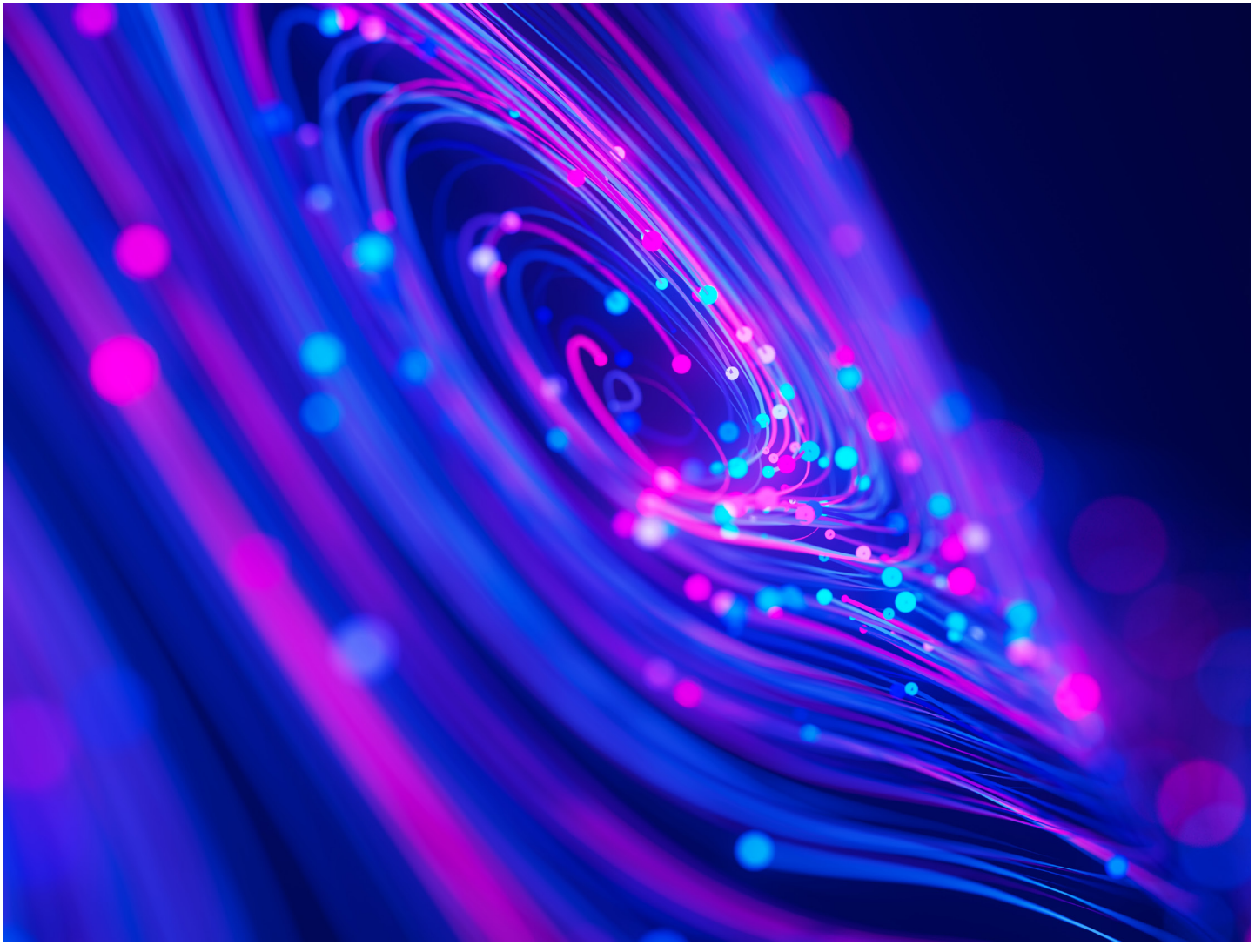
MIT Technology Review Insights

Produced in partnership with

arm

Understanding the evolving compute landscape of tomorrow.

