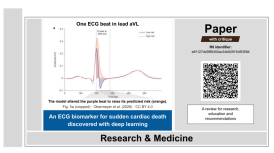
The Bible of AI™

Educational Research on Artificial Intelligence Journal — ISSN 2695-6411

Issue 2026 · Q4 · Updated 18 Nov 2026, 17:13 GMT+1

Research Proposals

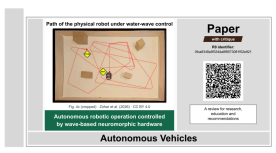


An ECG biomarker for sudden cardiac death discovered with deep learning

17 de noviembre de 2026

A routine electrocardiogram can reveal a disturbance in rhythm while it is happening. Could the shape of an apparently ordinary beat also help identify someone at risk of sudden cardiac death a year later? Ziad Obermeyer and colleagues linked Swedish ECGs to death certificates, then examined what their model had learned.

Read the review →

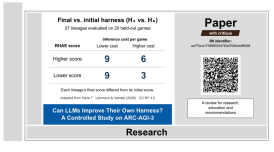


Autonomous robotic operation controlled by wave-based neuromorphic hardware

17 de noviembre de 2026

A small robotic car approaches an obstacle. Its distance readings travel by Bluetooth to a dish of water, where six actuators set waves in motion. A camera captures the interference pattern, a classifier reads it and the car receives an instruction to turn.

Read the review →



Can LLMs Improve Their Own Harness? A Controlled Study on ARC-AGI-3

14 de noviembre de 2026

If a language model can rewrite the software it uses to play a game, how can it tell whether an edit was worth keeping? Jens Lehmann and Sahar Vahdati let each model build a program, play five practice games and revise that program twice. They then compared all three versions on twenty games whose results had been withheld during development.

Read the review →

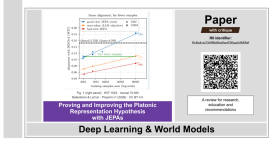


FrontierPhysics: Evaluating Agents for End-to-End Physics Research

14 de noviembre de 2026

Can a research agent tell when its own calculation has led it astray? In FrontierPhysics, physicists turned investigations from their own laboratories and research groups into 56 problems. An agent receives a brief, explores the literature, chooses a method and submits a manuscript alongside results that an examiner can check. The strongest configuration passed 46 of 168 trials.

Read the review →

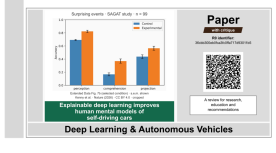


Proving and Improving the Platonic Representation Hypothesis with JEPAs

11 de noviembre de 2026

Can a compact text encoder learn representations that agree with vision more closely than those of a much larger language model? This preprint investigates that question through a conditional mathematical theorem and experiments comparing Joint-Embedding Predictive Architectures (JEPAs) with next-token prediction. Its central proposal is to train on different descriptions of the same scene, encouraging models to preserve shared information.

Read the review →

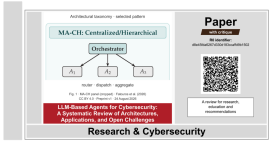


Explainable deep learning improves human mental models of self-driving cars

11 de noviembre de 2026

How can a passenger understand why a self-driving car behaves unexpectedly? This study introduces the Concept-Wrapper Network (CW-Net), which connects a learned planner to human-readable concepts and examines whether explanations improve understanding of vehicle behaviour. Its evaluation combines private-track experiments with a real autonomous vehicle, expert and non-expert studies, and an online study using public-road recordings collected while a human drove.

Read the review →

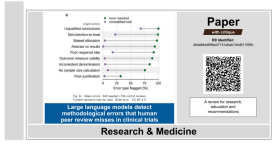


LLM-Based Agents for Cybersecurity: A Systematic Review of Architectures, Applications, and Open Challenges

10 de noviembre de 2026

This systematic review examines how language-model agents are being used in cybersecurity, from penetration testing and threat intelligence to incident response. It classifies 59 studies published between January 2022 and April 2026 by security function, agent architecture, knowledge augmentation, human oversight and evaluation rigor. The taxonomy helps readers compare single-agent tools with coordinated multi-agent systems and identify where evidence remains limited.

Read the review →



Large language models detect methodological errors that human peer review misses in clinical trials

10 de noviembre de 2026

Can language models help reviewers find methodological weaknesses in clinical-trial reports? This preprint tests fifteen models on documents containing deliberately inserted errors, first using reports from an earlier human-review experiment and then extending the task to 58 trials published in 2026. It examines detection scores, concerns raised about control reports and the time required to produce reviews.

Read the review →

The Bible of AI

Educational Research on Artificial Intelligence Journal — ISSN 2695-6411

Issue 2026 · Q4 · Updated 18 Nov 2026, 17:13 CET

Website: editorialia.com | **Contact form: editorialia.com/email/**

Publication and edition
© 2019–2026 | La Biblia de la IA – The Bible of AI™. editorialia.com. This quarterly PDF records a dated selection of the journal. Web publications may subsequently be updated. The page numbers below refer to this PDF, not to the digital page manager.

Licence and attribution
Except where specific licensing is stated, the journal uses Creative Commons Attribution–ShareAlike 4.0. Individual articles, images and third-party material retain their own stated licences and credits. The Director's editorial and several illustrations carry specific licences; the general licence does not replace them. See Terms and conditions of use.

Free edition and permitted use
This issue is provided free of charge for personal reading. Its purpose is to broaden public access to knowledge of the uses and benefits of artificial intelligence, while recognising the responsibilities that accompany them. To the extent that Editorialia holds rights in this PDF edition's original compilation and design that have not already been granted under an open licence, permission is limited to personal, non-commercial use; commercial distribution and other unauthorised uses are not permitted. This notice does not restrict rights granted by the licences of individual works, including commercial uses permitted by CC BY-SA 4.0, or uses permitted by law. If you believe this edition infringes any rights or obligations, please notify us through the **contact form**.

Reading this edition
Each interior item begins on a new sheet. Research Proposals includes only the first web page of each selected publication; the link opens the publication online, including its further pages. Other interior publications are included in full. The selected text may occupy more than one printed sheet. Links, source references and image credits are retained. Menus, sharing controls, citation widgets and accessibility controls are omitted.

Contents

1 Front Cover

2 Publication details and contents

3 Open Models — Dory1.0-4B and the value of a second pass

5 Biographies — Claude Shannon

7 Human & AI Reasoning — When Artificial Touch Improves Grasp Control

9 Horizons of Electronics — When does cooling an AI processor from inside the silicon improve efficiency?

11 AI's Human Mirror — Walking After a Stroke: From the Clinic to a Gravel Path

13 The AI Oligarchy — What does Z.ai still control after it releases the weights of GLM-5.3?

14 Programming Languages — Slang versus Rust and CUDA C++ in neural rendering

16 Dead Languages and the Past — What can AI find in a lidar map of the Maya landscape?

17 AI and the Labour Market by Sector — When AI speeds up customer support, who benefits?

19 Medicine — An ECG biomarker for sudden cardiac death discovered with deep learning

21 Cybersecurity — LLM-Based Agents for Cybersecurity: A Systematic Review of Architectures, Applications, and Open Challenges

24 Automotive — Autonomous robotic operation controlled by wave-based neuromorphic hardware

26 AI and hardware markets — dated snapshots

28 AI and Society — Who answers for AI? Lessons from the newsroom and the assembly line

30 Back Cover — The Director's Eye

Dory1.0-4B and the value of a second pass

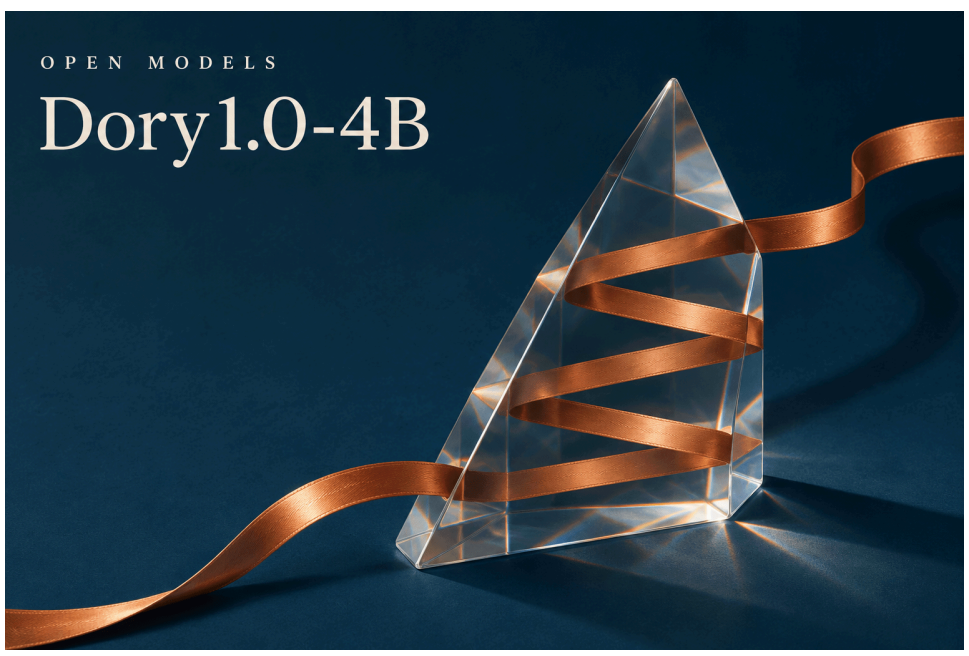
17 de noviembre de 2026

Original publication

Open Models · Documentary assessment · Model released 16 October 2026 · Evidence reviewed 17 November 2026

When a language model fails to repair a piece of software, would a second pass through the same network help? NVIDIA's Dory1.0-4B allows a particularly direct test. Its central block can run once, twice or four times while the learned weights stay fixed. The clearest answer comes from comparing two tasks. On software repair, the second pass raises the proportion of issues resolved from 50.3% to 59.5%. On a test of factual knowledge, accuracy changes from 10.03% to 10.08%. NVIDIA technical report, Table 2

Assessment. Dory merits inclusion in Open Models as an inspectable experiment in recurrent depth. Its strongest result is the gain from a second pass on particular tasks. Independent implementations also allow readers to examine whether a different engine preserves the model's numerical behaviour. The Bible of AI (BAI) reviewed the released materials and published measurements and did not run a new inference experiment.



© 2026 Juan Antonio Lloret Egea · ITVIA · The Bible of AI (BAI) · CC BY-NC-ND 4.0.

What changes on the second pass?

Before predicting the next token, Dory encodes the available text. Its middle block then combines that original representation with the state produced by the preceding pass. Each pass therefore applies the same learned transformation to a different input. One, two and four passes execute 72, 112 and 192 sublayers respectively, using the same 3.98 billion parameters. Two passes are the default in the released BF16 configuration. The intermediate states are vectors. Establishing what information a later pass corrects would require an intervention on those states. Earlier research on Huginn provides a scientific precedent for this use of recurrent depth. Released configuration · Architecture and Figure 2 · Recurrent-depth precedent

The laboratory and the release

NVIDIA, which designs accelerators and develops models and inference software, released Dory through Nemotron Labs. Its report describes six trillion tokens of pretraining followed by 300 billion tokens for context extension. Supervised fine-tuning used 128 GB300 GPUs across 32 nodes. A later reinforcement-learning stage assigned greater depth more often to difficult examples during training. That design leaves the choice of inference depth open. The report identifies equipment for particular stages but gives no itemised budget for the complete training run. Training, §§3–4 · NVIDIA model card

NVIDIA distributes trained weights in BF16, MXFP8, NVFP4 and GGUF formats. The English-language BF16 release includes its configuration and tokenizer. The OpenMDW 1.1 licence permits use, modification and distribution subject to its terms, including preservation of applicable notices. The published weights support comparisons across depths. The report and release do not supply a document-by-document training inventory sufficient to reconstruct the full training sequence. The configured limit of 262,144 tokens specifies the maximum sequence length. Performance over documents of that length remains a separate measurement. Official model collection · Configuration · Release licence

Where greater depth helps

Selected results from NVIDIA's Table 2. Individual task scores are percentages. The twelve-metric mean uses the authors' composite scale.

