PERSPECTIVE

How to Cite: Kalendar, R. (2026). Prospects for Artificial Intelligence: From Scaling Limits to Specialization, Neural Infrastructure and Autonomous Machine Building. Contig, 1(3), 202618. DOI: 10.66273/3134-6359.2026.1.3.006

Received: 06/10/2026 | Accepted: 10/10/2026 | Published online: 10/10/2026

Prospects for Artificial Intelligence: From Scaling Limits to Specialization, Neural Infrastructure and Autonomous Machine Building

Ruslan Kalendar*

Nazarbayev University, Astana, Kazakhstan

* Corresponding author: ruslan.kalendar@nu.edu.kz | ORCID: 0000-0003-3986-2460

Abstract

Artificial intelligence, like any technology, develops along an S-shaped curve and has a limit. In this Perspective we argue that the current paradigm of scaling neural networks is running into diminishing returns on compute, exhaustion of human-generated data, energy constraints and theoretical limits of architectures. Published estimates suggest that these constraints will become binding by the late 2020s to early 2030s. We show that saturation of general-purpose capabilities will redirect the development of AI along four lines. First, towards narrowly specialized models that penetrate governance, control, resource allocation, legislation, the economy, security and even the spiritual sphere. Second, towards the gradual absorption of the intermediate layer — programming languages, operating systems and drivers, as well as websites, search engines and the personal computer as we know it — by neural systems and AI agents that become a universal means of controlling machines and accessing information. Third, towards processor architectures that are specialized for AI and designed with the help of AI. Fourth, towards mechanical engineering — the design and manufacture of robots and autonomous machines. For each thesis we present published evidence, a testable prediction and the main objections: verifiability, energy efficiency, security and accountability.

Keywords: artificial intelligence; scaling laws; technology S-curve; specialized models; Software 2.0; neuromorphic computing; chip design; embodied AI; robotics; AI governance; agentic web

1. Introduction

Artificial intelligence (AI) has become a general-purpose technology, but, like any technology, it develops along an S-shaped curve and has a limit. We argue that approaching this limit will not stop the development of AI but will change its direction: from growth in general-purpose capabilities towards specialization and towards the absorption of the underlying technological layers.

Within a decade, deep learning has progressed from image recognition to general-purpose generative models (LeCun et al., 2015; Vaswani et al., 2017; Bommasani et al., 2021). Estimates indicate that large language models affect a substantial share of tasks in the economy, that is, they show the hallmarks of a general-purpose technology (Bresnahan & Trajtenberg, 1995; Eloundou et al., 2024). Public debate, meanwhile, oscillates between two extremes: extrapolation of exponential growth and expectation of a new "AI winter".

The history of technology suggests a third scenario. Every technological paradigm first grows at an accelerating pace and then reaches a plateau bounded by physics and economics (Foster, 1986). Yet the main economic impact often arrives precisely on the plateau, when the technology reorganizes production around itself. This was the case with electrification: electric motors were improved for decades, but productivity gains came only after factories were reorganized around electric drive (David, 1990; Perez, 2002).

In this Perspective we put forward five interrelated theses:

The limit. The current paradigm — scaling neural networks on human-generated data — is running into informational, energy and theoretical constraints. Saturation of general-purpose capabilities is expected in the foreseeable future.

Specialization. On the plateau, value shifts from general-purpose models to narrowly specialized models trained for a specific domain. AI penetrates every sphere: governance, control, resource allocation, legislation, the economy, security and even spiritual life.

Disappearance of the intermediate layer. Programming languages, operating systems and drivers are a historical intermediary between human intent and the machine. In the long run this layer is absorbed by neural systems, which themselves become a universal means of specifying and executing computation. The same fate awaits websites, search engines and the familiar personal computer: the Internet becomes an entry point to a personal AI service.

Hardware co-evolution. Processor architectures specialize for AI and are designed with the help of AI. General-purpose architectures recede into the background.

Mechanical engineering. The next stage is AI entering the physical world: the design and manufacture of robots and autonomous machines.

These theses form a single line of reasoning. The limit to growth "upwards" (generality) redirects development "outwards" (specialization) and "downwards" (into the software, hardware and machine-building layers). Sections 2–6 substantiate each thesis in turn with published evidence. Section 7 addresses the main objections, and Section 8 concludes. The forecasts in this article are the authors' position; we state them as testable hypotheses and indicate which observations could refute them.

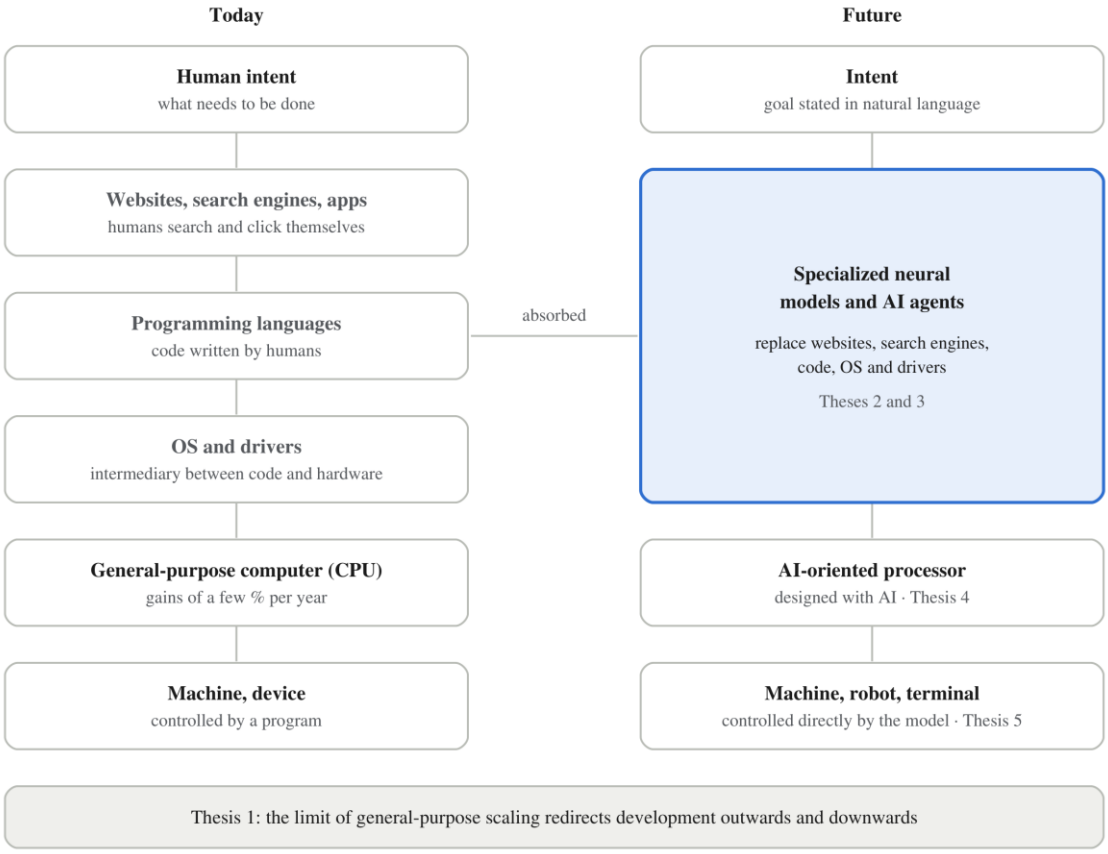


Figure 1. The technology stack: today and in prospect. The current (left) and prospective (right) technology stack. Websites and search engines, programming languages, operating systems and drivers are absorbed by specialized neural models and AI agents that run on AI-oriented processors and control the machine directly.

2. The Limits of Development: AI on the S-Curve

The current scaling paradigm is running into four constraints at once: diminishing returns on compute, data exhaustion, energy, and theoretical limits of architectures. According to published estimates, by the late 2020s to early 2030s these constraints will make further scaling of general-purpose models economically unjustified.