The road to artificial general intelligence



Artificial intelligence models that can discover drugs and write code still fail at puzzles a lay person can master in minutes. This phenomenon sits at the heart of the challenge of artificial general intelligence (AGI). Can today's AI revolution produce models that rival or surpass human intelligence across all domains? If so, what underlying enablers – whether hardware, software, or the orchestration of both – would be needed to power them?

Dario Amodei, co-founder of Anthropic, predicts some form of "powerful AI" could come **as early as 2026**, with properties that include Nobel Prize-level domain intelligence; the ability to switch between interfaces like text, audio, and the physical world; and the autonomy to reason toward goals, rather than responding to questions and prompts as they do now. Sam Altman, chief executive of OpenAI, believes AGI-like properties are already "**coming into view**," unlocking a societal transformation on par with electricity and the internet. He credits progress to continuous gains in training, data, and compute, along with falling costs, and a socioeconomic value that is "**super-exponential**."

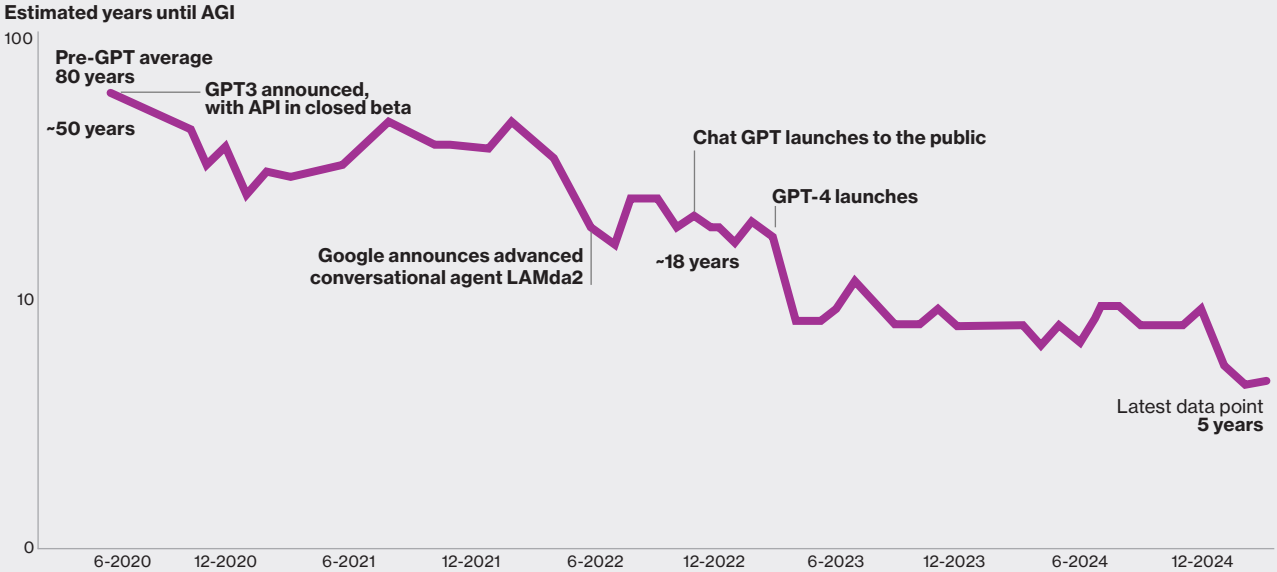
Key takeaways

1. While timelines are still unclear, AI that can match or exceed human intelligence in all domains is on the horizon. Breakthroughs in model adaptability, reasoning, and decision-making are needed to achieve this transformational potential.
2. Advances in the AI compute stack, including software, hardware, and infrastructure will enable artificial general intelligence, along with architectures that are optimized for computational efficiency and energy consumption.
3. Heterogeneous computing offers a balanced, scalable, and practical path towards AGI by leveraging the right processor for the use case and demand.