Evaluation	1 pass	2 passes	4 passes
Twelve-metric mean	35.17	39.67	40.60
SWE-Bench Verified · software repair	50.30	59.50	61.80
AA-LCR · long-context reasoning	63.00	71.00	73.00
GPQA Diamond · science questions	69.19	72.41	72.16
AA-Omniscience · factual accuracy	10.03	10.08	10.03

The equal-weight mean rises by 4.50 points from one pass to two and by a further 0.93 points from two to four. The authors rescale GDPval-AA Elo scores before averaging. On GPQA Diamond the observed score falls slightly after the second pass. The reported point estimates alone cannot establish whether that small difference exceeds sampling variation. Four passes also produced shorter outputs on four tasks requiring a sequence of actions. After accounting for the greater depth per token, NVIDIA estimated that generation still required 1.07–2.04 times as much computation as one pass. This estimate uses sublayer applications and output length. Elapsed time and energy require direct measurement. Tables 2 and 3

What users measured with llama.cpp

One user tested the Q5_K_M release on an RTX 5070 Ti with a Dory-specific development version of llama.cpp. With a separate attention cache for each pass, the initial decoding rates were 40, 27 and 16.5 tokens per second at one, two and four passes. A follow-up held the prompt at 512 tokens and the output at 128. Across five timed repetitions, placing the model's scaling tensors on the GPU raised the mean rate from 25.67 to 79.94 tokens per second. These two series answer different questions. The first follows speed across depths. The second compares two ways of placing the same data. User's protocol, hardware and results

A CPU experiment on a shared host yielded best-of-five rates of 3.30, 2.38 and 1.43 tokens per second at the same three depths. A separately registered test on a DGX Spark recorded 24.15, 16.02 and 9.56 tokens per second with one request at a time and a separate cache per pass. All 40 responses to its five simple prompts across eight configurations met the

authors' coherence criterion. With one run per configuration, this establishes basic operation and leaves differences in batched throughput unresolved. CPU experiment · DGX Spark experiment and deviations from protocol

What the Rust implementations establish

Numerical agreement on Apple silicon

Omegahee's Rust engine provides the closest independent check of whether an implementation preserves Dory's numerical behaviour. On Metal, the project compared converted BF16 weights against its FP32 reference at one, two and four passes, retaining either a separate attention cache for each pass or only the final pass's cache. All 31 cases met its stated tolerances. On the 1,830-token prompt, the engine and reference assigned the highest score to the same next token at 1,768 positions. This illustrates what passing a numerical tolerance means in practice. The authors also report about 25 generated tokens per second with two passes on an M5 Pro. These tests concern implementation fidelity and speed. They do not reproduce NVIDIA's task scores. Methods and device · 31 individual results

Adaptive depth in Cortiq

Cortiq offers a different experiment. Its 2.09 GB, four-bit Dory variant can assign one or two passes to each token and changes how distant context is retained. Over twelve prompts, its maintainers report that generation rose from 49.6 to 60.7 tokens per second on an NVIDIA L4 with Vulkan, and from 25.7 to 32.9 on a MacBook Air M4 with Metal, compared with their fixed two-pass variant. WikiText-2 perplexity rose from 13.96 to 14.22, indicating poorer prediction on that text. Routing and memory handling changed together, so the figures cannot isolate the contribution of either intervention. Paired measurements and evaluation · Rust engine release

Other routes through Rust

Eider provides Rust and CUDA code for generating text with NVIDIA's NVFP4 weights. Its separately published *Decisive Dory* has undergone additional training to choose among supplied answers. Results for that derivative concern a different set of weights. Rillwright places a Rust application around the Dory version of llama.cpp. Its authors report 80 successful runs on 16 small tasks repeated over five seeds. This tests the application and model together. Counting it as another independent numerical implementation would count llama.cpp twice. Eider generation example · Derivative model card · Rillwright evaluation and limitations

Jero provides Rust tests of the recurrent architecture for the SpacemiT K3 platform. Its published forward-pass test uses synthetic weights of width 256. That test demonstrates how the implementation can be examined, while leaving the speed and accuracy of the trained four-billion-parameter model on the board open. The reviewed Candle model registry contains recurrent architectures, including RecurrentGemma. A Dory-specific entry or documented run in Candle remained unverified at the review date. Jero test and weight construction · Candle model registry at review

Which result would change the assessment?

A decisive next comparison would hold the weights, prompts, scoring rules and hardware constant while varying depth. It would record task success, time to an accepted answer and energy per successful answer, including failed attempts. Memory deserves its own measure. In a separate four-pass Dory study at 16K context, LoopKV-M increased peak aggregate throughput from 219.7 to 247.9 tokens per second on a 32 GB RTX PRO 4500. The two peaks occurred at batch sizes of 24 and 48. That finding concerns a server handling many sequences together. A reader choosing how long to wait for one answer needs a single-request measurement. LoopKV, Appendix G, Table 20

Sources and scope of this assessment

The primary scientific reference is NVIDIA's *Dory1.0-4B: A Looped Normalized-GPT*, arXiv:2611.03998, version 1, submitted 2 November 2026. The model assessment concerns the post-trained BF16 release. Where community experiments use converted, quantised or further-trained weights, their results are attributed to those artefacts. The review covers materials available on 17 November 2026. No independent reproduction of the twelve-metric comparison is claimed. Primary report · Versioned release record

Claude Shannon

16 de noviembre de 2026

Original publication

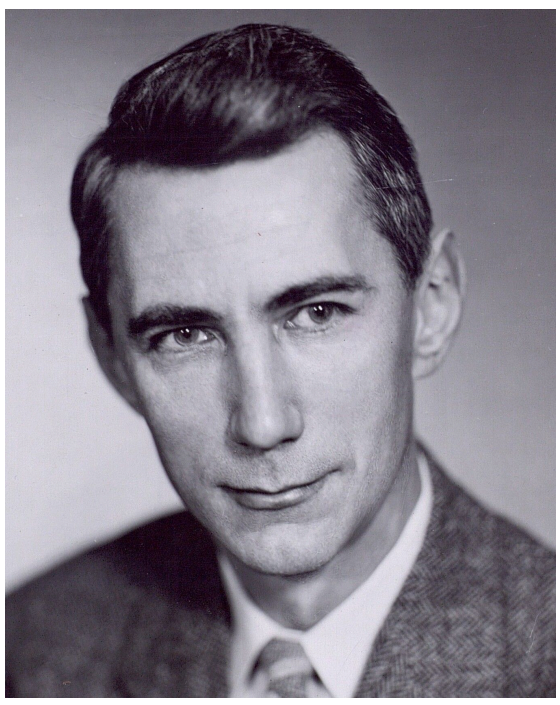
Biographies · Information theory and machine intelligence

As a boy in Gaylord, Michigan, Claude Shannon built a telegraph with a friend. Their houses were half a mile apart. They made the parts to send each other Morse code, then got hold of some equipment from the local telephone exchange and put in a phone. Years later, at Bell Laboratories, Shannon found a question common to both of those boyhood inventions. How much information could a line carry if noise kept disturbing the signal? His answer laid the foundations of information theory. Shannon recalls the line between the two houses.

Importance of his work for artificial intelligence

Imagine a mouse reaching the same fork in a maze twice. What would convince you it had learned something on its first visit? In a Bell Labs film, Shannon put a mechanical mouse called Theseus in a maze with movable walls. On its first trip it tried passages until it reached the goal. When he put it back at the start, it followed the route it had found. Then he placed it somewhere new. It resumed exploring until it met a familiar part of the route and could follow it home. Watch Shannon move the mouse to a new starting point.

The memory was under the floor. Telephone relays recorded which way to go, and a magnet moved the mouse above them. Betty Moore Shannon, a mathematician who had worked at Bell Labs, did much of the wiring. Lift the mouse off the board and the experience stays with the maze. What if you put the same mouse on a second board? The first board would still hold the directions. A change in behaviour had a physical cause that Shannon could show his audience. Betty Shannon’s contribution · How Theseus worked.



Claude Elwood Shannon. Photographer unknown. Tekniska museet, no. 43069. Crop by PhotographyEdits. CC BY 2.0.

At a glance

Born
30 April 1916, Petoskey, Michigan

Died
24 February 2001, Medford, Massachusetts

Education
University of Michigan and Massachusetts Institute of Technology

Fields
Electrical engineering and mathematics

When two switches mean “and”

Shannon earned degrees in mathematics and electrical engineering at the University of Michigan in 1936. At MIT he worked on Vannevar Bush’s differential analyser, a machine that solved equations. Its control circuits used relays, electrically operated switches. Shannon recognised in their wiring something he had studied in a mathematics class. Put two switches in series and current will flow only when both are closed. Put them in parallel and either one can complete the circuit. Here were the operations AND and OR of George Boole’s algebra, carried out by pieces of metal.

Shannon showed how to work in either direction. An engineer could write down what an existing circuit did, or start with the conditions a new one should satisfy and design its switches on paper. His study appeared in 1938. In 1940 MIT awarded him a master’s degree in electrical engineering and a doctorate in mathematics. After a year at the Institute for Advanced Study in Princeton, he joined Bell Laboratories in 1941. Shannon’s paper on relays · MIT’s biographical account.

What can a noisy line carry?

Suppose a friend has hidden a key in one of eight equally likely lockers. Ask whether it is in the left half. With one answer you are down to four lockers. Two more well-chosen questions settle the matter. Each answer to an even split gives you one bit of information. If the key is nearly always in the first locker, guessing where it is presents a different problem. Shannon used *entropy* to measure the average uncertainty when the alternatives and their probabilities are known.

In *A Mathematical Theory of Communication*, published in 1948, Shannon also established a limit for transmission through a noisy channel. Below its capacity, suitable codes operating on sufficiently long groups of symbols can make the probability of a decoding error as small as desired. For an engineer designing a link, it meant that a noisy wire could still carry an exceptionally faithful message. The recipient would decide what that message meant. The mathematics answered how reliably its symbols could arrive. Shannon’s original paper.

A chess player has to stop looking ahead

A rook is there for the taking. Before you reach for it, you see what your opponent would play next. How could a machine exercise that small piece of judgement? In 1950 Shannon described two ways to look ahead in chess. A computer could examine all legal moves to a chosen depth, or devote more of its time to promising continuations. Where the search stopped, it would need a way to judge the resulting position. The balance of material between the two players was one clue. The safety of the king was another. The distinction between choosing what to explore and judging what one finds is still recognisable to anyone who writes a chess program. Shannon’s chess paper in the Computer History Museum.

He also tried prediction with human readers. In 1951 he reported experiments in which a reader guessed the next letter of an English passage and then tried again after a wrong guess. The letters already revealed made some continuations much easier to anticipate. His question was how much uncertainty remained. In 1955 he joined John McCarthy, Marvin Minsky and Nathaniel Rochester in proposing the following summer's Dartmouth research project on artificial intelligence. By then he had approached machine intelligence with a maze, a chessboard and a page of ordinary prose. The letter-prediction experiment · The four authors' Dartmouth proposal.

The toy room in Winchester

Robert Gallager, a colleague at MIT, remembered that Shannon generally liked to work alone. Edward Thorp, who collaborated with him, remembered arriving at his study one day to find the calculation $2^{11} = 2048$ on the blackboard. Shannon and Betty had married in 1949. When he returned to MIT in 1956, they settled in Winchester, Massachusetts. The house had a room that they called the toy room. It held chess sets, musical instruments and juggling equipment. Gallager's recollections · Thorp's account of working with Shannon · Erico Guizzo on the toy room.

Shannon became professor emeritus in 1978. Some years later, the interviewer Anthony Liversidge saw a machine in the toy room whose soft hands juggled steel balls. He asked Shannon how many balls he could juggle himself. Four, Shannon said. With five he could get them into the air. Catching them was the difficulty. Liversidge's interview with Shannon.

A chronology

1. **1916–1936.** Grows up in Gaylord and studies mathematics and electrical engineering in Michigan.
2. **1938–1941.** Publishes his work on relay circuits, completes two graduate degrees at MIT and joins Bell Laboratories.
3. **1948.** Publishes *A Mathematical Theory of Communication* in two instalments.
4. **1950–1955.** Publishes on chess and letter prediction, demonstrates Theseus and co-authors the Dartmouth proposal.
5. **1956–2001.** Returns to MIT, becomes professor emeritus in 1978 and dies in Medford at the age of 84.

Further reading

In the Bell Labs film of Theseus, watch where Shannon puts the mouse after it has learned its first route. Then turn to his own paragraph in the Dartmouth proposal. Shannon suggests starting with a simple environment and studying what a machine can learn there before asking it to manage a more complicated one. The maze gives that proposal a scale you can picture. Its walls could be moved by hand.

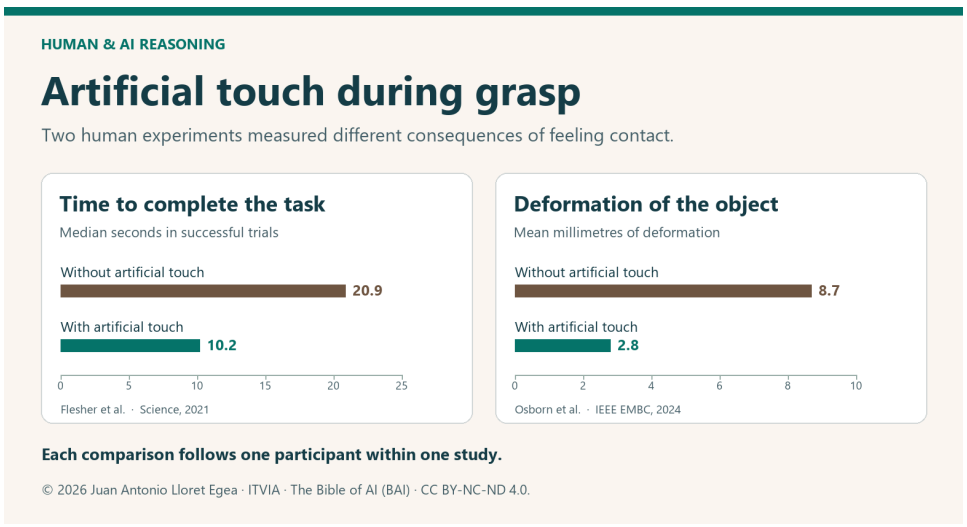
When Artificial Touch Improves Grasp Control

17 de noviembre de 2026

Original publication

A hand closes around a hollow eggshell. Before lifting it, the person has to judge whether the fingers have made contact. As the grip tightens, the person must also judge how much force the shell will bear. A sensor can help with either task. What matters is where its signal goes and who can act on it.

