areas can be tracked and improved through cybersecurity automation, especially to keep up with the growing amount of artificially generated content, as discussed further in our trend "*Defending reality*." Ethics requires a more nuanced approach, and businesses should keep up to date on the latest thinking on technology's potential harms to society.[11]

The benefits of a core wellness checkup across these five areas would be financial as well as intangible. For example, leaders who actively manage and reduce technical debt are expected to achieve at least 50% faster service delivery to the business.[12] And the time returned to developers could result in many more features being developed to generate revenue from customers or efficiency from employees. Perhaps, most of all, an accurate tracking system for technical debt could allow organizations some peace of mind in knowing when and how to prioritize their investments instead of scrambling to keep up with the market.

Next: The core heals itself

As modernization needs progress over the next decade, what if technology could become adaptive and resilient, able to "heal" its own outdated code or system without limited intervention?

The idea of self-healing systems is not new. In the natural world, whether at the micro level (such as a broken bone healing itself) or a macro level (for example, an entire ecosystem rebuilding itself after a forest fire), nature has shown us the pinnacle of resilient design. It's no surprise, then, that the field of biomimetics—design inspired by nature—has received more attention in recent years, and the applications within technology have already begun.[13]

For instance, self-healing raw materials such as ion gels have used clotting properties to heal damaged robot pieces, such as arms and hands, when the robot senses a cut in its material.[14] This same process is being replicated with electrical circuits as well: When an electrical circuit is damaged, a capsule of liquid metal can be released automatically into the circuit to repair the electric connection.[15]

Crucially, self-healing systems are slowly graduating out of the world of atoms and into the world of bits. Consider the example of adaptive AI, which has

progressed from human-initiated machine learning to unsupervised machine learning.[16] This AI can not only solve challenges but also, in studying those challenges, learns to teach and reprogram itself by developing more advanced problems.

Along these lines, core modernization solutions are also poised to become adaptive. AI embedded into core systems can currently diagnose tech debt accumulation in tech stacks and support engineers as they write the necessary code to modernize (while also streamlining remedial tasks and compliance, which often pile up when tech debt increases).[17] In fact, a recent global survey by Deloitte indicated that around 60% of organizations are already using AI to optimize code and identify bugs, while 50% are using it to manage code environments.[18]

Similar to a physician in training, these AI solutions for tech wellness are still prone to error and misdiagnoses; for example, they may be less effective at refactoring than debugging,[19] but as they spend more time in their residency of core systems, they're bound to improve. One day, AI could diagnose inefficiencies, develop a solution, and implement the solution without ever needing the support of human engineers.

As such innovations continue, longevity could be built into the five areas of core modernization from the onset. As more of the technology stack becomes software-defined, efforts to embed fault expectancy, monitoring, and self-healing could improve the "aging process" of our technology assets.[20] As with human wellness, the goal of tech wellness would be for core systems to age gracefully, with built-in supports and checkups to allow them to fulfill their purpose.