Optimism is not confined to founders. **Aggregate forecasts** give at least a 50% chance of AI systems achieving several AGI milestones by 2028. The chance of unaided machines outperforming humans in every possible task is estimated at 10% by 2027, and 50% by 2047, according to **one expert survey**. Time horizons

Figure 1: Forecasting the future of AGI

According to community predictions, AI systems that can achieve several AGI milestones seem within the range of possibility by 2030.



Source: Compiled by MIT Technology Review Insights with data from Metaculus via 80,000 Hours, 2025

"Intelligence is the ability to recombine what you already know into new patterns, to solve novel problems efficiently."

François Chollet, Co-Founder, Ndea, and Co-Creator, ARC Prize

shorten with each breakthrough, from 50 years at the time of GPT-3's launch to five years by the end of 2024 (see Figure 1). "Large language and reasoning models are transforming nearly every industry," says Ian Bratt, vice president of machine learning technology and fellow at Arm.

From autonomy to intelligence

History is, of course, full of technology predictions proven wrong with the course of time. In 1901, aviation pioneer Wilbur Wright told his brother Orville that **human flight** would take 50 years. It only took two. Conversely, cognitive and computer scientist Marvin Minsky predicted the **arrival of human-like computer intelligence by 1978**. Even today's AI models, impressive as they are, cannot rival humans across all domains.

OpenAI defines AGI as the autonomy **to outperform humans** in the realm of economically valuable work. On that reading, there are signs of its arrival in specialized fields, from law to medicine, as well as into our daily lives as consumers. "We're looking at a world where everything we do, when we're driving our car or talking to our TV or computer, is augmented by highly intelligent systems," says Bratt.

This acceleration in autonomous capability requires tooling and hardware that support model training across datasets, systems, and domains. This is because such systems operate with minimal human intervention, using feedback loops guided by objectives. It also needs innovations in large-scale training to balance generalizability with the ability to perform specialized tasks at a high level.

To perform at scale, autonomous systems must process diverse inputs and perform across platforms. A **heterogeneous computing** approach leverages the

Capabilities AI must master

AGI aims to mimic the cognitive abilities of the human brain, able to understand or learn any intellectual task that a person can. Still currently theoretical, true AGI should be able to transfer knowledge and skills learned in one domain to another, adapting to unknown situations and drawing on common sense. According to McKinsey, AI systems still fall short of humans in a number of different abilities, including:

Visual perception

AI systems still have poor color consistency and can be easily confused.

Audio perception

Current AI is not as capable of extracting and processing sound, including spatial characteristics.

Fine motor skills

AI-powered robots are still unable to achieve the same level of fine motor skills, like performing surgery.

Natural language processing

AI tools still lack a full comprehension of meaning, context, and nuance.

Problem-solving

AGI systems need to be able to assess problems without explicit programming from humans.

Navigation

We are still far off robot systems that can navigate autonomously with no human instruction.

Creativity

Current AI creativity still does not rival that of humans. AI also needs to rewrite and improve its own code.

Social and emotional engagement

AGI needs skills like interpreting facial expressions or changes in tone. Empathy is still a distant prospect.

Source: Compiled by MIT Technology Review Insights with data from McKinsey & Company, 2025

right tool for the right job. This combines processors such as central processing units (CPUs) and graphic processing units (GPUs), and specialized AI accelerators like neural processing units (NPUs), and tensor processing units (TPUs), with emerging architectures. Each processor delivers the right computation for the task, contributing a balanced, scalable, and practical path towards an AGI future, with each computational engine contributing to the right use case and demand.