2.1. Diminishing returns to scaling

Scaling laws show that the loss of a language model decreases as a power law with the number of parameters, the amount of data and the compute budget (Kaplan et al., 2020). The compute exponent is small ($\alpha_C \approx 0.050$), which implies that a tenfold increase in compute reduces the loss by only about 11%. Subsequent work refined the balance: parameters and data should grow in proportion, and a 70-billion-parameter model outperformed a 280-billion-parameter model at the same compute budget (Hoffmann et al., 2022).

A power law in itself contains no ceiling, but it means exponential costs for linear progress. Training compute for leading models doubled roughly every 20 months before 2010 and every 6 months in the deep-learning era (Sevilla et al., 2022); according to the AI Index, every 5 months (Maslej et al., 2025).

For ImageNet, the required compute grew roughly as the ninth power of performance. Reducing the error from 11.5% to 5% would require about 10^5 times more compute (Thompson et al., 2020).

Equally important is what scaling laws do not promise. "Emergent" jumps in capability turn out in most cases to be an artefact of discontinuous metrics; with continuous metrics, growth is smooth and predictable (Schaeffer et al., 2023). Hence there is no basis for expecting a qualitative leap from scale alone.

2.2. Data exhaustion

The stock of publicly available human-generated text is estimated at about 300 trillion tokens. If current trends continue, it will be fully used for training between 2026 and 2032 (median 2028) (Villalobos et al., 2024). Human data cannot be replaced by synthetic data without loss. Training without filtering on data generated by previous generations of models leads to irreversible "model collapse", in which the rare events of the original distribution disappear (Shumailov et al., 2024). The informational resource of the paradigm is therefore finite.

2.3. Energy and the physics of computation

In 2024, data centres consumed about 415 TWh (roughly 1.5% of global electricity); the base case projects about 945 TWh by 2030 (IEA, 2025). The largest training runs already draw more than 100 MW, and their power demand is growing about 2.2-fold per year. By 2030 a single run may require 4–16 GW, the output of several large power plants (You et al., 2025). Training emissions have grown from 588 t CO_2 for GPT-3 to 8,930 t for Llama 3.1 405B (Maslej et al., 2025; cf. Strubell et al., 2019).

A systematic analysis by Epoch AI compared four constraints on the scale of a single training run by 2030 (Sevilla et al., 2024). Power supply is likely to bind first:

Table 1.

Constraint	Resource estimate for 2030	Feasible training run, FLOP
Power supply	1–5 GW at a single site; 2–45 GW distributed	10^{28} – $2{\cdot}10^{30}$
Chip manufacturing	20–400 million H100 equivalents	10^{29} – $5{\cdot}10^{30}$
Data	400 trillion – 20 quadrillion effective tokens	$6{\cdot}10^{28}$ – $2{\cdot}10^{32}$
Latency wall	physical limit of parallelism	$3{\cdot}10^{30}$ – 10^{32}

The authors conclude that a run of about $2{\cdot}10^{29}$ FLOP (roughly 10,000 times GPT-4) is probably feasible by 2030. But this is precisely what marks the horizon: the lower bounds of all four constraints lie within one to two orders of magnitude of this value. Moreover, hardware no longer guarantees cheaper computation: Dennard scaling ended in the mid-2000s, and Moore's law is slowing (Theis & Wong, 2017; Leiserson et al., 2020).

2.4. Theoretical constraints

The universal approximation theorem guarantees that a neural network can represent any function (Hornik et al., 1989). But it is an existence theorem: it says nothing about how efficiently such a network can be trained. For large language models it has been shown formally that hallucinations cannot be eliminated completely if the model is used as a general problem solver (Xu et al., 2024).

Transformers solve multi-step compositional tasks mainly by matching previously seen fragments rather than by systematic reasoning. GPT-4, for example, multiplied three-digit numbers correctly in only 59% of cases (Dziri et al., 2023; cf. Marcus, 2018). Finally, the "no free lunch" theorems state that, averaged over all problems, no optimization algorithm is better than any other (Wolpert & Macready, 1997). A gain is possible only by tailoring to a class of problems — and that is exactly specialization (Section 3).

2.5. Signs of a plateau

The first signs of saturation are already visible. Leading models are converging in quality: within a year, the gap between the first- and tenth-ranked models on Chatbot Arena narrowed from 11.9% to 5.4%, and the gap between open- and closed-weight models from 8.0% to 1.7% (Maslej et al., 2025). According to the International AI Safety Report, since 2025 a substantial part of capability gains has come not from larger models but from post-training and inference-time compute; diminishing economic returns may slow scaling (Bengio et al., 2026).

We interpret the shift to "reasoning" models and inference-time scaling as a classic jump to the next S-curve (Foster, 1986). A smaller model with additional inference-time compute can outperform a model 14 times larger (Snell et al., 2025). Reinforcement learning on reasoning tasks raised accuracy on the AIME 2024 benchmark from 15.6% to 77.9% (Guo et al., 2025). However, the new curve shifts costs from training to every individual query and does not remove the energy constraints of Section 2.3.

Sutton's "bitter lesson" holds that general methods that leverage growing computation have always beaten methods that rely on domain knowledge (Sutton, 2019). We suggest that this lesson holds as long as computation becomes exponentially cheaper. Once the growth of available compute and energy slows, the balance will shift towards specialization and domain knowledge.

Testable prediction. If by 2030 models trained with a budget of the order of 10^{29} FLOP show a qualitative increase in capability beyond the extrapolation of power laws, Thesis 1 should be considered refuted.

3. Specialization: From a Universal Model to an Ecosystem of Narrow Models

As the growth of general-purpose capabilities slows, competitive advantage passes to narrowly specialized models. At the same time, AI moves beyond the IT sector and becomes embedded in public administration, control, resource allocation, legislation, the economy, security and even the spiritual sphere.

3.1. Why specialization is inevitable

Already today, the most significant scientific results of AI have been obtained by specialized models. AlphaFold was the first to predict protein structures routinely with atomic accuracy (Jumper et al., 2021). GraphCast produces a 10-day global weather forecast in under a minute and outperforms the operational ECMWF HRES system on 90% of 1,380 verification targets (Lam et al., 2023). Domain-specific language models in finance and medicine outperform general-purpose models of comparable size in their fields (Wu et al., 2023; Singhal et al., 2023).

Specialization makes it possible to substitute data quality for scale. The phi-1 model, with 1.3 billion parameters and trained on about 7 billion tokens of "textbook-quality" data, solved 50.6% of HumanEval problems (Gunasekar et al., 2023). Mixture-of-experts architectures implement specialization within a single network and speed up pre-training up to sevenfold at the same compute (Shazeer et al., 2017; Fedus et al., 2022). Knowledge distillation transfers the capabilities of a large model into a compact one (Hinton et al., 2015). For agentic systems it has already been proposed to use small language models by default, reserving large models for rare general-purpose tasks (Belcak et al., 2025).

Economics reinforces this trend. Between November 2022 and October 2024, the cost of inference at GPT-3.5 level fell more than 280-fold, and leading models converged in quality (Maslej et al., 2025). The general-purpose model is becoming a commodity with shrinking margins, and value is shifting to domain data, integration and accountability for outcomes. Theoretically, this follows from the "no free lunch" theorems (Wolpert & Macready, 1997): for any fixed class of problems there will always be a specialized system that solves it more cheaply than a general-purpose one.

We expect a two-tier structure. A small number of foundation models (Bommasani et al., 2021) will become infrastructure, like the electricity grid. On top of them, a multitude of specialized models will grow — for each industry, process and device.

3.2. Penetration into every sphere

AI is already applied in all the spheres named in Thesis 2, including those traditionally considered exclusively human:

Table 2.