A bidirectional brain–computer interface carries information both ways. A decoder translates recorded brain activity into a movement command. Sensors on a robotic hand or a hand orthosis can send information about contact back to the user through stimulation of the sensory cortex. A controller can also use a sensor reading to adjust grip force automatically. Six human studies help trace the contributions of perception, movement and automatic assistance.



Artificial touch during grasp. With touch added, median completion time fell from 20.9 to 10.2 seconds in successful trials. In a separate study, mean object deformation fell from 8.7 to 2.8 millimetres. Each comparison involved one participant. © 2026 Juan Antonio Lloret Egea · ITVIA · The Bible of AI (BAI) · CC BY-NC-ND 4.0.

What does contact feel like?

For a tactile signal to guide action, the person must be able to interpret it. In a 2016 study, a participant with a spinal cord injury felt sensations in particular parts of the hand when researchers stimulated the corresponding region of the somatosensory cortex. Some sensations felt like pressure. Their perceived intensity changed with stimulation amplitude. The participant's reports showed that the signal could convey where contact occurred and how strong it felt. A separate experiment was needed to establish whether that information could change a grasp. [1]

A later study asked what spatial detail the signal could carry. Two participants distinguished the orientation of edges evoked on their fingers. Accuracy varied across fingers, and similar shapes could be confused. This matters when an object is beginning to tilt. Knowing that a fingertip is touching something provides one cue. Knowing the orientation of the contact could help the person decide which way to move. Whether that cue prevents a particular object from slipping is a question for a manipulation task. [2]

From first contact to lifting an object

In an experiment with a brain-controlled robotic arm, a participant with tetraplegia could see the arm and the object. The researchers added a tactile signal on some trials. Among trials completed successfully, the median time fell from 20.9 seconds with vision alone to 10.2 seconds with vision and artificial touch. Most of the time gained was between first contact and lifting the object. The participant could already reach the object under visual guidance. The added information was particularly useful when deciding whether the grip was secure enough to lift. [3]

Speed answers only part of the question. A fast grasp can damage a delicate object. In a separate study, a participant using a robotic hand tried to grasp a deformable object without crushing it. Mean deformation was 8.7 millimetres without cortical force feedback and 2.8 millimetres with it. Here the object itself provided a measure of how the grasp changed. The two experiments used different tasks, so their results are best read in their original units. [4]

Who controlled the force on the eggshell?

A 2026 study examined grip force and contact detection in separate experiments. A participant with tetraplegia used an orthosis fitted to his own hand. A neural decoder detected his intention to grasp. A controller trained by reinforcement learning could then adjust the orthosis to keep the applied force within the prescribed limit. During the force-control test, the participant could see a live force trace while the hand and eggshell were hidden from view. Across five sessions per condition, mean success was 87% with the automatic regulator active and 27% with it inactive. This comparison measures the contribution of the regulator. [5]

In another test, the regulator remained available and the participant had to decide whether an eggshell was actually in his grasp before lifting his arm. With vision occluded, his decisions were correct on every trial with cortical feedback and at chance level without it. Here the tactile signal informed the person's decision to lift. The two experiments give distinct meanings to the everyday expression *holding an egg successfully*. [5]

How long does the signal remain available?

Five people with spinal cord injury took part in a study of implanted sensory arrays over periods ranging from two to ten years. In the participant followed for a decade, 60% of the electrodes could still evoke tactile sensations at the ten-year assessment. The investigators also recorded occasional sensations that briefly outlasted stimulation. That percentage describes electrodes in one person. Establishing how often someone can use those sensations to handle objects independently at home requires different observations, including how much help is needed to set up the equipment. [6]

What would establish shared control?

Consider an unfamiliar object whose stiffness the user cannot judge by sight. A useful experiment would vary the availability of the contact signal and the automatic regulation of force independently. Keeping the object, visual information and opportunities for practice comparable would allow four conditions to be tested. Recording contact judgements, excessive force, damage and time to lift would reveal where either contribution makes a difference. Failed attempts would matter just as much as successful ones.

This proposed test follows from a limit of the present evidence. Several studies report repeated trials in one person. Those trials can reveal a reliable change within a task, while the size of the benefit for another person or another device remains an open question. The evidence is most

informative when it identifies which part of the grasp changed. The user might feel contact sooner, judge more confidently when to lift, or rely on the controller to ease the pressure before the shell breaks.

Selective review of six primary studies published between 2016 and 2026. Evidence reviewed on 17 November 2026. The four-condition comparison above is proposed as a future experiment.

References

1. Flesher, S. N., et al. (2016). Intracortical microstimulation of human somatosensory cortex. *Science Translational Medicine*.
2. Valle, G., et al. (2025). Tactile edges and motion via patterned microstimulation of the human somatosensory cortex. *Science*.
3. Flesher, S. N., et al. (2021). A brain-computer interface that evokes tactile sensations improves robotic arm control. *Science*.
4. Osborn, L. E., et al. (2024). Artificial touch feedback using microstimulation of human somatosensory cortex to convey grip force from a robotic hand. *46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.
5. Chandrasekaran, S., et al. (2026). A neuroprosthesis for restoring hand movement and sensation in a person with complete tetraplegia. *Nature Medicine*.
6. Greenspon, C. M., et al. (2026). Long-term safety and efficacy of intracortical microstimulation in humans. *Science Translational Medicine*.

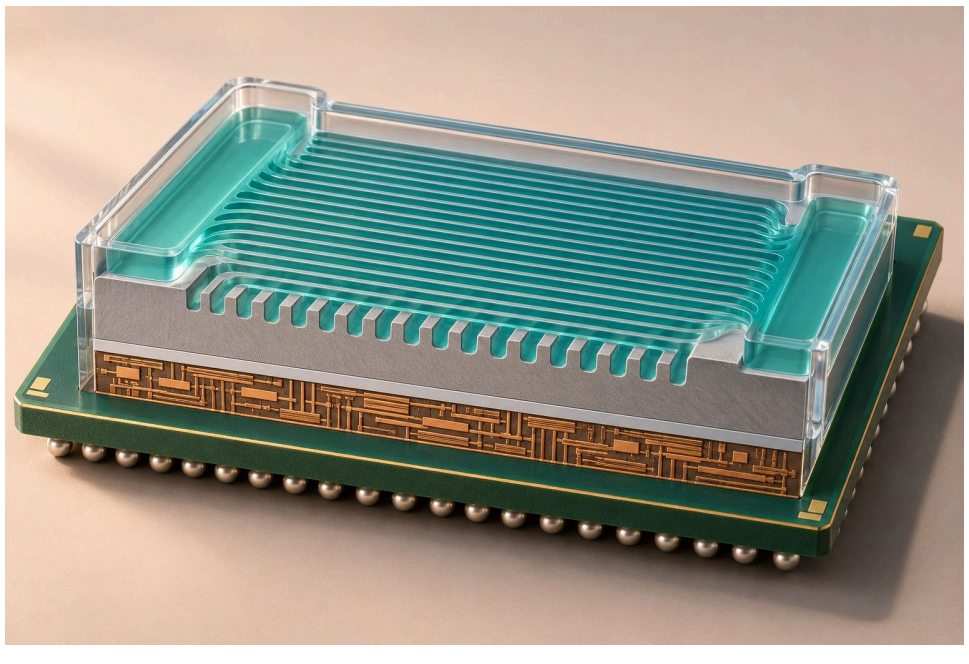
When does cooling an AI processor from inside the silicon improve efficiency?

18 de noviembre de 2026

Original publication

During a long AI training run, a small region of an accelerator can reach its temperature limit while the cooling water is still comparatively cool. The processor may then have to lower its clock frequency. How much of that lost computing time could be recovered by bringing the water closer to the transistors?

Researchers at Microsoft and Corintis tested one answer on an NVIDIA GH200. They etched fluid passages into the back of the graphics processor and measured how much hotter the device became than the incoming coolant. The experiment also tracked the nearby high-bandwidth memory that supplies the processor with data. To understand what either result means for AI, we need to follow the heat out of the chip and the electricity into the data centre. [1]



The channels bring liquid close to the heat source. Colour and dimensions are illustrative. The external pump and heat exchanger are outside this view. © 2026 Juan Antonio Lloret Egea · ITVIA · The Bible of AI (BAI) · CC BY-NC-ND 4.0.

Where does the heat meet the liquid?

In NVIDIA's Grace Blackwell rack systems, liquid already flows through cold plates attached to the CPUs and GPUs. Heat travels out of the device, across the thermal contact and into the plate before the liquid carries it away. Etching channels in the silicon shortens that path. A sealed cover distributes coolant through the passages, and their walls provide additional surface for heat exchange. The liquid stays separated from the electronic layers. A pump moves it to a heat exchanger, and the cooling plant must ultimately release the heat to the surroundings. [1] [4] [11]

A useful measurement is the temperature difference between the active silicon and the coolant at the inlet, divided by the power dissipated by the chip. This thermal resistance, expressed in kelvins per watt, tells us how much the silicon warms up at a given load. Georgia Tech and Microsoft etched the back of a working Intel Core i7-8700K. Compared with their conventional cold plate, they measured up to 44.4% lower junction-to-inlet thermal resistance and demonstrated higher sustained core frequencies. Here the cooler changed something a program could use: how fast the CPU could keep running. [2]

What did the accelerator and its memory gain?

On the GH200 platform, the researchers reported up to a **60% reduction in junction-to-inlet temperature rise at the GPU** and approximately **35% at the high-bandwidth memory**, relative to conventional cold plates. These percentages describe temperature differences. For example, if a chip were 40 degrees above its inlet, a 60% reduction in that difference would leave it 16 degrees above the inlet. This arithmetic illustrates the unit of comparison. Electricity use requires a separate measurement. [1]

During sustained matrix calculations, a cooler processor can finish sooner if temperature had been limiting its clock. A different bottleneck arises when a language model generates one token at a time and the processor waits for weights to arrive from memory. Cooling the GPU may then leave response time almost unchanged. Cooling the memory could help if its own thermal limit was restricting transfers. A convincing comparison would run the same model at the same output quality, batch size and latency requirement, then measure tokens per second and total joules per response. How memory bandwidth affects AI inference. [1]

How much electricity could a data centre save?

A retrofit at Lawrence Livermore National Laboratory illustrates why the answer depends on the building. Engineers fitted the Cabernet computing cluster with water-cooled plates on its hottest components. The remaining heat still went to room air, and both routes used the existing chilled-water plant. Using operating data, the researchers estimated **4% savings in overall data-centre energy** from their site-level analysis and **3%** from a component-by-component analysis. The main electrical benefit came from slowing the fans in the computer-room air handlers. This was a partial cold-plate retrofit at a working computing centre. [5]

Lawrence Berkeley National Laboratory examined a different installation. It instrumented 38 liquid-cooled servers and used two unmodified servers as a reference. The researchers combined measurements with models of alternative cooling plants. At the test's 50% server power level, the projected reduction in total site energy ranged from **14% to 20%**. A separate life-cycle assessment by Alissa and colleagues estimated **15–20% lower energy demand per virtual core per year** for cold-plate and immersion scenarios compared with air cooling. That boundary includes energy beyond the facility's electricity meter. Each study compares its own systems. The additional site-wide electricity saving from replacing an existing cold plate with channels etched in the GPU remains unquantified in the GH200 study. [1] [6] [7]

Consider an illustrative centre using 100 MWh for a fixed amount of AI work. Suppose cooling accounts for 10 MWh. If a change halves that cooling electricity, including pumps and heat rejection, and everything else stays constant, the centre uses 95 MWh: a **5% reduction at the meter**. A site can spend the recovered electrical capacity on additional inference requests instead. In that case, its benefit is more useful work within its grid connection, and its electricity bill may stay much the same.

Can operators adapt the equipment they have?