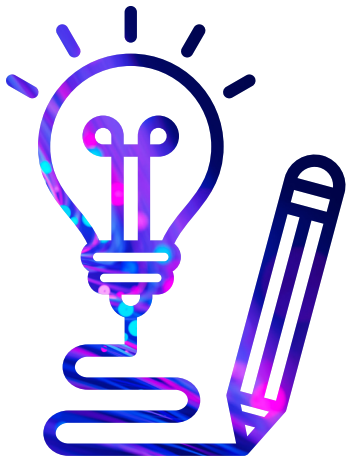
To orchestrate higher levels of autonomy also requires software tools and frameworks to manage, coordinate, and optimize the various processing chips that characterize a heterogeneous environment. AI software is continually improving to simplify work for developers, allowing them to write programs that target multiple hardware backends without rewriting code for each and access optimized performance for their apps and workloads. Orchestration tools are therefore increasingly critical to streamline workflow development, helping AI applications distribute and manage workloads dynamically across computing environments. Finally, the development of standards and protocols is helping AI applications interact with external tools, systems, and data sources, another key layer of capabilities for an AI-infused future.

Defining intelligence

A higher level of autonomy is not the same as true intelligence, according to François Chollet, co-founder of Ndea and co-creator of the ARC Prize, an intelligence benchmark that assesses the ability to adapt to reasoning tasks that have not been seen before.

"Intelligence is the ability to recombine what you already know into new patterns, to solve novel problems efficiently," Chollet argues. On ARC-AGI-2, a benchmark test, pure large language models (LLMs) score 0%, he explains. AI reasoning systems score single digit percentages. Humans can solve nearly every task. "The test shows how much more adaptable human intelligence is compared to the most powerful models we have today," Chollet says.

Bratt sees parallels between AI and the development dynamics of the human brain, codified in the human plasticity curve, which captures how brain functionality develops in the early years. The first phase is sensory systems of vision and sound, followed by language, and eventually by cognition, such as the ability to engage in



"Assessing intelligence only on narrow definitions of human achievement misses a spectrum of capabilities."

Rumman Chowdhury, Chief Executive Officer, Humane Intelligence

forward-looking planning. Just like the infant human brain, the first AI networks were focused on vision and sensor systems, followed by natural language user interfaces, and we are now at the cusp of cognition. AI is "still in the toddler phase" says Bratt, and like humans, the cognitive leap will take far longer.

Achieving human-like intelligence requires progress across domains. One is inference, the process where trained models apply learned knowledge to new data and make predictions or decisions based on acquired knowledge. Inference enables models to adapt pre-conceived knowledge to new problems across various contexts.

A second is adaptability and knowledge transfer. While AI systems of the past excelled in specialized domains, they struggled when faced with unfamiliar problems or moving across challenges like learning to play a new board game after reading the rules, solving mathematical problems, or explaining social situations within the same framework. Achieving this requires breakthroughs in machine learning, reasoning, memory, and decision making.

Some of these capabilities were demonstrated at scale in LLMs like GPT-4, where capabilities in understanding and generating text demonstrated a form of general intelligence in language processing. Small language models offer similar capabilities with reduced

computational demands to be more efficient and enable deployment on devices with limited resources, such as smartphones and edge devices.

The development of multimodal models that integrate text, vision, and other data types further enhances the generalization capabilities of contemporary AI, where models can understand and generate content across different modalities to bring AI closer to human-like understanding and interaction. However, these models still fall short of true generalized intelligence because they lack the ability to autonomously generate new knowledge or adapt outside their training data.

"Base models have no fluid intelligence whatsoever," argues Chollet. Despite advances in LLMs and scaling, they "have no ability on their own to recombine what they know into the solution of a new problem" and he explains that current models "are very good at memorizing patterns, but not very good at coming up with new patterns on the fly."

The industry may even need a broader framing of what computer intelligence means, according to Rumman Chowdhury, chief executive officer of Humane Intelligence, a non-profit that focuses on model evaluation. "We have methodologies for defining intelligence, whether animal, insect, mycelial, even extra-terrestrial, and AI does not meet these," she argues. "Animal intelligence assessments are not based on whether monkeys can pass the Medical College Admission Test (MCAT), because assessing intelligence only on narrow definitions of human achievement misses a spectrum of capabilities," Chowdhury adds. Measurement benchmarks that include fluid intelligence such as adaptability and flexibility, social intelligence such as social understanding, and embodied intelligence, such as environmental perception, could provide a better measure of AGI, while being receptive to novel forms of intelligence distinct from humans.