Sphere	Characteristic example	Source
Public administration	By 2020, 64 of 142 US federal agencies (45%) had used AI in 157 use cases	Engstrom et al., 2020; Coglianese & Lehr, 2017
Legislation	In 2023 the Porto Alegre city council unanimously passed an ordinance whose text was written by ChatGPT; GPT-4 passed the US Uniform Bar Exam	Jeantet & Savarese, 2023; Katz et al., 2024
Resource allocation	Reinforcement learning designs tax policies that improve the equality–productivity trade-off in simulated economies	Zheng et al., 2022
Control of physical systems	A neural controller commands all magnetic coils of the TCV tokamak and shapes the plasma as specified	Degrave et al., 2022
Economy and healthcare	The FDA approved 223 AI-enabled medical devices in 2023, compared with 6 in 2015	Maslej et al., 2025; Topol, 2019
Security	AI changes the threat landscape in the digital, physical and political domains — both as a means of defence and as a means of attack	Brundage et al., 2018
Spiritual sphere	The robot preacher Mindar (Kyoto), the blessing robot BlessU-2 (Wittenberg), a church service written by ChatGPT (2023)	Jackson et al., 2023; Löffler et al., 2021; Simmerlein, 2025

Growth is especially notable in legislation. In 2024, AI was mentioned 1,889 times in the legislative proceedings of 75 countries, more than nine times as often as in 2016; US federal agencies issued 59 AI-related regulations (Maslej et al., 2025). The European Commission is studying the integration of language models into its LEOS legislative drafting system (Fitsilis & Mikros, 2024). However, general-purpose models hallucinate on at least 58% of verifiable legal questions (Dahl et al., 2024). Specialized legal tools reduce this figure to 17–33%, but not to zero (Magesh et al., 2025). This is a direct argument for Thesis 2: in high-stakes domains, only a specialized and verifiable system is fit for purpose.

Regulation itself pushes towards specialization. The EU Artificial Intelligence Act classifies systems not by technology but by area of application (Regulation (EU) 2024/1689, 2024). Eight areas are classified as high-risk, including critical infrastructure, law enforcement, the administration of justice and democratic processes. A model with a clearly delimited scope of application is easier to certify than a general-purpose one. The "society-in-the-loop" principle requires that stakeholder values be embedded in algorithmic governance (Rahwan, 2018); this is achievable for a narrow task but hardly for a "model of everything".

The spiritual sphere is instructive because it reveals both the possibilities and the limits of penetration. In three studies (Japan, Singapore, the USA), robot preachers were perceived as less credible, which reduced religious commitment and donations (Jackson et al., 2023). Religious institutions respond with their own normative frameworks. By 2024 the Rome Call for AI Ethics (2020) had been signed by

representatives of the Abrahamic religions and of eleven other world religions (Pontifical Academy for Life, 2020). The Vatican note *Antiqua et nova* (2025) states that the intelligence of AI is functional and that moral responsibility remains with humans (Dicastery for the Doctrine of the Faith, 2025). In other words, even here AI enters as a specialized tool with a delimited role rather than as a universal substitute for the human being.

Testable prediction. By 2030, most AI queries in the corporate and public sectors will be served by specialized or small models rather than by the largest general-purpose ones. The opposite outcome would refute Thesis 2.

4. The Disappearance of the Intermediate Layer: From Programs to Neural Control

Programming languages, operating systems and drivers arose as a translator between human intent and the machine. When the machine itself understands intent, and the neural network itself becomes the executable "program", the translator loses its function. Websites and search engines belong to the same layer (Section 4.4). This is the most radical of our theses, and we state it as a long-term hypothesis.

4.1. Programming as a historical intermediary

The history of programming is a steady rise in the level of abstraction: machine code, assembly language, high-level languages, frameworks. Each step brought the way a task is specified closer to natural language. The logical end point of this sequence is the specification of intent without any formal language at all. This is how the concepts of "the end of programming" (Welsh, 2023) and "Software 2.0" — in which neural-network weights found by optimization replace hand-written code (Karpathy, 2017) — are formulated.

Automated code writing is advancing faster than any other application of AI. In 2021 Codex solved 28.8% of HumanEval problems on the first attempt (Chen et al., 2021), and AlphaCode reached the upper half of participants in Codeforces competitions (Li et al., 2022). On SWE-bench, built from real GitHub issues, the best model of 2023 resolved only 1.96% of tasks (Jimenez et al., 2024). By 2026, developers report scores above 90% on SWE-bench Verified (Anthropic, 2026), although the validity of this benchmark is itself now being questioned (OpenAI, 2026). In industry, Google reported that AI generated more than a quarter of new code in October 2024 and about 75% in April 2026 (Pichai, 2024, 2026). These are company self-reports without a public methodology, but the direction of travel is clear.

It is particularly significant that AI is already improving low-level algorithms, including the infrastructure of AI itself. Sorting routines discovered by AlphaDev have been incorporated into the standard LLVM libc++ library (Mankowitz et al., 2023). AlphaTensor and FunSearch discovered new matrix multiplication algorithms and combinatorial constructions (Fawzi et al., 2022; Romera-Paredes et al., 2024). AlphaEvolve found an algorithm for multiplying 4×4 complex matrices with 48 multiplications — the first improvement on Strassen's result in 56 years — and sped up one of the kernels used to train the Gemini model by 23% (Novikov et al., 2025). Thus even the development and maintenance of AI itself is gradually passing to AI.

4.2. The neural network as an execution environment

There are no theoretical obstacles to such a transition: recurrent neural networks are computationally universal and can simulate any Turing machine (Siegelmann & Sontag, 1995). The differentiable neural computer learns by itself to read and write an external memory and solves graph problems (Graves et al., 2016).

In practice, neural networks are already displacing classical components of system software. Learned indexes are up to 70% faster than cache-optimized B-trees and require an order of magnitude less

memory (Kraska et al., 2018). A small neural network inside the operating-system kernel predicts slow SSD I/O operations with 87–97% accuracy and reduces average latency by 9.6–79.6% (Hao et al., 2020). In cars and robots, a single network maps camera pixels directly to steering commands and actuator torques, without a hand-crafted processing pipeline (Bojarski et al., 2016; Levine et al., 2016).

Moreover, a neural network can replace a program entirely. GameNGen runs the game DOOM entirely inside a diffusion model at 20 frames per second, without a game engine (Valevski et al., 2025). NeuralOS generates the graphical interface of the Ubuntu operating system, including launching applications, directly from mouse and keyboard input (Rivard et al., 2026). The "language model as an operating system" concepts treat the model as the kernel, the context as memory, tools as peripherals and applications as agents programmed in natural language (Packer et al., 2023; Ge et al., 2023).

4.3. The trajectory of transition

We distinguish three phases, which will arrive at different times in different industries:

Assisted programming (today). Humans write code with the help of AI and are responsible for it. Programming languages, operating systems and drivers remain in their current form.

Code as the machine's internal representation. AI writes and maintains code itself; humans set goals and verify results. Code becomes what assembly language is today: it exists, but almost nobody writes it by hand. A precursor of this phase is robots that write their own control programs from natural-language commands (Liang et al., 2023).

Neural control. Neural systems map intent directly onto machine behaviour. The functions of application programs, operating systems and drivers are absorbed into model weights and hardware architecture (Section 5).

The transition will not be linear. A controlled experiment in 2023 showed a 55.8% speed-up in development (Peng et al., 2023). A randomized trial in 2025 on mature projects, by contrast, found that experienced developers were 19% slower (Becker et al., 2025). About 40% of generated programs in security-relevant scenarios contained vulnerabilities (Pearce et al., 2022). These findings do not refute the direction, but they show that the second phase will require new verification tools (Section 7).

Testable prediction. By 2035, the share of production code written by hand by humans at leading technology companies will fall below 10%. Mass-produced devices will appear in which control is implemented by a neural network without a conventional operating system and drivers. The absence of both developments would weaken Thesis 3.

4.4. An Internet without websites and a computer without programs

The same process of absorption will affect the Internet and the personal computer itself. Websites, search engines and personal pages are the same kind of intermediate layer between human intent and information or services as programming languages are between intent and the machine. When an AI assistant itself finds and compares information, gives an answer and carries out an action, this layer is no longer needed.