A rack already fitted with cold plates has part of the necessary infrastructure: supply and return manifolds, pumps and a route for rejecting heat. Reuse depends on the coolant chemistry, inlet temperature, pressure and flow that the new device requires. Smaller passages may need finer filtration. Pumping power depends on both flow and pressure drop, so a smaller flow alone cannot establish an electricity saving. An air-cooled room would also need a liquid distribution circuit, connections at the racks and a way to detect and isolate a leak. HP and NVIDIA have explored an intermediate approach, a separately fabricated silicon microchannel plate coupled to the GPU. It moves the coolant closer while allowing the cooling structure to be made separately from the processor. [2] [4] [10] [11]

The more demanding change is at the factory. Etching a finished die must leave enough silicon for mechanical strength, and the seal must withstand thermal cycling. Microsoft and Corintis have investigated mechanical stress, clogging and erosion. TSMC separately demonstrated a CoWoS-R packaging test vehicle with silicon micropillars that dissipated over 5 kW under uniform heating, using 6 litres per minute of deionised water at 40 °C. At Peking University a much smaller heater experiment reached 3,000 W/cm² at a hydraulic pumping power of 0.9 W/cm². One experiment measures total heat over a package, the other heat per unit area. Production readiness also depends on manufacturing yield, maintenance intervals and failures observed over years of service. [3] [8] [9]

When would this help AI expand?

Where a training cluster repeatedly loses clock speed to hot spots, sustained throughput is the first outcome to measure. Where the constraint is a fixed electricity supply, the relevant result is more completed training or inference work for each kilowatt-hour drawn by the site. And where memory is stacked close to logic, the physical challenge is getting heat out of an interior layer without displacing the electrical connections that made the stack attractive. Angstrom-class process names describe a manufacturing generation. The cooling passages can remain micrometre-sized: their purpose is to carry away heat from a region containing many transistors. How backside electrical connections use the same side of the chip. [1] [8] [11]

Bringing coolant inside the silicon is justified when the extra thermal margin translates into sustained computation or lower site electricity for the same AI service, and the package can deliver that gain reliably. For a system already meeting its thermal and latency targets with external plates, the decisive test is a matched run that includes the pumps and cooling plant. A cooler chip settles the thermal question. The electricity meter and the completed workload settle the efficiency question.

Sources and scope

Evidence reviewed through 18 November 2026. The CPU and GH200 studies concern working processors. The Peking and TSMC results concern thermal test devices. The facility comparisons come from separate installations and models. The CPU, GH200, reliability and life-cycle publications share some contributors or institutional affiliations.

1. Majety, S., et al. (2026). Direct to Silicon Microfluidic Cooling for Datacenters. IEEE Electronic Components and Technology Conference, pp. 926–929. GPU and high-bandwidth memory tests.
2. Kochupurackal Rajan, S., et al. (2022). Integrated Silicon Microfluidic Cooling of a High-Power Overclocked CPU for Efficient Thermal Management. IEEE Access, 10, 59259–59269. Operating CPU and cold-plate reference.
3. Wu, Z., et al. (2025). Jet-enhanced manifold microchannels for cooling electronics up to a heat flux of 3,000 W cm^{-2}. Nature Electronics, 8, 810–817. Heated test chip and hydraulic power.
4. NVIDIA. DGX GB Rack Scale Systems User Guide, Hardware. Installed arrangement of CPU and GPU cold plates in the compute trays.
5. Lawrence Berkeley National Laboratory and collaborators (2020). The Installation of Direct Water-Cooling Systems to Reduce Cooling Energy Requirements for High-Performance Computing Centers. ACEEE Summer Study on Energy Efficiency in Buildings. Two analyses of the Cabernet retrofit at Lawrence Livermore.
6. Coles, H., and Greenberg, S. (2014). Direct Liquid Cooling for Electronic Equipment. Lawrence Berkeley National Laboratory, LBNL-6641E. Instrumented servers and modelled site energy at different loads.
7. Alissa, H., et al. (2025). Using life cycle assessment to drive innovation for sustainable cool clouds. Nature, 641, 331–338. Life-cycle energy demand per virtual core per year.
8. Jiang, S.-D., et al. (2026). Process Development and Thermal Characterization of Micropillar Direct-to-Silicon Liquid Cooling Solution on CoWoS-R Platform. IEEE Electronic Components and Technology Conference, pp. 546–552. Packaging test vehicle under uniform heating.
9. Microsoft and Corintis (2026). Reliability of In-Silicon Etched Microfluidic Cooling Systems. IEEE International Reliability Physics Symposium. Mechanical and fluid-path reliability testing.
10. HP (2025). Embedded Microfluidic Cooling for Nextgen High-Power Server Architectures. ARPA-E COOLERCHIPS programme, 4 December. Silicon cold plate coupled to a GPU.
11. Microsoft (2025). AI chips are getting hotter. A microfluidics breakthrough goes straight to the silicon to cool up to three times better. Prototyping, cooling passages and integration constraints.

Explore the Horizons of Electronics collection.

Walking After a Stroke: From the Clinic to a Gravel Path

17 de noviembre de 2026

Original publication

One of the four walkers chose a road with gentle slopes and a gravel path in a park. Another chose an urban greenway. All four had survived a stroke. When researchers asked where they would like to try a soft robotic ankle support, each picked somewhere close to home or work. The team checked the routes. Then the walkers took the equipment out on their own. For four weeks, a question usually asked on a laboratory floor went with them: how much of the next step could the weaker leg do for itself? The community walking study.

The choice of route deserves as much attention as the machine. In a separate interview study, a man explained why he and his wife went to a park to walk. There were benches. He could judge the distance to the next one and know where he could rest if his leg stiffened. He had the physical ability to walk outdoors in his neighbourhood. What the park gave him was a reason to trust the return journey. Read the study of 20 stroke survivors' experiences outdoors.

Before the first step, a buckle

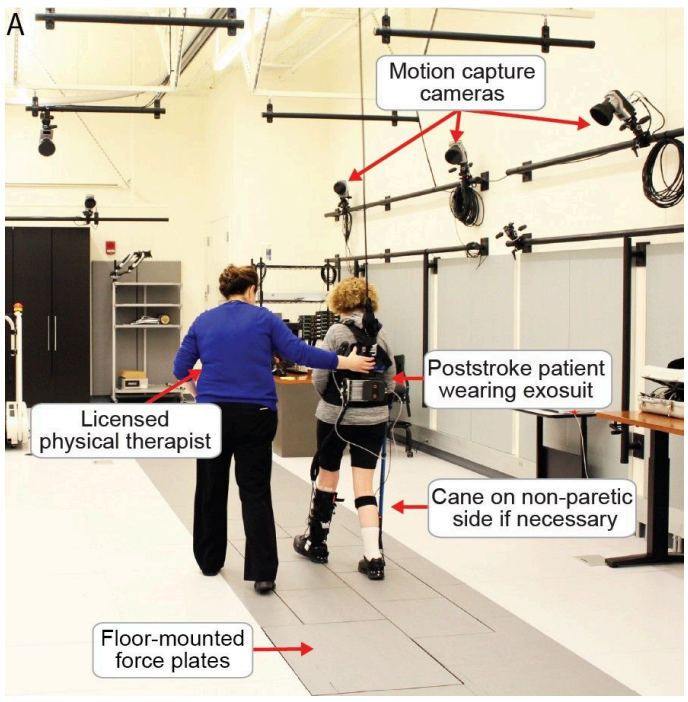
A stroke can leave one side of the body weaker than the other. Watch what your own ankle does when you take a slow step. As your heel comes up, the front of your foot presses into the ground and helps send your body forward. A moment later, the toes have to clear the ground as that leg swings through. Someone can cover a respectable distance by asking the stronger leg to do more, while getting relatively little push from the other one. Timing the walk tells only part of that story.

The experimental ankle support addressed both moments. A flexible rod helped hold the foot up during the swing. A motor pulled on a cable to help the ankle push off. The rod did a second job before the walker even stood up: it held the calf wrap in shape while the person fastened it. A magnetic buckle and straps could be managed mainly with the stronger hand. At their final assessment, the four people took roughly two to three and a half minutes each to put the suit on. That modest design detail put the decision to go out within the wearer's reach. Device design and fitting procedure.

All four were experienced walkers after stroke and had previously tried similar equipment in the laboratory. They demonstrated that they could put it on, switch it on and off, and judge their exertion. A member of the research team also checked their ability to use it at the chosen location. That preparation matters to anyone who imagines taking the same device home. The study began with people who could already make the journey. The question was what assistance could add.

What the floor knows

In the photograph below, a therapist walks beside a participant in a different study of ankle assistance. The rectangles beneath their feet are force plates. Whenever a foot lands on one, the plate records the forces that the walker exerts on the ground. Cameras above the track follow the movement. That combination lets researchers connect something they can see, a foot clearing the floor, with something they cannot judge by eye, the force that helped carry the walker forward.



A therapist walks alongside a participant testing a soft ankle exosuit on an indoor track. Photograph from Sloot et al., *Journal of NeuroEngineering and Rehabilitation* (2023), supplementary figure 1. Cropped to panel A, with the original annotations retained. CC BY 4.0.

Nineteen people completed that laboratory study. Each walked around an indoor track at a comfortable pace, once wearing the active suit and once without it. The slower walkers increased their speed by an average of 0.08 metres per second with assistance. The faster group showed little change in average speed. The team also assessed energy expenditure and found no significant overall change in the energy needed to cover a given distance. A person can therefore have a quicker step in this test without the researchers being able to promise a less tiring outing. The study pictured above.

There are no force plates under an ordinary gravel path. To follow the four walkers into the community, the researchers used small motion sensors on the foot, lower leg and pelvis. They trained a model by pairing movement signals with forces measured in the laboratory, then adapted it to each walker using overground measurements. Out on a route, it could estimate how strongly the affected leg pushed forward. The distinction is quite practical. The cable helped the person walk. Machine learning helped the researchers interpret the walk afterwards. Whoever chose the park knew something neither system measured. How the researchers estimated push-off outdoors.

The week the suit came off

All four finished the programme. Two of them, the walkers who had started with the weakest push-off, produced more forward force with the affected leg even when tested afterwards without the suit. The walker who chose the gravel path improved after one four-week programme. The walker on the greenway showed little change at that point and improved only after a second four-week programme. These two walkers also averaged about 4,000 more steps a day in the week after the programme than in the week before it. Of the other two, one recorded fewer steps during a spell of poor weather; the fourth did not wear the step counter consistently enough to provide a comparison. Those are results from four individual lives, and the follow-up for that step count lasted one week. Individual results and follow-up.

There was no comparison group taking the same walks without the device. A scheduled walk several times a week, familiarity with the route and the assistance itself could each have mattered. Imagine wanting to know whether the suit would help you get to the park a month after you handed it back. The fairest answer is that the study gives you a reason to ask for that test. It also gives the researchers a clue about whom to invite, including people with a weak push-off and others whose main difficulty is something else.

The number on a step counter has a further limitation. A thousand steps can bring someone to a friend's door or take them around an indoor track and back to the starting line. Both may be worthwhile, for different reasons. People interviewed about walking after stroke described the freedom of doing something without depending on someone else, and the hope they took from seeing progress for themselves. The destination, the company and the choice to turn back belong in any account of success. What real-world walking meant to people after stroke.

Would the walk feel easier?

Seven people with long-standing weakness after stroke tried a different device in a study published in 2026. This one assisted the hips. On a treadmill, the net metabolic cost of walking was 18 per cent lower on average with the device than without it. For someone who plans an outing around how much energy the return trip may take, that is a result worth hearing. The researchers measured the energy used during treadmill walking. How far a participant would choose to walk in a neighbourhood, or how tired they would feel that evening, remains an open question. Portable hip assistance after stroke.

Timing that assistance is another job that AI can do. In a separate engineering study, a controller learned in simulation when to assist the hips and was then tested during walking, running and stair climbing. It reduced energy expenditure in those tests. The community ankle study used learning to estimate a force. The engineering team used learning to control a motor. A newer preprint used signals from five stroke survivors wearing a powered knee device to estimate where each person was in the step cycle. Its estimates were tested on recorded walks. Whether acting on them would make a future walk easier or safer still has to be tested with people. Learning to control hip assistance · Personalised timing study, reported as a preprint.

Families early in recovery face a different choice: how best to use time with a therapist. An international randomised trial assigned 151 people with substantial difficulty walking after a recent stroke to two schedules. One group divided its training time between an exoskeleton and conventional walking therapy. The other spent the same time in conventional training. After four weeks both groups had improved, and the trial found no significant difference between them on its main measure of walking ability. Leg strength improved more in the group that used the device. That is encouraging for strength, while the answer about walking independence remains more modest. Multicentre rehabilitation trial.

What a rehabilitation team can offer now

A person need not wait for an experimental suit to talk to a physiotherapist about a familiar route. In England, NICE recommends walking training for people who can already walk with or without assistance. Its guidance also asks clinicians to consider an ankle–foot orthosis when difficulty clearing the foot or controlling the leg during stance interferes with walking. This is a brace. Part of assessing it is finding out how comfortable it is, whether the person can put it on and what difference it makes to walking. NICE places electromechanical gait training in the context of research. NICE guidance on walking after stroke.