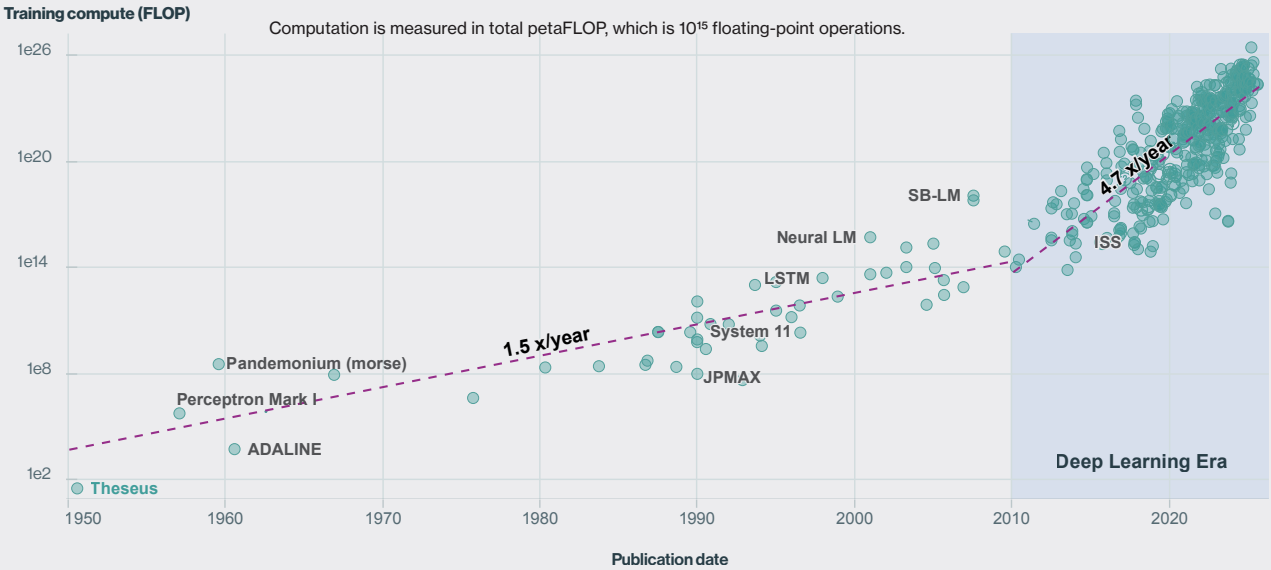
The compute imperative

One debate that runs through the AGI timeline discussion is whether adding more computation can solve the intelligence problem. Before 2010, AI compute requirements doubled every 21 months, following a predictable trajectory. When deep learning arrived, that timeline collapsed to just **5.7 months** (see Figure 2).

Large-scale AI models use about 100 times more compute than contemporaneous AI models. AGI could **require computing power of more than 10^16 teraflops**. Some estimates place the cost in excess of the entire US GDP by **2037**. "As algorithms improve and workloads scale, the need for efficient silicon keeps rising," Bratt says.

Figure 2: Growth in AI compute demand

There has been exponential growth of computation in the training of notable AI systems since 2010.

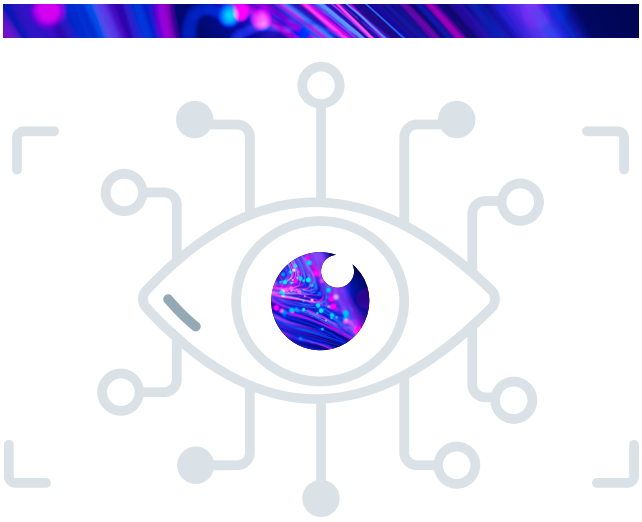


Source: Compiled by MIT Technology Review Insights with data from Epoch, 2025

Progress in AI accelerators, including NPUs, TPUs, and emerging architectures, is needed to optimize computational efficiency and energy consumption, while specialized hardware must constantly evolve to keep pace.