The shift is already measurable. When Google shows an AI summary, users click on result links in 8% of visits compared with 15% without one, and on links inside the summary in only 1% (Chapekis & Lieb, 2025). Human pageviews of Wikipedia in 2025 fell by about 8% compared with 2024; the Wikimedia Foundation attributes this in part to search engines answering questions directly with generative AI (Miller, 2025). Activity on Stack Overflow fell by about 25% within six months of the release of ChatGPT (del Rio-Chanona et al., 2024). As early as 2024, Gartner predicted a 25% drop in traditional search volume by 2026 (Gartner, 2024). In 2024, automated traffic exceeded human traffic for the first time in a decade, accounting for 51% of all web traffic (Imperva, 2025). AI crawlers take content while returning almost no visitors: in July 2025 there were between 195 and more than 38,000 crawler requests per referral to a site, depending on the platform (Tomé, 2025). Meanwhile, by October 2025 more than 800 million people were using ChatGPT every week (Bellan, 2025), and 15% of people under 25 were getting news from AI chatbots weekly (Newman et al., 2025).

An organization's representative website is turning into its AI agent. Open protocols already allow organizations to expose data and services directly to AI systems: the Model Context Protocol (Anthropic, 2024), Agent2Agent (Surapaneni et al., 2025) and NLWeb, which adds to any website a conversational interface accessible to AI agents (Microsoft, 2025). Researchers describe an "agentic web" whose main users are not people but AI agents interacting with one another (Yang et al., 2025; Chen et al., 2025). Search engine optimization is already giving way to optimization for generative engines (Aggarwal et al., 2024). In the logic of Thesis 2, an organization's representative website becomes its specialized model — an agent that answers questions and performs transactions on its behalf. Standard websites, personal pages and search engines are passing into history. The Internet becomes a network of agents and, for a person, an entry point to their own AI service on a computer, tablet or phone.

The computer itself is changing too. The familiar way of working with it — files, folders, windows, applications and a browser — took shape around the graphical interface and manual control of programs. Generative AI introduces a new interface paradigm in which the user specifies the desired outcome rather than a sequence of actions (Nielsen, 2023). Even the graphical interface of an operating system can already be generated by a neural network (Rivard et al., 2026). As early as 1991, Weiser wrote that the most profound technologies "disappear", weaving themselves into everyday life (Weiser, 1991). We expect the personal computer as a universal tool for manually operating programs to lose its purpose, because the familiar ways of using it will disappear. The device will become a terminal for accessing a personal AI service, and computation will move to specialized hardware (Section 5).

Testable prediction. By 2035, most information queries will be served by AI assistants without a visit to a website, and large organizations will provide their services primarily through AI agents accessible via open protocols. If the share of clicks from search to websites stays at its 2024 level, this part of Thesis 3 will be refuted.

5. The Hardware Foundation: Processors for AI, Designed by AI

General-purpose processors no longer deliver their former performance gains. Growth now comes from architectures specialized for AI, and these are increasingly designed with the participation of AI itself. This closes a loop of co-evolution: AI designs the chips on which the next generation of AI is trained.

5.1. The end of general-purpose growth

Dennard's scaling rule made it possible to shrink transistors at constant power density (Dennard et al., 1974). It ceased to hold in the mid-2000s (Theis & Wong, 2017). Since then, the performance of general-purpose processors has grown by only a few per cent per year, compared with about 50% per year in 1986–2002. This is why Hennessy and Patterson identified domain-specific architectures as the main path forward (Hennessy & Patterson, 2019).

The main constraint of modern architectures is not arithmetic but data movement. An access to external DRAM costs two to three orders of magnitude more energy than an arithmetic operation (Horowitz, 2014; Sze et al., 2017). The von Neumann architecture, with its separation of processor and memory, is poorly suited to neural networks, in which billions of weights are read at every step.

5.2. Architectural specialization

Google's first tensor processing unit (TPU) was 15–30 times faster than contemporary CPUs and GPUs on production inference workloads and 30–80 times more energy-efficient (Jouppi et al., 2017). TPU v4 connects 4,096 chips through optical circuit switches and is 1.2–1.7 times faster than the NVIDIA A100 while using 1.3–1.9 times less power (Jouppi et al., 2023). Further examples include accelerators with optimized dataflows (Chen et al., 2017), deterministic streaming processors scheduled at compile time (Abts et al., 2020) and wafer-scale processors with 2.6 trillion transistors and 850,000 cores (Lie, 2023).

The next step is to abandon the von Neumann principle itself. In-memory computing performs matrix–vector multiplication where the weights are stored (Sebastian et al., 2020). IBM's analogue AI chip with 35 million phase-change memory devices reaches 12.4 TOPS/W on speech recognition with accuracy close to that of software (Ambrogio et al., 2023; cf. Ambrogio et al., 2018). The TrueNorth neuromorphic chip, with one million neurons, recognizes objects in a video stream while consuming about 63 mW (Merolla et al., 2014; see also Davies et al., 2018). The prospects for scaling neuromorphic systems and photonic processors are being actively investigated (Kudithipudi et al., 2025; Shen et al., 2017). In such architectures the neural network is not executed by a program but is physically embodied in hardware — and this is exactly what makes the intermediate software layer of Section 4 redundant.

The market already reflects this shift. NVIDIA's data-centre revenue grew from US$115.2 billion in fiscal year 2025 (February 2024 – January 2025) to US$193.7 billion in fiscal year 2026 (NVIDIA, 2025, 2026). In the fourth quarter of 2025, servers with embedded GPUs accounted for more than half of server market revenue (IDC, 2026). At the same time, spending on general-purpose servers is still growing in absolute terms (Dell'Oro Group, 2026): general-purpose processors are receding in share rather than disappearing.

Specialization also has a downside — the "hardware lottery". Ideas often win not because they are better but because they suit the existing hardware (Hooker, 2021). Hardware tailored to transformers may slow the transition to the next S-curve. In our reasoning, this is one more mechanism that locks in the limit of the current paradigm (Section 2).

5.3. AI designs processors

A reinforcement-learning method for placing macro blocks on a chip produces a floorplan in under six hours (Mirhoseini et al., 2021). According to its developers, it has been used in every TPU generation since 2020 (Goldie et al., 2024; Google DeepMind, 2024). The result is disputed: independent evaluations found that carefully tuned simulated annealing and commercial electronic design automation tools achieve comparable or better results (Cheng et al., 2023; Markov, 2024). A commercial reinforcement-learning design optimization tool had already been used in 100 commercial chip tape-outs by February 2023 (Synopsys, 2023).

Language models are also reaching higher levels of design. Fine-tuned models write hardware descriptions in Verilog and outperform general-purpose models in doing so (Thakur et al., 2023, 2024). ChipNeMo, a model adapted for chip design, outperformed GPT-4 in two of three engineering use cases (Liu et al., 2023). Both examples also support Thesis 2: in hardware design, the specialized model wins.

Testable prediction. By 2035, mass-produced processors will appear for which an AI system proposed not only the physical layout but also the microarchitecture, with humans performing only verification. The share of analogue, neuromorphic or photonic accelerators in model inference on edge devices will become significant.

6. Mechanical Engineering, Robots and Autonomous Machines

The logical continuation of this trajectory is AI entering the physical world. Neural models control robots directly, AI designs machines, and production becomes autonomous. All the preceding theses converge here: a specialized model (Thesis 2) without intermediate code (Thesis 3) on specialized hardware (Thesis 4) controls a machine built by other machines.

6.1. Overcoming Moravec's paradox

For a long time, perception and movement remained harder for machines than abstract reasoning: a computer could play checkers but could not match a one-year-old child in motor skills (Moravec, 1988). The embodied approach held from the outset that intelligence must be built from the bottom up, from direct coupling with the world (Brooks, 1991).

Deep learning has substantially narrowed this gap. Convolutional policies map images directly to motor torques (Levine et al., 2016). Policies trained in simulation transfer to real legged robots (Hwangbo et al., 2019). The autonomous quadrotor Swift won 15 of 25 races against three world champions (Kaufmann et al., 2023). Fine bimanual manipulation tasks are learned with 80–90% success from only 50 demonstrations on hardware costing about US$20,000 (Zhao et al., 2023).

6.2. Foundation models for robots

Vision–language–action (VLA) models take an image and a natural-language command as input and output control actions. This is exactly what the third phase of Section 4.3 looks like: there is no human-written control program between intent and the machine's movement.