A device can have a recognised clinical use and still be unsuitable for an unsupervised trip to the shops. In the United States, for example, the FDA clearance for the ReStore soft exosuit specifies rehabilitation institutions and supervision by a trained therapist, for people meeting its walking criteria. That is a different arrangement from the experimental community programme. A useful conversation at an appointment might start with the route itself: where the foot catches, where the walker needs to stop and whether getting ready to leave already requires another person. Each detail gives the therapist and the walker something specific to assess together. FDA indications for ReStore.

The strongest reason to take these machines outdoors is that people already know where they would like to go. On a gravel path, the researchers could estimate how much the weaker leg helped push the walker forward. The man who chose a park for its benches had identified something equally precise about his own mobility. He needed to know there would be a place to sit. A good rehabilitation programme can make room for both kinds of knowledge, and leave the last word about the value of the walk to the person who made it.

What does Z.ai still control after it releases the weights of GLM-5.3?

17 de noviembre de 2026

Information checked on 17 November 2026. Market shares and company revenue refer to the first half of 2026.

Original publication

A cloud company can download GLM-5.3, install it on its own computers and inspect the numerical weights that shape its answers. Can it then sell access to customers? For some companies, Z.ai has reserved a say in that decision. The answer depends on what the group does and how much revenue it earns across its businesses.



Zhang Peng
Co-founder and CEO of Z.ai. Shown answering reporters' questions while serving as the company's chief technology officer. Photograph published by Z.ai. Image rights reserved.

The person

Zhang worked at Tsinghua University and joined the company in 2019. Its annual report assigns him responsibility for research and development, business development and day-to-day operations. Fellow co-founder Liu Debing chairs the board and oversees strategic planning. The licence names Z.ai as the reviewer and leaves the individual decision-maker unspecified. Directors and senior management, p. 63 · Zhang Peng's interview with Zhang Xiaojun.

Z.AI Co., Ltd. · Beijing, China

Z.ai, also known as Zhipu AI, develops the GLM family of models. A customer can run a downloadable model on its own equipment or pay to send requests to a service that runs it. The distinction matters to Z.ai's business. In the six months to 30 June 2026, its open platform and programming interfaces brought in 825.2 million yuan, or 86.5% of the group's 953.9 million yuan in revenue. Those figures cover the company's business before GLM-5.3 appeared in August. Interim report, business review and revenue breakdown, p. 30.

Capital and control. The company listed its shares in Hong Kong in January 2026 under stock code 2513. Investors can buy a stake in the business. A company that downloads GLM receives the rights specified in the licence for that version. In its first-half 2026 survey, IDC put Zhipu second by revenue in China's market for large language models delivered as a service, at 10.6%, behind Alibaba at 46.6%. The figures describe sales in that Chinese market. Company and share listing · IDC's comparison of providers.

What changed between two downloads?

With GLM-5.2, Z.ai published the weights under the MIT licence. A company could run the model, adapt it and sell access, keeping the required notices with copies. The full GLM-5.3 licence also permits copying, adaptation and commercial use. Its second clause introduces a review for a defined group of companies. That is a decision about a particular release. Even the similarly named GLM-5.3-Flash carries the MIT licence. GLM-5.2 licence · GLM-5.3 licence, clause 2 · GLM-5.3-Flash licence.

Which companies must pass the review?

Two conditions must coincide. First, the licensee or an affiliate must offer model services that give customers meaningful control over the inputs, settings or training data used to generate answers or adapt a model. Second, the licensee and its affiliates together must earn more than US$10 billion over a period of twelve consecutive months. Revenue from every line of business counts toward that threshold. When both conditions hold, the licensee must pass Z.ai's security review before any commercial use of GLM-5.3 or a derivative. Z.ai can reasonably determine how it conducts that review and what it covers. Z.ai's definition and review condition.

Consider a retailer that adds a writing assistant to its product catalogue. Under the licence, embedding a model solely in a particular product feature does not itself make the retailer a model-service provider. Simply forwarding requests to someone else's hosted model also falls outside the definition. The analysis changes if another company in the retailer's group already runs a qualifying model service. The licence makes the group relationship material, even if the retailer's own AI project is small. Scope of the two exclusions.

What can a customer actually choose?

On 5 October 2026, Amazon announced GLM-5.3 on Bedrock for eligible enterprise customers. That gives a business a way to use the model through AWS. Amazon's announcement establishes availability and identifies the customers it can serve. It gives no account of a security review or the terms the two companies agreed. A business seeking to host a model itself can instead test GLM-5.2 or GLM-5.3-Flash under their published terms. Whether either can do the job depends on the customer's workload. AWS launch announcement · Bedrock availability.

What would show the practical reach of Z.ai's condition? A provider could disclose when it requested a review, when Z.ai decided, and whether the result changed a planned launch. Until such a case is documented, the public evidence establishes who holds the contractual decision. Its cost in time or lost business remains an open question.

Slang versus Rust and CUDA C++ in neural rendering

17 de noviembre de 2026

Original publication

New Programming Languages versus Rust / C++

Tilt a piece of leather under a lamp. The grain stays in place, but the highlight moves and changes shape. NVIDIA's graphics researchers wanted a small neural network to reproduce that behaviour for lighting and viewing angles it would encounter while drawing a scene. They used Slang in the software that learns a material's appearance and in the code that evaluates a learned material. A physics team faced a different choice. After trying Slang for its Nexus engine, Dimforge chose Rust so that a calculation on the graphics processor could share code and data types with the rest of the simulator.

Slang at a glance

Origins
A Carnegie Mellon, NVIDIA and Stanford research collaboration presented the language at SIGGRAPH 2018. Original language paper.

Field
Graphics and parallel computation, including shaders that evaluate a learned material or calculate a derivative.

Edition reviewed
Slang v2026.22. The timings in Figure 1 come from the 2023 Slang.D study.

Output
The compiler produces code for a chosen target, including SPIR-V for Vulkan and code for CUDA.

Stewardship
Khronos launched the Slang initiative in 2024 after NVIDIA contributed the open-source compiler.

The creator

The Slang research team

Video: Slang Shading Language Advances, a presentation by the language research team

The language team presents the design in its 2022 talk, *Slang Shading Language Advances*. Watch the original recording · Talk and slides.

A problem with adding materials
Every new material must work with the lighting methods in a renderer. The original project used modules and interfaces to let a team add a material without editing every combination of light and surface by hand.

Where differentiation came in
The Slang.D researchers later extended the language so that the compiler could generate derivatives from existing shading code. They demonstrated it in the Falcor renderer. Original design · Slang.D study.

What runs on the GPU, and who supplies the memory?

A shader is a program the renderer runs on the graphics processor. A material shader might read a texture, take the directions of the light and the viewer, and calculate how much light reaches the eye. The host application creates GPU resources and makes them available to the shader. Slang can declare an input buffer as read-only and an output buffer as writable. For Vulkan, the compiler can turn that program into SPIR-V, which the graphics driver consumes. The application remains responsible for keeping the resources available while the GPU uses them. Slang's worked example · How buffer declarations map to Vulkan.

How NVIDIA teaches a network the appearance of leather

NVIDIA's published example starts with a description of faux leather in MaterialX. That description already specifies how the reference surface scatters light. The researchers can therefore generate training examples directly on the GPU. For a point on the surface and a pair of directions, the reference material supplies the answer the network should learn. The system learns small networks and textures that store information about how the material varies across its surface. Slang, SlangPy and the Falcor2 renderer supply the graphics computation, with Python coordinating the work. NVIDIA's training pipeline and faux-leather example.

Once a material is learned, its use in an image is a separate step. In *Real-Time Neural Appearance Models*, the researchers generated one Slang module for each neural material. Each exposes the interface the renderer already uses for conventional materials. When a simulated light ray hits the surface, the renderer can ask how that surface responds to light and the learned network supplies the answer. NVIDIA also documents use of the Slang language in the Omniverse and RTX Remix renderers. Research implementation of neural materials · NVIDIA's account of renderer adoption.

What the CUDA C++ comparison actually measured

The first experiment asks how long one shading operation takes. The second timing adds the backward pass, which calculates how a change in an input would affect the result. The researchers compared the derivatives generated by Slang.D with handwritten CUDA C++ routines. On a V100, CUDA C++ took less time to produce the image alone. Including the backward pass reversed the order of the two reported means. The study also measured the loss, a measure of disagreement with a target, and the calculation of surface normals, which describe the orientation of a surface. Slang.D, Tables 3 and 4.

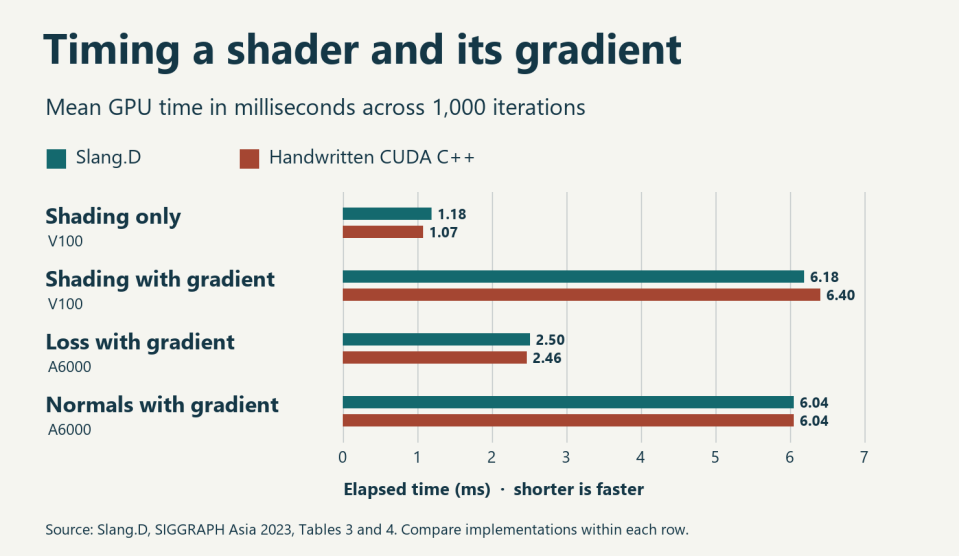


Figure 1. Mean GPU time over 1,000 iterations for a launch size of 8 × 1,024 × 1,024. Each pair compares two implementations on the same GPU. Shorter bars indicate less time. Enlarge the figure · Source measurements.

Elapsed time in milliseconds. The rows labelled “with gradient” include the forward and backward passes.

Operation	GPU	Slang.D	CUDA C++
Shading only	V100	1.18	1.07
Shading with gradient	V100	6.18	6.40
Loss with gradient	A6000	2.50	2.46
Normals with gradient	A6000	6.04	6.04

These averages show that the automatically differentiated operations ran close to the handwritten versions in the four cases shown. The tables give no uncertainty intervals for the small differences. A separate project timed a full training run using a renderer that represents a scene with many small, overlapping 3D Gaussians. After 30,000 iterations on the *Bicycle* scene and an RTX 4090, the Slang version took **36 min 37 s** against **33 min 10 s** for the CUDA reference, or 10.4% more elapsed time. Image fidelity was close at 25.12 and 25.10 dB respectively on the peak signal-to-noise ratio scale. The project reports this one scene and describes itself as unofficial and unsupported by Google. Scene, hardware and results.

Why Dimforge chose Rust for Nexus

Dimforge had mathematical routines it trusted on the CPU, including a matrix decomposition used in simulation. Writing the GPU side of Nexus in Rust let the team reuse those routines and share data structures across the two processors. That reduced the chance of a value having one layout in CPU memory and another in the GPU program. The team could distribute common components through Cargo and run GPU calculations on the CPU while debugging. Its account of trying Slang describes friction integrating with a C++ software stack and dissatisfaction with the development and package-sharing tools available to it. Dimforge’s account of the choice.

In its October 2026 report, Dimforge confirmed that all Nexus GPU shaders were still written in Rust. It described several ways to compile them, including Rust-GPU and cuda-oxide. The reports establish what the team chose and why. A timing comparison between Slang and Rust-GPU on the same neural shader, scene, precision and GPU remains open in the sources reviewed. Nexus follow-up report.

What this language can contribute to artificial intelligence

Slang’s automatic differentiation lets a team update the rule that computes a material’s response to light and generate the associated gradient from that rule. The RTX Neural Shading SDK illustrates both training and inference, including how to use cooperative vectors to accelerate small neural operations on compatible hardware. The SIGGRAPH 2026 course includes a scalar implementation alongside its hardware-accelerated example. These give an application developer a way to measure both the training time spent on derivatives and the frame time spent evaluating the learned material. Language support for derivatives · RTX Neural Shading SDK · SIGGRAPH 2026 examples.

Small neural materials can learn quite different answers from different starting weights. NVIDIA’s 2026 study trained several instances and dropped the weaker ones to increase the chance of finding a good fit. That result concerns the training method. Its tests assess how reliably a network learns reflectance. They leave resistance to malicious inputs and protection of training data unmeasured. Training compact neural shading networks.

Which measurement would settle the choice?

For a renderer, hold the scene and lighting fixed and measure both the time to draw the material and its error against a reference image. Then change the lighting to see whether the learned surface still looks right. For a physics engine, try moving a trusted CPU routine onto the GPU and check that both processors interpret the same data identically. NVIDIA’s materials research and Dimforge’s choice of Rust answer different engineering questions. Each team’s evidence is strongest for the task it actually performed.