"If we want to enable that far future where everything is cognitively enhanced, it's going to be a huge step function increase [in compute]," predicts Bratt. "In that future world, we're going to have AI running at all different layers of the stack, from the tiniest devices all the way up to the cloud. If we want to enable that, we as a community and ecosystem must have a relentless focus on improving the energy efficiency of our platforms and unlocking that with software and software enablement."

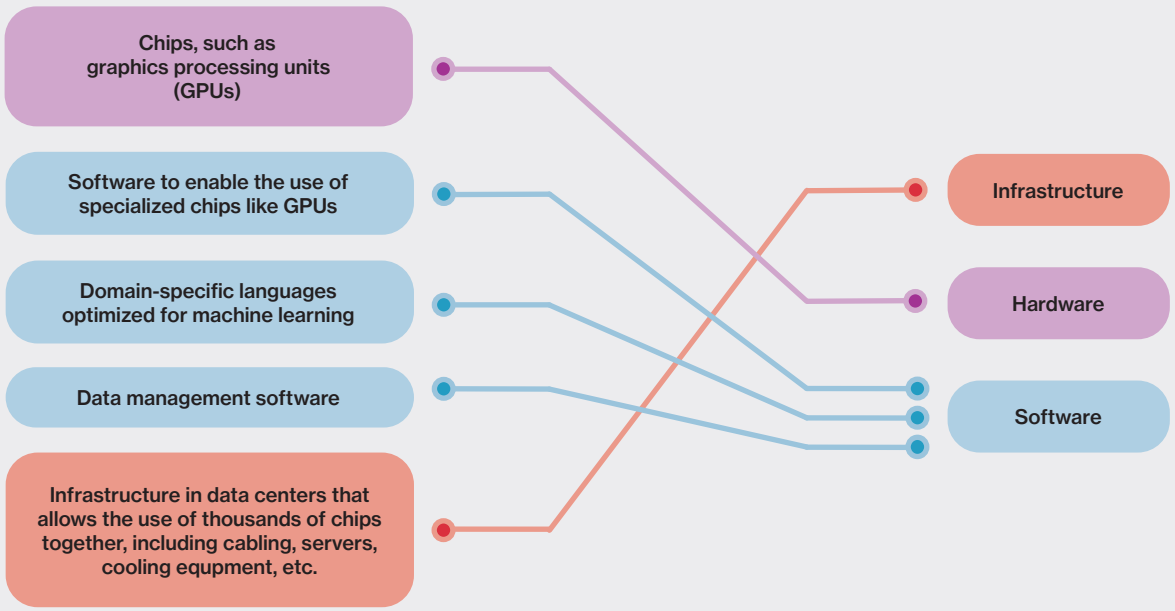
Breakthroughs in software and hardware, algorithms and chips – the AI compute stack – will be crucial (see Figure 3). But rather than trying to achieve AGI with compute alone, success will require more efficient, distributed, and context-aware compute architectures that are optimized not just for speed, but for latency, bandwidth, and energy. Bratt notes that the future of AI processing will likely entail a heterogeneous approach, with lightweight models for some workloads and a constant improvement in compute efficiency. "It has to be a combination of hardware and software. We need to collaborate with and support the whole ecosystem".



"If we want to enable that far future where everything is cognitively enhanced, it's going to be a huge step function increase in compute."