PaLM-E (562 billion parameters) was the first to embed continuous sensor signals into a language model for robot action planning (Driess et al., 2023).

RT-2 encodes actions as text tokens and transfers web knowledge into control; success in unseen scenarios rose from 32% to 62% (Brohan et al., 2023).

Open X-Embodiment pooled data from 22 robot embodiments across 21 institutions (527 skills, more than 1 million episodes); the shared model was on average 50% more successful than embodiment-specific ones (O'Neill et al., 2024).

OpenVLA (7 billion parameters) outperformed a closed model seven times its size by 16.5 percentage points in task success (Kim et al., 2024).

π_0 was trained on more than 10,000 hours of data from 7 robot configurations and controls them at up to 50 Hz (Black et al., 2025).

Gemini Robotics and **GR00T N1** learn new tasks from about a hundred demonstrations and are trained on synthetic data: 780,000 trajectories for a humanoid were generated in 11 hours (Gemini Robotics Team, 2025; Bjorck et al., 2025).

We emphasize the pattern: a general robotics model serves as the foundation, but a fine-tuned specialized version runs on each particular machine (Firoozi et al., 2025). In road transport, autonomy already delivers measurable benefits. Over 56.7 million rider-only miles, Waymo vehicles were involved in injury crashes significantly less often than human drivers under comparable conditions; for intersection crashes the reduction was 96% (Kusano et al., 2025).

6.3. Machines that build machines

The idea of an automaton capable of building machines goes back to von Neumann (von Neumann, 1966). The first practical demonstration came in 2000: robot bodies and their neural controllers were co-evolved by an evolutionary algorithm and automatically fabricated on a 3D printer (Lipson & Pollack, 2000). Today, deep generative models are used in topology optimization, shape synthesis and materials design (Regenwetter et al., 2022). Digital twins link design with operation (Tao et al., 2019), and the concept of "Industry 5.0" places at its centre not technology but people and sustainability (Xu et al., 2021).

Autonomous laboratories carry the same principle into science (Wang et al., 2023). A mobile robotic chemist independently performed 688 experiments in 8 days and found a photocatalyst six times more active than the starting formulation (Burger et al., 2020). The A-Lab story, however, shows the cost of haste. After criticism, the number of successfully synthesized target compounds was corrected from 41 of 58 to 36 of 57, and the compounds were no longer described as new to science (Szymanski et al., 2023; Leeman et al., 2024).

The scale of robotization is growing rapidly. In 2024, 542,000 industrial robots were installed worldwide — twice as many as ten years earlier (IFR, 2025). In 2025 installations exceeded 600,000, the global operational stock reached 5 million units, and China accounted for 59% of installations (IFR, 2026a). Series production of humanoids has also begun: about 7,000 full-size humanoid robots were sold in

2025 (IFR, 2026b). This is still the early part of the S-curve — the same phase in which language models found themselves at the turn of the 2010s and 2020s.

Testable prediction. By 2035, most newly installed robots will be configured through demonstration and natural-language commands via foundation models rather than programmed in industrial robot languages.

7. Objections and Open Questions

The most serious objections are directed against Thesis 3. A neural network is hard to verify, it is less energy-efficient at exact computation, and it bears no legal responsibility. Below we consider seven objections and indicate how they modify our forecasts.

7.1. Verifiability. For classical software there are methods for proving correctness (Hoare, 1969). The correctness of the seL4 microkernel, comprising 8,700 lines of C, has been proved by machine (Klein et al., 2009). No comparable guarantees exist for neural networks, and complete elimination of hallucinations is theoretically impossible (Xu et al., 2024). Formal layers will therefore persist longest in safety-critical systems. We expect that human code writing will disappear, while formal specifications will remain as machine-generated and machine-checked artefacts. Verification of neural systems is the main precondition for realizing Thesis 3.

7.2. Energy efficiency of exact computation. Adding two numbers in a processor register costs a fraction of a picojoule (Horowitz, 2014), whereas a language model performs about 2N operations per output token, where N is the number of parameters (Kaplan et al., 2020), and still makes mistakes in multi-digit multiplication (Dziri et al., 2023). It follows that what will disappear is not exact computation but the intermediate program between it and intent. Deterministic units will become part of specialized hardware (Section 5), which the neural system invokes directly.

7.3. Technical debt. In real ML systems, the model code itself is a small fraction; most of the complexity and hidden debt accumulates in the surrounding infrastructure (Sculley et al., 2015). Replacing code with weights moves complexity into data but does not eliminate it. This is an argument for self-maintaining AI systems, but also the reason why the transition will take decades rather than years.

7.4. Security and dual use. About 40% of AI-generated programs in security-relevant scenarios contained vulnerabilities (Pearce et al., 2022). Autonomous machines and AI in infrastructure management enlarge the attack surface (Brundage et al., 2018; Bengio et al., 2026). The deeper AI penetrates the physical world (Thesis 5), the higher the cost of error and the stricter the certification requirements will become.

7.5. Accountability and legitimacy. In governance, law and the spiritual sphere, a decision must be not only good but also legitimate. Regulators and religious institutions agree that responsibility remains with humans (Regulation (EU) 2024/1689, 2024; Dicastery for the Doctrine of the Faith, 2025). Evidence of lower trust in robot preachers (Jackson et al., 2023) shows that social acceptance may lag behind technical capability. In these spheres we therefore expect AI to act as a specialized adviser and executor rather than as the subject of decisions.

7.6. A new paradigm may push the limit back. The history of AI teaches that general methods that leverage computation win sooner or later (Sutton, 2019). Epoch AI's analysis allows for training runs roughly 10,000 times larger than GPT-4 by 2030 (Sevilla et al., 2024). A fundamentally new architecture could start a new S-curve (Foster, 1986), while hardware specialization for the current paradigm could slow it down (Hooker, 2021). Thesis 1 therefore refers to the current scaling paradigm, not to AI as such. Theses 2–5 remain valid even if scaling continues: they describe the direction in which the technology diffuses, not its ceiling.

7.7. Who will create knowledge? If websites and forums disappear, people will have less incentive to publish knowledge openly; the decline of activity on Stack Overflow is a first signal (del Rio-Chanona et al., 2024). Yet it is human data that feeds AI, and training on a model's own outputs leads to model

collapse (Shumailov et al., 2024; Villalobos et al., 2024). The agentic Internet will therefore require new mechanisms for rewarding authors; first attempts, such as charging AI crawlers for access, have already appeared (Allen & Newton, 2025). Another risk is the concentration of access to information in the hands of a few AI service operators. This is an argument for open protocols and for many specialized agents rather than a single universal one.

8. Conclusion

AI is a technology with a limit, and the current scaling paradigm will in all likelihood reach it by the late 2020s to early 2030s. Yet a limit to growth "upwards" means not a halt but a redirection of development. Outwards — through specialization and penetration into every sphere of life. Downwards — through absorption of the intermediate layer of programs, websites and search engines, specialization of hardware and entry into mechanical engineering.

The resulting picture is a short chain, "intent → specialized neural model → AI-oriented hardware → machine", in place of today's multi-layer stack of websites and search engines, programming languages, operating systems, drivers and general-purpose computers. So far this chain has been realized only in fragments: in robots with VLA models, in neural game engines, in AI assistants that replace search and in processors designed with the participation of AI. But every one of its links has already been demonstrated.

This Perspective points to six priorities for research and practice:

Verification of neural systems — methods for proving and certifying the behaviour of models that replace code.

Energy-efficient specialized hardware — in-memory, analogue, neuromorphic and photonic architectures.

Data-efficient training of domain models — quality and structure of data rather than volume.

Frameworks for governance and accountability for AI in public administration, law, security and the spiritual sphere.

The knowledge economy of the agentic Internet — reward mechanisms for authors and open protocols, without which AI will run out of fresh human data.

Reform of engineering education — a shift of emphasis from programming to problem specification, verification, hardware and mechanical engineering.