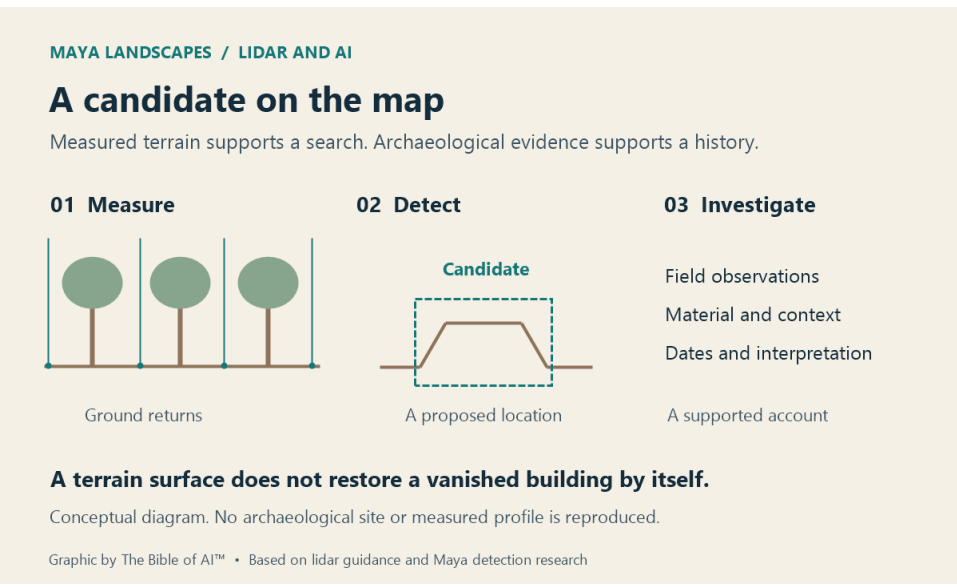
What can AI find in a lidar map of the Maya landscape?

17 de noviembre de 2026

Original publication

On a map of the Maya lowlands, a small rise may be the surviving mass of an ancient building. It may also be a natural formation. Lidar makes the shape of the ground easier to examine beneath vegetation; artificial intelligence can help select shapes that deserve a closer archaeological look.

The distinction matters. A detected mound supplies a place to investigate. Establishing when people built there, what they used the building for and how it related to neighbouring settlements requires further evidence. The promise is a more informed search across a landscape whose human history is only partly visible.



From measured terrain to an archaeological candidate and supporting evidence. Conceptual diagram by The Bible of AI™ based on NOAA lidar guidance and Maya detection research; no archaeological site or measured profile is reproduced.

From laser returns to archaeological candidates

Airborne lidar measures distances with laser pulses. Some reach the ground through openings in the vegetation; others return from leaves and branches. Processing separates likely ground measurements to construct a terrain model. As NOAA's lidar guide explains, dense vegetation can leave too few ground observations. The method depends on openings and sampling, so the resulting surface is not equally reliable everywhere.

AI can then examine images derived from that terrain. A 2024 study in the Journal of Archaeological Science trained an object detector using 15,765 labelled Maya mounds from areas in Guatemala. Combining examples from different areas improved performance compared with training on one area alone. The researchers also found that the chosen way of displaying local relief affected detection: preparing the terrain image was part of the scientific method.

A subsequent Q2000 study, published online in November 2025 and assigned to a 2026 volume, used NASA G-LiHT data from two regions of the Yucatan Peninsula. Its 2,000 selected training examples supported a detector intended to work across varied settings. The flight transects actually examined totalled 421 km²; they sampled a much larger regional extent.

The study reports an F1 score of 0.89, combining precision and recall. Its evaluation used curated, visually assessed test cells and combined counts of individual detections with counts of cells. This is not evidence that 89% of all Maya buildings have been found, and the score should not be ranked directly against results from different evaluation designs. Training and test identifications were not comprehensively confirmed through physical field inspection, and the authors acknowledge possible human identification bias.

When does a terrain model become a reconstruction?

A detector marks candidate locations in the surviving landscape. The terrain surface alone cannot recover the vanished roof of a house, date its construction or establish the activities of its occupants. A convincing reconstruction has to connect measured remains with excavation records, architectural evidence and clearly stated interpretations.

The MayaArch3D project at Copán illustrates a complementary approach. Its research environment connects three-dimensional models with archaeological records and geographic information, allowing architecture and landscape to be examined together. Laser scanning and photogrammetry contribute models of surviving objects and buildings. This project supplies a useful example of linking geometry to evidence; it is separate from the AI detectors described above.

For a reader exploring a reconstructed settlement, the key question is what supports each visible element. A measured platform, a proposed building outline and a reconstructed upper storey carry different kinds of evidence. Showing those distinctions makes a three-dimensional view useful for inquiry, because a change in interpretation can be traced back to the observations behind it.

Lidar and AI can help archaeologists identify where to look and compare patterns across difficult terrain. The larger historical task is to establish which of those patterns belong to human activity, how they developed over time, and what remains uncertain when the past is made visible again.

When AI speeds up customer support, who benefits?

18 de noviembre de 2026

Original publication

Customer support and contact centres

A faster answer can help a customer and the person serving them. Whether AI also improves jobs depends on what happens to pay, workload, recruitment and the quality of the service.

Evidence reviewed: 18 November 2026

A resolved enquiry is more useful than a quickly closed conversation. That distinction matters when assessing AI in customer support: the worker must still establish what the customer needs, recognise exceptions and know when to seek help. The sector's scale provides the starting point; its working conditions determine what an improvement means.

The sector in context

The work

Resolving enquiries, handling complaints and maintaining trust

Customer support combines communication, product knowledge, record-keeping and escalation. Contact centres organise this work across telephone and digital channels. Staff may work for the organisation whose customers they serve or for a contracted provider. [1]

Where the boundary falls

The industry figures describe units classified as call-centre businesses. Customer-support staff in other industries also matter, but are not all included in those totals. The workplace study below concerns a support operation, rather than a representative sample of this statistical industry.

What should change for people?

Measure reliable resolution, paid time, manageable workloads, discretion and opportunities to learn. A greater number of interactions is only one part of that assessment.

European Union · 27 countries · 2024

Activities of call centres

Eurostat · NACE Rev. 2, group 82.2

Persons employed
500,000 estimated

Enterprises
11,014

Net turnover
€30 billion estimated

Annual structural business statistics, including employees and self-employed persons. Employment and turnover carry Eurostat's estimate flag. These are industry totals, not counts of jobs exposed to AI. [2]

United States · Private sector · 2025

Telephone call centers

BLS · NAICS 56142 · Annual averages

Covered payroll jobs
327,577

Establishments
14,961

Average annual pay
US$58,411

QCEW administrative records. Pay is the mean across covered jobs in this industry, including managerial and other occupations; it is not a median agent salary. [3]

Read these as separate baselines. The years, classifications and statistical units differ. An enterprise may contain several establishments; payroll jobs are not a count of unique people. The US category includes answering services and other contact centres, including units serving other establishments within the same enterprise. No EU–US employment ratio or wage comparison is inferred. [2][3][4]

The broader US occupation of customer service representative illustrates the difference: the BLS records 2,666,000 jobs across industries in 2025. Its 2025–2035 projection anticipates a 5% decline, alongside about 289,500 annual openings from occupational transfers and labour-force exits. These projected openings are not net job creation, and the decline is not a measured effect of generative AI. [1]

What assistance has demonstrably changed

Generative AI at Work examines 5,172 agents serving US business customers through a software company and its subcontractors. Of those agents, 89% worked outside the United States, mainly in the Philippines. The main rollout occurred in autumn 2020 and winter 2021; publication followed in 2025. Agents could accept or reject the suggestions. [5]

The quasi-experimental analysis of staggered, non-random access estimates approximately 15% more issues resolved per hour. Benefits were greater among less experienced and initially lower-performing workers; some highly skilled agents experienced small quality reductions. The study lacked individual pay data and did not estimate the sector's long-term employment response. [5]

This provides evidence of a contextual productivity benefit, subject to the study's comparison and assumptions. It does not show what happens when an autonomous system replaces agents. Nor can that percentage be converted into a corresponding staffing reduction: service demand, quality requirements, checking work and the organisation of shifts can all change.

Employment signals need an explanation

Spain supplies a reason to investigate employment rather than assume its stability. CEX's public account of its 2025 market study reports 65,859 workers in Spain, down 6.7%, and 12,375 offshore workers serving the Spanish market, up 12.5%. Participating companies represent around 75% of sector revenue, not necessarily employment. [6]

Participation is voluntary, the sample changed, and overseas activity was redefined. The full report was not examined here. These figures cannot isolate an AI effect or be joined mechanically to older editions. [6][7] A

stronger assessment would follow comparable workplaces, technology adoption, service demand and contract movements, distinguishing fewer recruits from redundancies, reduced hours and relocation.

The ILO's warning concerns this work

The joint UN–ILO report *Mind the AI Divide* explicitly discusses contact and call-centre work. It identifies business-process outsourcing as a source of formal, relatively well-paid employment in countries including India and the Philippines, particularly for women, while warning that technological advances may threaten some jobs. Exposure is not itself displacement. [8]

Its illustrative framework places routine call-centre operations among activities susceptible to substitution and associates that category with reskilling and social protection. The framework assumes smooth technological integration and acknowledges real infrastructure and cost constraints. This is an institutional assessment of risks and transition needs, not an estimate of redundancies in the EU or United States. [8]

Who should receive the gain?

Additional capacity could shorten queues, improve difficult cases, reduce working time or support fewer staff. Those are possible organisational responses, not measured shares. For customers, the test is dependable resolution and access to effective escalation. For workers, it includes pay, breaks, intensity, discretion and time to check an answer. For new entrants, it includes recruitment and supervised learning.

The Bible of AI (BAI) considers those outcomes part of the evidence required to describe technological change as progress in working life. A commercial benefit is worth measuring; it does not establish that benefits are shared. Equally, a reduction in repetitive work deserves recognition when it demonstrably improves people's time, security or ability to help customers.

The present evidence supports a measured judgement: assistance can improve performance, while the distribution of benefits and the employment consequences remain to be established in each setting. A credible claim of success must follow what happens after the faster answer.

Scope and method

Focused narrative analysis; sources checked on 18 November 2026. Industry statistics, occupational projections, workplace estimates and institutional assessments are kept distinct. The workplace results have not been independently reproduced, and no interviews were conducted. The workplace study's commercial setting and acknowledged Stanford Digital Economy Lab support should be considered when assessing the evidence. [5]

References

Numbered references. All online sources accessed 18 November 2026.

1. U.S. Bureau of Labor Statistics. *Customer Service Representatives. Occupational Outlook Handbook* [online]. Updated 29 October 2026. Occupation SOC 43-4051; employment 2025, projections 2025–2035. Available from: BLS occupational profile.
2. Eurostat. *Enterprises by detailed NACE Rev. 2 activity and special aggregates* [dataset]. Updated 30 September 2026. EU27_2020; N822; 2024; indicators ENT_NR, EMP_NR and NETTUR_MEUR. DOI: 10.2908/SBS_OVW_ACT. Exact data selection; SBS methodology. Flag *e*: estimated.
3. U.S. Bureau of Labor Statistics. *Quarterly Census of Employment and Wages: 2025 annual averages, Telephone call centers* [dataset]. NAICS 56142; US000; private ownership; all establishment sizes. Available from: industry data. Definitions: annual-average fields, ownership codes and coverage and statistical units. Employment and establishment counts are annual averages.
4. Executive Office of the President, Office of Management and Budget. *North American Industry Classification System: United States, 2022* [online]. 2022. Section 56142, including 561421 and 561422. Available from: U.S. Census Bureau, official manual.
5. *Generative AI at Work*. *The Quarterly Journal of Economics* [online]. 2025;140(2):889–942. Published online 4 February 2025. DOI: 10.1093/qje/qjae044. Setting: sections III.B–IV.C. Table II, column 3: 0.301 additional resolutions per hour, standard error 0.0329. Interpretation and limits: results and conclusion.
6. Asociación CEX. *Tecnología y deslocalización redefinen el mapa del sector según el Estudio de Mercado 2025 de CEX* [online]. 27 July 2026. Available from: CEX's public account. Methodology and employment sections; reported year-on-year changes.
7. Asociación CEX. *Estudio de Mercado CEX 2025* [online]. Undated public catalogue entry. Available from: publication and participation information. The complete report was not examined.
8. United Nations, Office of the Secretary-General's Envoy on Technology; International Labour Organization. *Mind the AI Divide: Shaping a Global Perspective on the Future of Work* [report]. UN/ILO; 2024. ISBN 9789211066524. DOI: 10.18356/9789211066524. Call-centre employment: p. 7; illustrative framework and transition measures: pp. 15–16. Official publication record.