Ian Bratt, Vice President of ML Technology and Fellow, Arm

Figure 3: The AI compute stack



Source: Compiled by MIT Technology Review Insights with data from AI Now Institute, 2025

Chollet believes further intellectual progress will also be critical. “I don’t think compute is the bottleneck. The bottleneck is really ideas. When you have the right ideas, they’re going to be compute efficient by definition, because intelligence is efficiency. So the good news is that the compute that we are producing will be reusable.”

AGI might even require new breakthroughs akin to the deep learning revolution or the language and image prediction capabilities behind generative AI today, especially in the areas of reasoning, memory, and learning. “AGI, defined as something that can perform across domains of human intelligence and knowledge and act in novel ways to novel problems, doesn’t work well when you are working on a map of probabilities based on multidimensional representations of how language works,” argues Henry Ajder, an AI expert and founder of Latent Space Advisory, referring to the technical foundation of today’s large language models. “Reaching AGI could be about a new architecture, just like transformers led to significant advances in generative AI.”

The path to AGI will likely require a synthesis of different philosophies and approaches, in turn requiring collaboration across the industry to create a unified ecosystem. Advances in specialized hardware and newer computing approaches, like heterogeneous compute, could make scaling more efficient. Breakthroughs in architectural design could unlock new forms of reasoning and understanding. And the debate about the shape and meaning of AGI itself drives progress. As Chowdhury’s call for novel intelligence frameworks suggests, the journey to AGI may teach us as much about the nature of intelligence as it does about building smarter machines.

“Reaching AGI could be about a new architecture, just like transformers led to significant advances in generative AI.”

Henry Ajder, AI Expert and Founder, Latent Space Advisory

"The road to artificial general intelligence" is an executive briefing paper by MIT Technology Review Insights. Laurel Ruma was the editor of this report, and Nicola Crepaldi was the publisher. MIT Technology Review Insights has independently collected and reported on all findings contained in this paper. We would like to thank the sponsor, Arm, as well as the following experts for their time and insights:

Henry Ajder, AI Expert and Founder, Latent Space Advisory

Ian Bratt, Vice President of ML Technology and Fellow, Arm

François Chollet, Co-Founder, Ndea, and Co-Creator, ARC Prize

Rumman Chowdhury, Chief Executive Officer, Humane Intelligence

About MIT Technology Review Insights

MIT Technology Review Insights is the custom publishing division of MIT Technology Review, the world's longest-running technology magazine, backed by the world's foremost technology institution – producing live events and research on the leading technology and business challenges of the day. Insights conducts qualitative and quantitative research and analysis in the US and abroad and publishes a wide variety of content, including articles, reports, infographics, videos, and podcasts. This content was researched, designed, and written entirely by human writers, editors, analysts, and illustrators. This includes the writing of surveys and collection of data for surveys. AI tools that may have been used were limited to secondary production processes that passed thorough human review.