Specialization also has an important practical consequence. If value shifts from the largest general-purpose models to domain data and integration, then countries, universities and companies without access to gigawatt-scale data centres will be able to compete in their own fields. The limit of general-purpose AI thus becomes the beginning of its democratization.

Author Contributions

R.K.: conceptualization, writing — original draft, writing — review and editing.

Use of Generative AI

Generative AI tools were used to assist with drafting, translation and verification of bibliographic references. The author reviewed and edited all content and takes full responsibility for it.

Data Availability Statement

No new data were created or analysed in this work. All sources are listed in the References.

Funding

This work received no specific funding.

Ethics approval and consent

Not applicable.

Competing Interests

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Abts, D., et al. (2020). Think fast: A tensor streaming processor (TSP) for accelerating deep learning workloads. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)* (pp. 145–158). doi:10.1109/ISCA45697.2020.00023
2. Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5–16). doi:10.1145/3637528.3671900
3. Allen, W., & Newton, S. (2025, July 1). *Introducing pay per crawl: Enabling content owners to charge AI crawlers for access*. Cloudflare Blog. blog.cloudflare.com
4. Ambrogio, S., et al. (2018). Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature, 558*(7708), 60–67. doi:10.1038/s41586-018-0180-5
5. Ambrogio, S., et al. (2023). An analog-AI chip for energy-efficient speech recognition and transcription. *Nature, 620*(7975), 768–775. doi:10.1038/s41586-023-06337-5
6. Anthropic. (2024, November 25). *Introducing the Model Context Protocol*. anthropic.com
7. Anthropic. (2026, April 7). *Project Glasswing*. https://www.anthropic.com/glasswing
8. Becker, J., Rush, N., Barnes, E., & Rein, D. (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity. *arXiv:2507.09089*. doi:10.48550/arXiv.2507.09089
9. Belcak, P., et al. (2025). Small language models are the future of agentic AI. *arXiv:2506.02153*. doi:10.48550/arXiv.2506.02153
10. Bellan, R. (2025, October 6). Sam Altman says ChatGPT has hit 800M weekly active users. *TechCrunch*. techcrunch.com
11. Bengio, Y., et al. (2026). *International AI Safety Report 2026* (DSIT 2026/001). *arXiv:2602.21012*. doi:10.48550/arXiv.2602.21012
12. Bjorck, J., et al. (2025). GR00T N1: An open foundation model for generalist humanoid robots. *arXiv:2503.14734*. doi:10.48550/arXiv.2503.14734
13. Black, K., et al. (2025). π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems XXI*. doi:10.15607/RSS.2025.XXI.010

14. Bojarski, M., et al. (2016). End to end learning for self-driving cars. *arXiv:1604.07316*. doi:10.48550/arXiv.1604.07316
15. Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. *arXiv:2108.07258*. doi:10.48550/arXiv.2108.07258
16. Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'Engines of growth'? *Journal of Econometrics, 65*(1), 83–108. doi:10.1016/0304-4076(94)01598-T
17. Brohan, A., et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning* (PMLR 229, pp. 2165–2183). *arXiv:2307.15818*. doi:10.48550/arXiv.2307.15818
18. Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1–3), 139–159. doi:10.1016/0004-3702(91)90053-M
19. Brundage, M., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv:1802.07228*. doi:10.48550/arXiv.1802.07228
20. Burger, B., et al. (2020). A mobile robotic chemist. *Nature, 583*(7815), 237–241. doi:10.1038/s41586-020-2442-2
21. Chapekis, A., & Lieb, A. (2025, July 22). *Google users are less likely to click on links when an AI summary appears in the results*. Pew Research Center. pewresearch.org
22. Chen, M., et al. (2021). Evaluating large language models trained on code. *arXiv:2107.03374*. doi:10.48550/arXiv.2107.03374
23. Chen, W., et al. (2025). Internet of Agents: Weaving a web of heterogeneous agents for collaborative intelligence. In *International Conference on Learning Representations (ICLR 2025)*. *arXiv:2407.07061*. doi:10.48550/arXiv.2407.07061
24. Chen, Y.-H., Krishna, T., Emer, J. S., & Sze, V. (2017). Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE Journal of Solid-State Circuits, 52*(1), 127–138. doi:10.1109/JSSC.2016.2616357
25. Cheng, C.-K., Kahng, A. B., Kundu, S., Wang, Y., & Wang, Z. (2023). Assessment of reinforcement learning for macro placement. In *Proceedings of the 2023 International Symposium on Physical Design* (pp. 158–166). doi:10.1145/3569052.3578926
26. Coglianese, C., & Lehr, D. (2017). Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal, 105*, 1147–1223.
27. Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis, 16*(1), 64–93. doi:10.1093/jla/laae003
28. David, P. A. (1990). The dynamo and the computer: An historical perspective on the modern productivity paradox. *American Economic Review, 80*(2), 355–361.
29. Davies, M., et al. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro, 38*(1), 82–99. doi:10.1109/MM.2018.112130359
30. Degrave, J., et al. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature, 602*(7897), 414–419. doi:10.1038/s41586-021-04301-9
31. del Rio-Chanona, R. M., Laurentsyeva, N., & Wachs, J. (2024). Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus, 3*(9), pgae400. doi:10.1093/pnasnexus/pgae400
32. Dell'Oro Group. (2026, March 17). *Data center capex surges 57 percent in 2025 as AI deployments accelerate* [Press release]. delloro.com
33. Dennard, R. H., et al. (1974). Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE Journal of Solid-State Circuits, 9*(5), 256–268. doi:10.1109/JSSC.1974.1050511
34. Dicastery for the Doctrine of the Faith & Dicastery for Culture and Education. (2025). *Antiqua et nova: Note on the relationship between artificial intelligence and human intelligence*. Vatican City, 28 January 2025. vatican.va
35. Driess, D., et al. (2023). PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning* (PMLR 202, pp. 8469–8488).

36. Dziri, N., et al. (2023). Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems, 36*, 70293–70332. doi:10.52202/075280-3081
37. Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science, 384*(6702), 1306–1308. doi:10.1126/science.adj0998
38. Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). *Government by algorithm: Artificial intelligence in federal administrative agencies*. Report to the Administrative Conference of the United States. acus.gov doi:10.2139/ssrn.3551505
39. Fawzi, A., et al. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature, 610*(7930), 47–53. doi:10.1038/s41586-022-05172-4
40. Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research, 23*(120), 1–39.
41. Firoozi, R., et al. (2025). Foundation models in robotics: Applications, challenges, and the future. *International Journal of Robotics Research, 44*(5), 701–739. doi:10.1177/02783649241281508
42. Fitsilis, F., & Mikros, G. (2024). *AI-based solutions for legislative drafting in the EU*. European Commission, Interoperable Europe.
43. Foster, R. N. (1986). *Innovation: The attacker's advantage*. Summit Books.
44. Gartner. (2024, February 19). *Gartner predicts search engine volume will drop 25% by 2026, due to AI chatbots and other virtual agents* [Press release]. gartner.com
45. Ge, Y., et al. (2023). LLM as OS, agents as apps: Envisioning AIOS, agents and the AIOS-agent ecosystem. *arXiv:2312.03815*. doi:10.48550/arXiv.2312.03815
46. Gemini Robotics Team. (2025). Gemini Robotics: Bringing AI into the physical world. *arXiv:2503.20020*. doi:10.48550/arXiv.2503.20020
47. Goldie, A., et al. (2024). Addendum: A graph placement methodology for fast chip design. *Nature, 634*(8034), E10–E11. doi:10.1038/s41586-024-08032-5
48. Google DeepMind. (2024, September 26). *How AlphaChip transformed computer chip design*. deepmind.google
49. Graves, A., et al. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature, 538*(7626), 471–476. doi:10.1038/nature20101
50. Gunasekar, S., et al. (2023). Textbooks are all you need. *arXiv:2306.11644*. doi:10.48550/arXiv.2306.11644
51. Guo, D., et al. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature, 645*(8081), 633–638. doi:10.1038/s41586-025-09422-z
52. Hao, M., et al. (2020). LinnOS: Predictability on unpredictable flash storage with a light neural network. In *14th USENIX Symposium on Operating Systems Design and Implementation* (pp. 173–190).
53. Hennessy, J. L., & Patterson, D. A. (2019). A new golden age for computer architecture. *Communications of the ACM, 62*(2), 48–60. doi:10.1145/3282307
54. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv:1503.02531*. doi:10.48550/arXiv.1503.02531
55. Hoare, C. A. R. (1969). An axiomatic basis for computer programming. *Communications of the ACM, 12*(10), 576–580. doi:10.1145/363235.363259
56. Hoffmann, J., et al. (2022). An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems, 35*, 30016–30030. (Preprint: Training compute-optimal large language models, *arXiv:2203.15556*.) doi:10.48550/arXiv.2203.15556
57. Hooker, S. (2021). The hardware lottery. *Communications of the ACM, 64*(12), 58–65. doi:10.1145/3467017
58. Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366. doi:10.1016/0893-6080(89)90020-8
59. Horowitz, M. (2014). Computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers* (pp. 10–14). doi:10.1109/ISSCC.2014.6757323