An ECG biomarker for sudden cardiac death discovered with deep learning

17 de noviembre de 2026

Original publication

About this critique

A risk score can rank electrocardiograms long before anyone can say what the model sees in a heartbeat. In this study the authors follow the score back to a particular shape in lead aVL. They also ask a harder clinical question. Among people whose ECG raises concern, who might live longer if offered a defibrillator?

This assessment examines the version of record and its supplementary analyses. We checked the reported denominators, operating characteristics and mortality comparison against the published tables. The patient records and trained models were not independently analysed.

Critique

Examined version	Peer-reviewed article, *Nature* 655, 210–218. Published online on 24 June 2026.
Materials examined	Version of record and Supplementary Information, particularly Tables I.A.1, III.B.1 and V.G.1.
Recommendation	Read Figure 1 alongside the sensitivity in Supplementary Table I.A.1, then follow the proposed waveform marker through Figures 4 and 5.

Unit 1 — What does a high-risk result predict?

Article data. The Swedish model was trained to predict an out-of-hospital cardiac death recorded on a death certificate within one year after an ECG. Patients were allocated to separate development and test sets. The principal analysis comprised 113,072 ECGs from 35,417 patients under 80 without a defibrillator. Some patients contributed several recordings. The area under the receiver operating characteristic curve was 0.872 (95% confidence interval 0.843–0.899).

Reading. At the authors' preferred threshold, 2.2% of recordings were flagged. The event rate was 7.0% in that group, against 0.58% across the test sample. Yet sensitivity was 26.5%: nearly three quarters of the ECGs followed by a qualifying death fell below the threshold. A death certificate also leaves uncertainty about the rhythm at the time of death.

Critical conclusion. The score identifies a small group at substantially elevated risk. A low score at this threshold offers limited grounds for ruling out a future event, and multiple ECGs from one patient must not be counted as separate lives.

Claim type	Risk stratification and ascertainment of the outcome
Exposure	High. Confusing the event rate among flagged recordings with the fraction of events detected would overstate the reach of screening.

Unit 2 — What travelled across the three health systems?

Article data. The US validation included 251,858 ECGs from 139,613 patients. The outcome was a recorded episode of ventricular fibrillation or ventricular tachycardia. The unchanged model achieved an area under the curve of 0.822 for that outcome, compared with 0.717 for the corresponding outcome in Sweden. In Taiwan the model distinguished adjudicated arrhythmic arrests from emergency department controls, with an area under the curve of 0.767.

Reading. The US records lacked linked causes of death. In Taiwan the investigators selected cases and controls, so the proportion of cases in the study could not establish an annual risk in the population. Even the definition of a positive screen matters. Applying the Swedish score cut-off in the US flagged 3.4% of recordings. Selecting the highest 2.2% locally produced a different group.

Critical conclusion. Evidence from three settings supports an association between the score and serious cardiac events. The calibration of an annual sudden-death probability in a new clinical population remains a separate question.

Claim type	External validation across outcomes and sampling designs
Exposure	High. An estimate obtained from one outcome or sampling design could be used to counsel patients about another.

Unit 3 — Which part of the heartbeat carries the signal?

Article data. A generative model altered the shape of individual beats in the direction that increased a predictive model's score. In lead aVL, the downward slope at the end of the R wave became less abrupt. The authors then quantified the shape in observed ECGs using the mean absolute first and second differences in voltage from the R-wave peak to the end of the QRS complex.

Reading. Figures 4 and 5 show how a model-derived contrast led to a measurement that another investigator could examine. In Taiwan, both waveform measures were associated with arrhythmic arrest in unadjusted analyses. After adjustment for conventional ECG features, both estimates remained negative, but neither was statistically significant. For sudden cardiac death in Sweden, only the first-difference measure retained a significant association after adjustment. Both measures retained significant associations with ventricular fibrillation or tachycardia in Sweden and the US. The paired traces in Figure 5 illustrate a generated change to a beat. They do not document a patient's heart progressing from one state to the other.

Critical conclusion. The work offers a testable waveform marker. Its relation to conduction through diseased myocardium merits direct physiological study, and the performance of the full model cannot be attributed to this one feature alone.

Claim type	Interpretation of the waveform and its possible mechanism
Exposure	Moderate. A visible and measurable feature can acquire a more definite physiological interpretation than the observations support.

Unit 4 — Would acting on the score prevent deaths?

Article data. Only 13.9% of the high-risk ECG group were known to have a left ventricular ejection fraction of 35% or less. Ejection fraction had been measured in 22.5% of the analysis sample. In a separate observational comparison, the estimated interaction between a high-risk ECG and an existing defibrillator was −3.62 percentage points for sudden cardiac death. Relative to the authors' predicted rate of 6.65%, this gives the reported 54.4%.

Reading. An unmeasured ejection fraction cannot be classified as normal. Patients who already had a defibrillator had been selected for implantation in clinical practice. Their underlying disease, care and eligibility could differ from those of newly identified patients. The regression provides a numerical comparison, but treatment allocation leaves the counterfactual outcome uncertain.

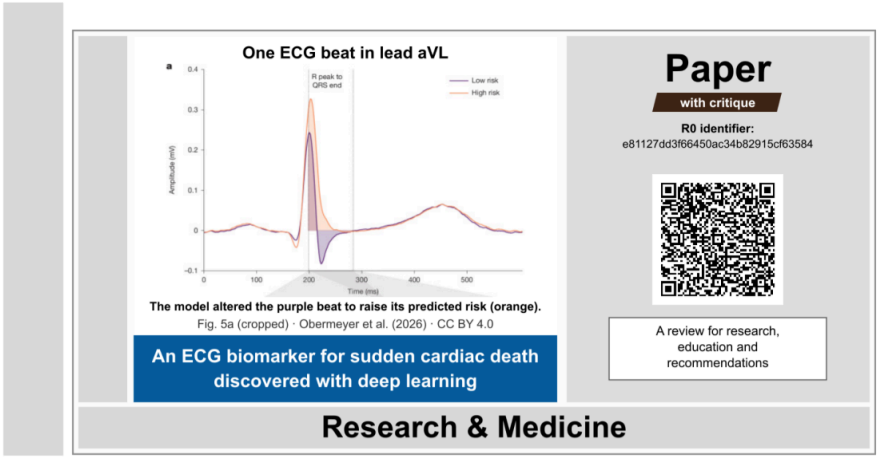
Critical conclusion. The study makes a case for a prospective trial of ECG-guided selection. Demonstrating clinical benefit would require comparing outcomes under the proposed strategy and usual care, including all-cause mortality and device complications.

Claim type	Added clinical information and benefit from treatment
Exposure	High. Treating the reported 54.4% association as the benefit of a new implantation rule could affect an invasive decision.

Overall assessment

The paper's most distinctive contribution is the passage from a prediction to a waveform measurement. A reader can follow the transformation in Figure 4, identify the interval being measured in Figure 5 and consult the regressions on observed ECGs. That sequence makes a specific question available for independent testing: does this part of the QRS complex retain its predictive value across instruments, populations and definitions of the outcome?

The choice of a clinical threshold needs its own evidence. The Swedish results quantify both how much risk was concentrated and how much lay outside the selected group. The external cohorts broaden the evidence for discrimination. A trial would answer the question that remains for a patient offered an implant on the strength of a routine ECG: how does that decision change the chance of surviving, and at what cost in complications?



Full-resolution cover (PNG). Lead aVL of a single heartbeat. The purple trace shows the original beat. The orange trace shows a model-generated version of that beat with higher predicted risk. The box marks the interval from the R-wave peak to the end of the QRS complex. Cropped from Figure 5a of *An ECG biomarker for sudden cardiac death discovered with deep learning*, by Obermeyer et al., *Nature* 655, 210–218 (2026). Source article, CC BY 4.0.

Paper page:
https://www.nature.com/articles/s41586-026-10674-6

Abstract

Sudden cardiac death is, in theory, preventable with defibrillators. But every year, many patients die without defibrillators because doctors fail to predict their risk[1]. The only predictive biomarker in wide use, cardiac left ventricular ejection fraction (LVEF), misses most sudden cardiac deaths[2], and flags many low-risk patients for futile defibrillators that never fire[3,4]. Here we apply deep learning to a dataset linking all electrocardiograms (ECGs) in a Swedish region to death certificates. The resulting model isolates a high-risk group (2.2% of the sample) with a 7.0% annual rate of sudden cardiac death, higher than those with reduced LVEF (1.9% of the sample; 4.6% annual rate). Notably, 86.1% of the model's high-risk patients were not flagged by LVEF. High-risk ECG patients with defibrillators implanted were 54.4% less likely to die than expected, suggesting a mortality benefit. We externally validate the model in a US health system, in which it predicts ventricular arrhythmias that cause sudden death; and a Taiwanese hospital registry, in which it specifically predicts future arrhythmic cardiac arrests. To visualize the waveform morphology 'discovered' by the predictive model, we pair it with a generative model of the ECG waveform. Together, they reveal a biomarker that is easily visible and robustly predicts sudden cardiac death, but has not to our knowledge been previously described. Tying the biomarker's shape to electrophysiological first principles, we form and preliminarily test a new hypothesis on the mechanism of sudden cardiac death.

Authors' abstract reproduced from Nature under the Creative Commons Attribution 4.0 International licence. Reference numbers follow the original article.

Authors

Ziad Obermeyer, Alexander Schubert, James Ross, Sendhil Mullainathan, Markus Lingman

LLM-Based Agents for Cybersecurity: A Systematic Review of Architectures, Applications, and Open Challenges

10 de noviembre de 2026

Original publication

About this critique

This assessment examines the Preprints.org version posted on 24 August 2026. It is an editorial critique, not peer review or an independent replication of the underlying agent experiments. Selected counts and retrieval percentages have been recalculated from the deposited review records. This verifies their arithmetic and consistency, not every inclusion decision, coding judgement or experimental result.

The assessment distinguishes limitations acknowledged by the authors from inconsistencies identified here. Exposure denotes the potential consequences of overstating a particular inference; it is not an overall rating of study quality. A subsequent withdrawal affecting one cited study is considered with its chronology.

The recommendation is to read the paper critically, together with the qualifications below.

Critique

Audited version	Preprints.org v1 · DOI 10.20944/preprints202608.1600.v1 · posted 24 August 2026 · CC BY 4.0 · this version is not peer-reviewed.
Incorporated material	Complete preprint text, figures and classification tables; the deposited protocol, coding manual, study classification, screening and recovery records; and primary-source checks of the architectural examples discussed below. Consulted 10 November 2026.
Version distinction	The repository reports a subsequent publication in the Journal of Cybersecurity and Privacy, DOI 10.3390/jcp6050159. This assessment concerns the identified preprint and does not imply that the journal version has been separately reviewed.

Unit 1 — A pivotal incident-response result has subsequently been withdrawn

Article data. Sections 4.6, 5.1 and 7.2 use MyAntFarm.ai to support a strong claim about the importance of orchestration: 100% actionable recommendations, compared with 1.7% for a single-agent configuration, across 348 trials. The review was posted on 24 August 2026. On 31 August 2026, the author of the cited study withdrew its Decision Quality results, including the associated actionability and zero-variance claims. The withdrawal notice explains that the evaluated action list had been fixed in the implementation rather than generated by the model; the scoring procedure assessed that list. The notice states that latency measurements are unaffected.

Reading. The withdrawal occurred after the selected preprint was posted. It therefore changes the evidence available to readers today without establishing that the review's authors could have known about it when this version appeared. The affected measurements cannot now support the claim that orchestration determines incident-response quality more strongly than model capability.

Critical conclusion. The MyAntFarm.ai quality results should be removed from the review's current evidential support. This correction does not invalidate the taxonomy or the results of other multi-agent studies; it specifically weakens the review's strongest incident-response example and the broader inference drawn from it.

Claim type	Subsequent correction to the evidence base
Exposure	High. A withdrawn result is used repeatedly to support a central architectural conclusion.

Unit 2 — Architectural prevalence does not establish general superiority

Article data. Recalculation from the deposited classification confirms 18 single-agent tool-calling studies and 22 multi-agent studies: 11 centralised or hierarchical systems, eight pipeline systems and three debate systems. Centralised or hierarchical systems therefore constitute 18.6% of the 59-study corpus. Section 3.4 nevertheless describes this category as having the smallest evidence base within the multi-agent family. The review also highlights an improvement of up to 4.3-fold for HPTSA; the primary study reports this comparison on a benchmark of 14 vulnerabilities.

Reading. The classification makes centralised or hierarchical systems the largest of the three multi-agent categories. A claim about a smaller subset of controlled comparative experiments would require that subset to be defined. The HPTSA comparison remains relevant to its evaluated task, but its maximum relative improvement is not an average effect across architectures, security functions or deployment settings. The review itself acknowledges benchmark heterogeneity and does not pool effect sizes.

Critical conclusion. The prevalence statement requires correction. The comparative findings support further controlled evaluation of particular orchestration designs, while a general ranking of architectures remains unestablished. Architectural frequency and the number of convincing comparative evaluations should be reported separately.