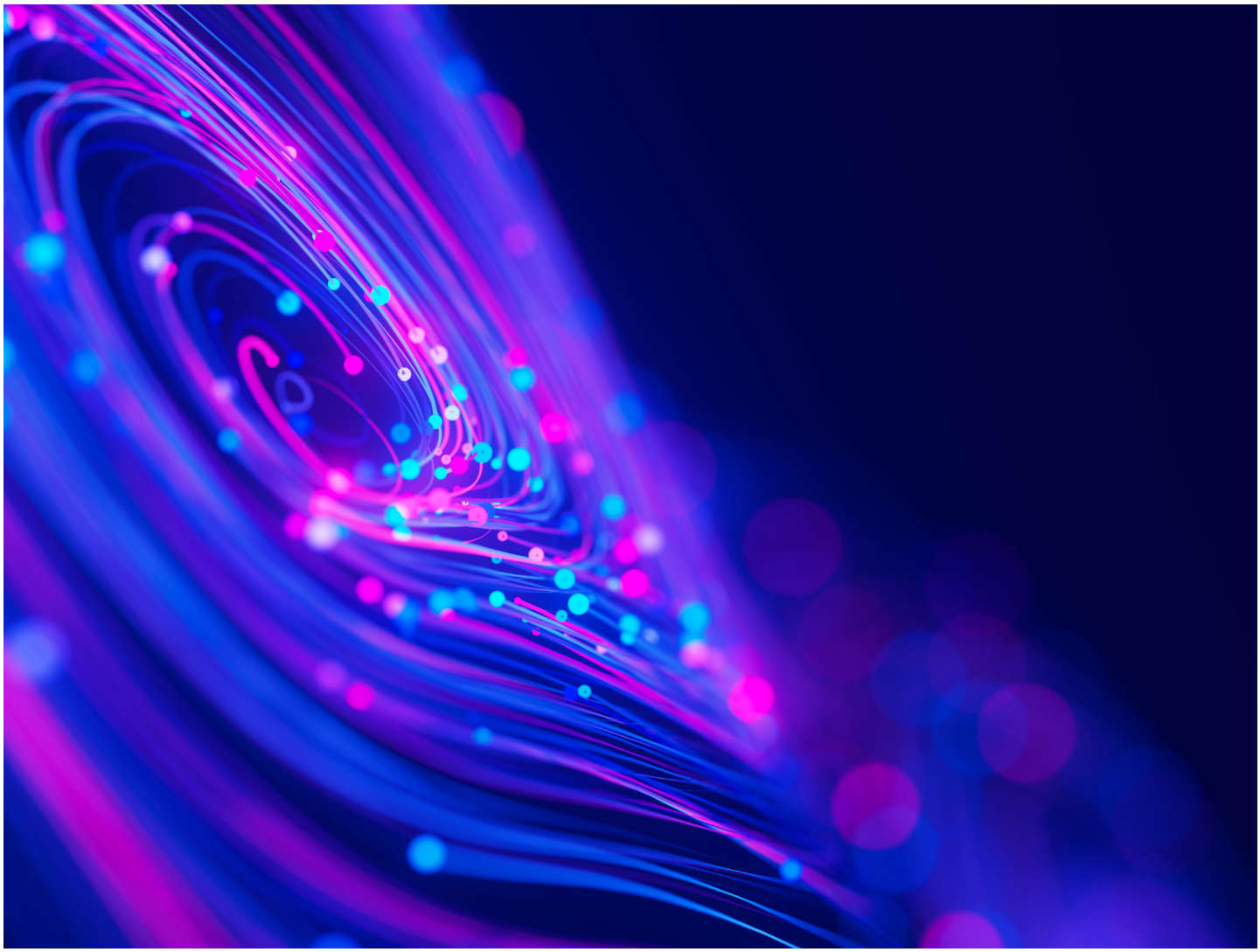
From the sponsor

Arm is the industry's highest-performing and most power-efficient compute platform with unmatched scale that touches 100% of the connected global population. To meet the insatiable demand for compute, Arm delivers advanced technology solutions that allow the world's leading companies to unleash the unprecedented experiences and capabilities of AI.

The Arm architecture provides the ideal foundation for advanced workloads, with our flexible compute platform offering the heterogeneous computing capabilities to run AI on the CPU or accelerator technologies. Our CPU designs can process a broad range of the AI inference workloads most used across billions of devices, from sensors and smartphones to the cloud and datacenter. Integrating Arm's CPU designs with our leading-edge GPUs offers acceleration capabilities for a broad set of on-device AI use cases. Moreover, the Arm Ethos-NPU runs alongside Cortex-A based systems to deliver accelerated performance for edge AI workloads. The Arm compute platform also provides our industry partners with the flexibility to create their own customized, differentiated silicon solutions for AI. Together with the world's largest computing ecosystem and 20 million software developers, we are building the future of AI on Arm. See how Arm and its ecosystem of partners are accelerating AI everywhere, from cloud to edge. Visit **arm.com/ai**.

arm

Illustrations
Cover and spot art by Adobe Stock.



MIT Technology Review Insights

www.technologyreview.com

insights@technologyreview.com