60. Hwangbo, J., et al. (2019). Learning agile and dynamic motor skills for legged robots. *Science Robotics, 4*(26), eaau5872. doi:10.1126/scirobotics.aau5872
61. IDC. (2026, March 12). *Worldwide quarterly server tracker, Q4 2025* [Press release]. idc.com
62. IEA. (2025). *Energy and AI*. International Energy Agency. iea.org
63. IFR. (2025, September 25). *Global robot demand in factories doubles over 10 years* (World Robotics 2025) [Press release]. ifr.org
64. IFR. (2026a, September 24). *Five million robots now operate in factories globally* (World Robotics 2026) [Press release]. ifr.org
65. IFR. (2026b, September 30). *Global sales of professional service robots surge 24%* [Press release]. ifr.org
66. Imperva. (2025, April 15). *2025 Imperva Bad Bot Report: AI fuels rise of hard-to-detect bots that now make up more than half of global internet traffic* [Press release]. Thales. cpl.thalesgroup.com
67. Jackson, J. C., Yam, K. C., Tang, P. M., Liu, T., & Shariff, A. (2023). Exposure to robot preachers undermines religious commitment. *Journal of Experimental Psychology: General, 152*(12), 3344–3358. doi:10.1037/xge0001443
68. Jeantet, D., & Savarese, M. (2023, November 30). Brazilian city enacts an ordinance that was secretly written by ChatGPT. *Associated Press*.
69. Jimenez, C. E., et al. (2024). SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR 2024)*. *arXiv:2310.06770*. doi:10.48550/arXiv.2310.06770
70. Jouppi, N. P., et al. (2017). In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture* (pp. 1–12). doi:10.1145/3079856.3080246
71. Jouppi, N. P., et al. (2023). TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In *Proceedings of the 50th Annual International Symposium on Computer Architecture* (pp. 1–14). doi:10.1145/3579371.3589350
72. Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. doi:10.1038/s41586-021-03819-2
73. Kaplan, J., et al. (2020). Scaling laws for neural language models. *arXiv:2001.08361*. doi:10.48550/arXiv.2001.08361
74. Karpathy, A. (2017, November 11). *Software 2.0*. Medium. karpathy.medium.com
75. Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A, 382*(2270), 20230254. doi:10.1098/rsta.2023.0254
76. Kaufmann, E., et al. (2023). Champion-level drone racing using deep reinforcement learning. *Nature, 620*(7976), 982–987. doi:10.1038/s41586-023-06419-4
77. Kim, M. J., et al. (2024). OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning* (PMLR 270, pp. 2679–2713). *arXiv:2406.09246*. doi:10.48550/arXiv.2406.09246
78. Klein, G., et al. (2009). seL4: Formal verification of an OS kernel. In *Proceedings of the 22nd ACM Symposium on Operating Systems Principles* (pp. 207–220). doi:10.1145/1629575.1629596
79. Kraska, T., et al. (2018). The case for learned index structures. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD)* (pp. 489–504). doi:10.1145/3183713.3196909
80. Kudithipudi, D., et al. (2025). Neuromorphic computing at scale. *Nature, 637*(8047), 801–812. doi:10.1038/s41586-024-08253-8
81. Kusano, K. D., et al. (2025). Comparison of Waymo rider-only crash rates by crash type to human benchmarks at 56.7 million miles. *Traffic Injury Prevention, 26*(sup1), S8–S20. doi:10.1080/15389588.2025.2499887
82. Lam, R., et al. (2023). Learning skillful medium-range global weather forecasting. *Science, 382*(6677), 1416–1421. doi:10.1126/science.adi2336

83. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. doi:10.1038/nature14539
84. Leeman, J., et al. (2024). Challenges in high-throughput inorganic materials prediction and autonomous synthesis. *PRX Energy, 3*(1), 011002. doi:10.1103/PRXEnergy.3.011002
85. Leiserson, C. E., et al. (2020). There's plenty of room at the Top: What will drive computer performance after Moore's law? *Science, 368*(6495), eaam9744. doi:10.1126/science.aam9744
86. Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research, 17*(39), 1–40.
87. Li, Y., et al. (2022). Competition-level code generation with AlphaCode. *Science, 378*(6624), 1092–1097. doi:10.1126/science.abq1158
88. Liang, J., et al. (2023). Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 9493–9500). doi:10.1109/ICRA48891.2023.10160591
89. Lie, S. (2023). Cerebras architecture deep dive: First look inside the hardware/software co-design for deep learning. *IEEE Micro, 43*(3), 18–30. doi:10.1109/MM.2023.3256384
90. Lipson, H., & Pollack, J. B. (2000). Automatic design and manufacture of robotic lifeforms. *Nature, 406*(6799), 974–978. doi:10.1038/35023115
91. Liu, M., et al. (2023). ChipNeMo: Domain-adapted LLMs for chip design. *arXiv:2311.00176*. doi:10.48550/arXiv.2311.00176
92. Löffler, D., Hurtienne, J., & Nord, I. (2021). Blessing robot BlessU2: A discursive design study to understand the implications of social robots in religious contexts. *International Journal of Social Robotics, 13*(4), 569–586. doi:10.1007/s12369-019-00558-3
93. Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies, 22*(2), 216–242. doi:10.1111/jels.12413
94. Mankowitz, D. J., et al. (2023). Faster sorting algorithms discovered using deep reinforcement learning. *Nature, 618*(7964), 257–263. doi:10.1038/s41586-023-06004-9
95. Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv:1801.00631*. doi:10.48550/arXiv.1801.00631
96. Markov, I. L. (2024). Reevaluating Google's reinforcement learning for IC macro placement. *Communications of the ACM, 67*(11), 60–71. doi:10.1145/3676845
97. Maslej, N., et al. (2025). *The AI Index 2025 annual report*. Stanford Institute for Human-Centered AI. *arXiv:2504.07139*. hai.stanford.edu doi:10.48550/arXiv.2504.07139
98. Merolla, P. A., et al. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science, 345*(6197), 668–673. doi:10.1126/science.1254642
99. Microsoft. (2025, May 19). *Introducing NLWeb: Bringing conversational interfaces directly to the web*. Microsoft Source. news.microsoft.com
100. Miller, M. (2025, October 17). *New user trends on Wikipedia*. Wikimedia Foundation, Diff. diff.wikimedia.org
101. Mirhoseini, A., et al. (2021). A graph placement methodology for fast chip design. *Nature, 594*(7862), 207–212. doi:10.1038/s41586-021-03544-w
102. Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.
103. Newman, N., et al. (2025). *Reuters Institute Digital News Report 2025*. Reuters Institute for the Study of Journalism, University of Oxford. reutersinstitute.politics.ox.ac.uk
104. Nielsen, J. (2023, June 18). *AI: First new UI paradigm in 60 years*. Nielsen Norman Group. nngroup.com
105. Novikov, A., et al. (2025). AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv:2506.13131*. doi:10.48550/arXiv.2506.13131
106. NVIDIA. (2025, February 26). *NVIDIA announces financial results for fourth quarter and fiscal 2025* [Press release]. sec.gov

107. NVIDIA. (2026, February 25). *NVIDIA announces financial results for fourth quarter and fiscal 2026* [Press release]. sec.gov
108. O'Neill, A., et al. (Open X-Embodiment Collaboration). (2024). Open X-Embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6892–6903). doi:10.1109/ICRA57147.2024.10611477
109. OpenAI. (2026, February 23). *Why we no longer evaluate SWE-bench Verified*. openai.com
110. Packer, C., et al. (2023). MemGPT: Towards LLMs as operating systems. *arXiv:2310.08560*. doi:10.48550/arXiv.2310.08560
111. Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022). Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy* (pp. 754–768). doi:10.1109/SP46214.2022.9833571
112. Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv:2302.06590*. doi:10.48550/arXiv.2302.06590
113. Perez, C. (2002). *Technological revolutions and financial capital: The dynamics of bubbles and golden ages*. Edward Elgar. doi:10.4337/9781781005323
114. Pichai, S. (2024, October 29). *Q3 earnings call: CEO's remarks*. Google. blog.google
115. Pichai, S. (2026, April 22). *Cloud Next 2026* [Keynote remarks]. Google. blog.google
116. Pontifical Academy for Life, et al. (2020). *Rome Call for AI Ethics*. Rome, 28 February 2020. romecall.org
117. Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5–14. doi:10.1007/s10676-017-9430-8
118. Regenwetter, L., Nobari, A. H., & Ahmed, F. (2022). Deep generative models in engineering design: A review. *Journal of Mechanical Design, 144*(7), 071704. doi:10.1115/1.4053859
119. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). *Official Journal of the European Union*, L, 2024/1689. eur-lex
120. Rivard, L., et al. (2026). NeuralOS: Towards simulating operating systems via neural generative models. In *International Conference on Learning Representations (ICLR 2026)*. *arXiv:2507.08800*. doi:10.48550/arXiv.2507.08800
121. Romera-Paredes, B., et al. (2024). Mathematical discoveries from program search with large language models. *Nature, 625*(7995), 468–475. doi:10.1038/s41586-023-06924-6
122. Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems, 36*, 55565–55581. doi:10.52202/075280-2425
123. Sculley, D., et al. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems, 28*, 2503–2511.
124. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R., & Eleftheriou, E. (2020). Memory devices and applications for in-memory computing. *Nature Nanotechnology, 15*(7), 529–544. doi:10.1038/s41565-020-0655-z
125. Sevilla, J., et al. (2022). Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). doi:10.1109/IJCNN55064.2022.9891914
126. Sevilla, J., et al. (2024, August 20). *Can AI scaling continue through 2030?* Epoch AI. epoch.ai
127. Shazeer, N., et al. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR 2017)*. *arXiv:1701.06538*. doi:10.48550/arXiv.1701.06538
128. Shen, Y., et al. (2017). Deep learning with coherent nanophotonic circuits. *Nature Photonics, 11*(7), 441–446. doi:10.1038/nphoton.2017.93
129. Shumailov, I., et al. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*(8022), 755–759. doi:10.1038/s41586-024-07566-y
130. Siegelmann, H. T., & Sontag, E. D. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences, 50*(1), 132–150. doi:10.1006/jcss.1995.1013

131. Simmerlein, J. (2025). Sacred meets synthetic: A multi-method study on the first AI church service. *Review of Religious Research, 67*(1), 126–145. doi:10.1177/0034673X241282962
132. Singhal, K., et al. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. doi:10.1038/s41586-023-06291-2
133. Snell, C., Lee, J., Xu, K., & Kumar, A. (2025). Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR 2025)*. *arXiv:2408.03314*. doi:10.48550/arXiv.2408.03314
134. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). doi:10.18653/v1/P19-1355
135. Surapaneni, R., Jha, M., Vakoc, M., & Segal, T. (2025, April 9). *Announcing the Agent2Agent Protocol (A2A)*. Google Developers Blog. developers.googleblog.com
136. Sutton, R. S. (2019, March 13). *The bitter lesson*. incompleteideas.net
137. Synopsys. (2023, February 7). *AI-designed chips reach scale with first 100 commercial tape-outs using Synopsys technology* [Press release]. news.synopsys.com
138. Szymanski, N. J., et al. (2023). An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature, 624*(7990), 86–91. doi:10.1038/s41586-023-06734-w (Author Correction: *Nature, 650*, E1, 2026, doi:10.1038/s41586-025-09992-y)
139. Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE, 105*(12), 2295–2329. doi:10.1109/JPROC.2017.2761740
140. Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2019). Digital twin in industry: State-of-the-art. *IEEE Transactions on Industrial Informatics, 15*(4), 2405–2415. doi:10.1109/TII.2018.2873186
141. Thakur, S., et al. (2023). Benchmarking large language models for automated Verilog RTL code generation. In *2023 Design, Automation & Test in Europe Conference (DATE)* (pp. 1–6). doi:10.23919/DATE56975.2023.10137086
142. Thakur, S., et al. (2024). VeriGen: A large language model for Verilog code generation. *ACM Transactions on Design Automation of Electronic Systems, 29*(3), 1–31. doi:10.1145/3643681
143. Theis, T. N., & Wong, H.-S. P. (2017). The end of Moore's law: A new beginning for information technology. *Computing in Science & Engineering, 19*(2), 41–50. doi:10.1109/MCSE.2017.29
144. Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. *arXiv:2007.05558*. doi:10.48550/arXiv.2007.05558
145. Tomé, J. (2025, August 29). *The crawl-to-click gap: Cloudflare data on AI bots, training, and referrals*. Cloudflare Blog. blog.cloudflare.com
146. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. doi:10.1038/s41591-018-0300-7
147. Valevski, D., Leviathan, Y., Arar, M., & Fruchter, S. (2025). Diffusion models are real-time game engines. In *International Conference on Learning Representations (ICLR 2025)*. *arXiv:2408.14837*. doi:10.48550/arXiv.2408.14837
148. Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.
149. Villalobos, P., et al. (2024). Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning* (PMLR 235, pp. 49523–49544).
150. von Neumann, J. (1966). *Theory of self-reproducing automata* (A. W. Burks, Ed.). University of Illinois Press.
151. Wang, H., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature, 620*(7972), 47–60. doi:10.1038/s41586-023-06221-2
152. Weiser, M. (1991). The computer for the 21st century. *Scientific American, 265*(3), 94–104. doi:10.1038/scientificamerican0991-94

153. Welsh, M. (2023). The end of programming. *Communications of the ACM, 66*(1), 34–35. doi:10.1145/3570220
154. Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82. doi:10.1109/4235.585893
155. Wu, S., et al. (2023). BloombergGPT: A large language model for finance. *arXiv:2303.17564*. doi:10.48550/arXiv.2303.17564
156. Xu, X., Lu, Y., Vogel-Heuser, B., & Wang, L. (2021). Industry 4.0 and Industry 5.0—Inception, conception and perception. *Journal of Manufacturing Systems, 61*, 530–535. doi:10.1016/j.jmsy.2021.10.006
157. Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv:2401.11817*. doi:10.48550/arXiv.2401.11817
158. Yang, Y., et al. (2025). Agentic Web: Weaving the next web with AI agents. *arXiv:2507.21206*. doi:10.48550/arXiv.2507.21206
159. You, J., Owen, D., Porter, D., & Wilson, T. (2025). *Scaling intelligence: The exponential growth of AI's power needs*. EPRI & Epoch AI.
160. Zhao, T. Z., Kumar, V., Levine, S., & Finn, C. (2023). Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems XIX*. doi:10.15607/RSS.2023.XIX.016
161. Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., & Socher, R. (2022). The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances, 8*(18), eabk2607. doi:10.1126/sciadv.abk2607

Author Information

Full Name	Ruslan Kalendar
Department	
Institution	Nazarbayev University, Astana, Kazakhstan
Email	ruslan.kalendar@nu.edu.kz
ORCID	0000-0003-3986-2460
Scopus ID	