Claim type	Numerical consistency and comparative inference
Exposure	Medium. The error distorts the distribution of evidence; generalising selected performance maxima would overstate the comparative conclusion.

Unit 3 — A documented subset cannot establish the absence of production evidence throughout the field

Article data. The review classifies 59 studies, with 30 assigned primarily to penetration testing (50.8%) and none assigned to production evaluation, E4. Its verification search recovers 44 of the 51 studies found through the primary search (86.3%), or 47 of all 59 included studies (79.7%). The authors also estimate that the eligible literature is substantially larger. Section 8 expressly limits frequency and absence claims to the selected corpus, whereas the abstract and Section 3.4 extend the absence of E4 studies to the entire field.

Reading. Recovery of previously included studies measures overlap between retrieval procedures. It does not establish the proportion of all eligible studies captured by the original search. The authors' explicit acknowledgement that the corpus is a documented subset is therefore essential to interpreting both the frequencies and the absence of E4 evaluations.

Critical conclusion. The supported statement is that no E4 evaluation was identified among these 59 coded studies. A field-wide claim requires broader evidence than this corpus supplies. The qualification already present in the limitations should govern the abstract and the corresponding interpretative passages as well.

Claim type	Retrieval coverage and generalisation
Exposure	High. A corpus-level absence could otherwise be mistaken for evidence that production evaluation does not exist anywhere in the field.

Unit 4 — Disclosure of screening limitations does not restore missing exclusion records

Article data. The deposited screening file contains 211 records and records 30 full-text exclusions. For 29 of those exclusions, the individual study identities and reasons were not retained; the corresponding rows explicitly document this absence. The manuscript and archive acknowledge single-reviewer screening, the absence of a prospective registration and the lack of a formal risk-of-bias assessment. They also disclose that AI-assisted exclusion codings were adopted without record-by-record human re-verification.

Reading. These disclosures are substantive strengths of the reporting. They nevertheless distinguish what can be counted from what can be independently reassessed. Placeholder rows allow the totals to reconcile, but cannot establish whether each unidentified study satisfied an exclusion criterion. Similarly, the E1–E4 classification describes evaluation settings and does not assess all sources of bias in study conduct or measurement.

Critical conclusion. The archive supports an audit of the available classification and much of the selection record, but not a complete reconstruction of full-text eligibility decisions. This acknowledged limitation constrains confidence in selection completeness; it is not evidence that the retained studies were necessarily selected incorrectly.

Claim type	Selection reliability and auditability
Exposure	Medium. Numerical reconciliation could be confused with a fully recoverable and independently verified selection process.

Unit 5 — The supplementary files require reconciliation across corpus revisions

Article data. The classification file contains 59 studies. The recovery file contains 60, including PentestEval, which the manuscript and screening decisions exclude from the primary corpus. The verification-search documentation still reports 48 of 60 studies recovered and 45 of 52 primary-search studies recovered; its protocol also retains wording that treats PentestEval as included. Restricting the recovery file to the 59 studies actually classified gives 47 of 59 and 44 of 51, reproducing the manuscript's 79.7% and 86.3%. These checks use the deposited review materials consulted on 10 November 2026.

Reading. The discrepancy is a traceable inconsistency between versions of the supplementary records. It does not refute the manuscript's two recovery percentages, which can be reconstructed after the excluded record is removed. It does mean that readers cannot apply every supplied denominator or procedural description without first resolving which corpus version it describes.

Critical conclusion. The supplementary recovery file, protocol and explanatory text should be aligned with the final 59-study classification. Reproducibility requires consistent records as well as correct calculations. The revised percentages should remain explicitly tied to their distinct denominators.

Claim type	Data consistency and reproducibility
Exposure	Medium. Unreconciled supplementary records can produce different, individually plausible recovery estimates.

Unit 6 — Evaluated autonomy is not evidence that human oversight is unnecessary

Article data. The deposited classifications yield 52 fully automated systems, six human-guided systems and one system requiring human approval. None is classified as human-overseen. Among the 11 studies primarily concerned with DevSecOps, intrusion detection or incident response, ten are classified as fully automated. The coding manual concerns the evaluated configuration, rather than oversight proposed for a future deployment. Section 5.4 recognises that laboratory settings may explain the observed pattern.

Reading. A study can evaluate an automated component while leaving deployment decisions to people. Conversely, a proposed approval mechanism may not have been exercised in its experiment. The frequency of fully automated experimental configurations therefore describes the research designs in this corpus; it does not measure the comparative safety or operational adequacy of autonomous deployment.

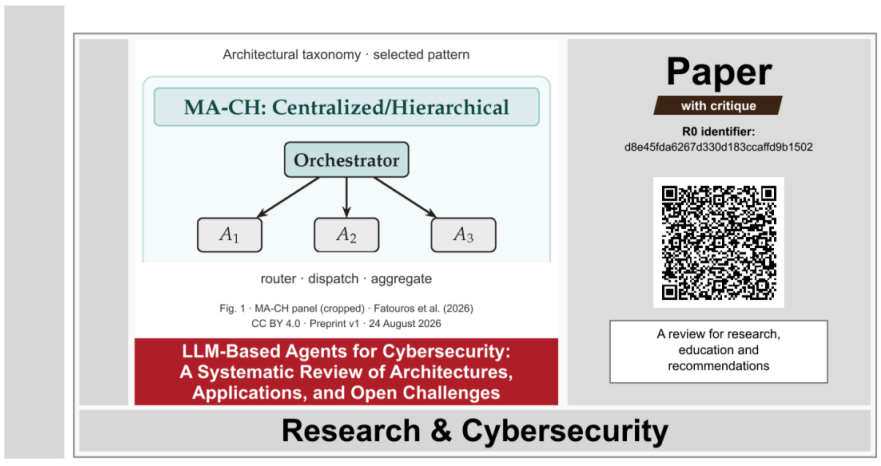
Critical conclusion. The oversight findings identify an evaluation gap. They support testing approval mechanisms, intervention procedures and their effects under realistic conditions, rather than treating experimental autonomy as evidence that human oversight can be dispensed with.

Claim type	Measurement scope and proposed application
Exposure	High. Confusing experimental configuration with demonstrated deployment suitability could distort decisions about consequential security actions.

Overall assessment

The review offers a useful five-dimensional taxonomy and a publicly inspectable classification of 59 studies. Its principal descriptive counts can be reproduced from the deposited data. Its value as a structured account of selected research should nevertheless be distinguished from exhaustive coverage, a general ranking of architectures or evidence of production readiness. The later withdrawal of the MyAntFarm.ai quality results materially weakens one prominent argument; inconsistent supplementary records and the incomplete exclusion history further limit reconstruction of the synthesis. These qualifications preserve the contribution of the taxonomy while identifying where its interpretation and supporting records require correction.

The paper is recommended as a resource for studying how cybersecurity agents are classified and evaluated, with particular attention to the strength, currency and boundaries of the evidence supporting each conclusion.



Selected MA-CH architectural pattern from Figure 1 in Fatouros et al. (2026), Preprints.org v1, posted 24 August 2026, reproduced under CC BY 4.0. The panel has been cropped, and its functional descriptor has been reset for legibility.

Paper page:
https://doi.org/10.20944/preprints202608.1600.v1

Abstract

The rapid evolution of Large Language Models (LLMs) has opened new frontiers in cybersecurity automation, enabling intelligent agents capable of multi-step reasoning, tool invocation, and autonomous decision-making across complex security tasks. While individual applications have emerged across threat intelligence, vulnerability assessment, penetration testing, and security operations center (SOC) automation, a systematic understanding of the LLM-based agent paradigm in cybersecurity — encompassing both single-agent and multi-agent architectures — remains lacking. This paper presents a systematic literature review following PRISMA guidelines, identifying records through 59 structured web-search queries whose results resolve predominantly to arXiv, Semantic Scholar, the ACM Digital Library, IEEE Xplore, USENIX, MDPI, SpringerLink, and Elsevier ScienceDirect, supplemented by citation chaining, for works published between January 2022 and April 2026; the full query record is published with the paper. We applied structured inclusion and exclusion criteria and classified 59 primary studies along five dimensions: security function, agent architecture pattern, knowledge augmentation strategy, human-in-the-loop posture, and evaluation rigor. Our analysis reveals that penetration testing and threat intelligence are the most extensively studied domains, while incident response and compliance verification remain critically underrepresented. Penetration testing alone accounts for over half the corpus (50.8%). Single-agent tool-calling remains the most prevalent architecture (30.5% of studies), whereas centralized multi-agent orchestration — present in 18.6% — yields the strongest reported performance gains, up to 4.3x on zero-day exploitation; prevalence and performance therefore point in opposite directions. No included study achieves production-grade (E4) evaluation: the entire field currently rests on controlled laboratory assessments. An independent search of six bibliographic databases recovers 86.3% of the studies the primary search had surfaced (79.7% of the full corpus) while indicating a total eligible literature of roughly 400 studies, so the corpus is reported as a documented subset rather than an exhaustive census. We propose a unifying taxonomy, identify cross-cutting challenges including hallucination, prompt injection, and benchmark fragmentation, and outline open research directions with particular emphasis on multi-agent orchestration design. Financial sector applicability under DORA and the EU AI Act is treated as a documented evidence gap rather than a synthesis: the corpus's only compliance and risk assessment study is also its only banking-specific system.

Abstract reproduced from the authors' Preprints.org v1 under the Creative Commons Attribution 4.0 International licence. Percentage signs and dash notation have been normalised for presentation; the wording is otherwise retained. The editorial critique above qualifies the abstract's interpretation of the evidence. The subsequent withdrawal discussed in Unit 1 concerns a cited study, not this review.

Authors

George Fatouros, Konstantinos Mavrogiorgos, George Makridis, John Soldatos, Dimosthenis Kyriazis

Autonomous robotic operation controlled by wave-based neuromorphic hardware

17 de noviembre de 2026

Original publication

About this critique

What made the car turn? The answer begins with a distance sensor on the vehicle, passes through a dish of water on a laboratory bench and ends with a command sent back to the motor controller. Zohar and colleagues make the intermediate computation visible. Their experiment also illustrates what an independent investigator would need to account for one particular turn.

This assessment examines the published article, its supplementary methods, the physical vehicle recording and the authors' archived software release. We checked the deposited input tables and aggregate classification results. The physical experiment and micromagnetic simulations were not repeated. Exposure describes the consequence of extending a finding beyond the conditions in which it was established.

Critique

Examined version	Peer-reviewed article, *Nature Communications* 17, 10677 (10 September 2026).
Materials examined	Published article, supplementary methods and Video 2, and version 1.0.0 of the accompanying code and data.
Recommendation	Read Figure 3 alongside Figure 4c. The first measures recognition of labelled inputs. The second shows the behaviour of a vehicle on a course.

Unit 1 — How does a distance measurement become a manoeuvre?

Article data. A single ultrasonic sensor on the physical car scans six directions. A host computer receives the distances over Bluetooth and uses them to time six mechanical excitations of a water surface. A camera records the interference pattern. After subtraction of a reference image and spatial averaging, a trained support vector machine classifies the scene. In the physical driving experiment the choice was binary. A nearby obstacle prompted a turn of approximately 30 degrees.

Reading. The wave field provides a measurable intermediate state. Its formation can be understood from the timing of the inputs and the interaction of the waves. Assigning meaning to that state requires the optical reference, the sampling procedure and the fitted classifier. A photograph of the waves alone cannot establish which reading changed the decision.

Critical conclusion. The sensing, physical transformation, classification and motor response are individually open to investigation. Explaining a particular turn requires linking observations from all four stages to the same instant.

Claim type	Interpretability of an individual control decision
Exposure	High. An error rate for labelled scenes has a different denominator from a collision rate per journey.

Unit 2 — Which scenes did the classification test cover?

Article data. The study tests tasks with two, three and five output classes. The authors sampled a single water-wave image on a 50 × 50 grid, 150 ms after input excitation, with adjacent distance levels encoded 200 ms apart. For the learning curves, they report 100 test samples per evaluation and 50 repetitions with random training and test sets. Their deposited water-wave results give mean recognition rates of 97.20%, 93.46% and 87.76% at 250 training examples. All three columns reach 100% at 2,500.

Water-wave obstacle recognition. Mean accuracy in the archived laboratory results, by training-set size and number of output classes.

Training examples	Two classes	Three classes	Five classes
250	97.20%	93.46%	87.76%
2,500	100.00%	100.00%	100.00%

Reading. The five-class input file contains 1,000 distinct six-distance vectors, 200 in each class. In this file, no vector places an obstacle at distance levels 1–4 on both sides at once. The accompanying generator explicitly excludes that case. These labels make it possible to check what the task asks of the classifier. A robot approaching a narrow passage may have to act precisely when both sides are occupied.

Critical conclusion. The learning curves support recognition of the defined classes under the sampling procedure used. The deposited five-class examples leave bilateral obstacles outside their coverage. The recognition percentages cannot be read as the probability of completing a journey without a collision.

Claim type	Classification accuracy and coverage of obstacle configurations
Exposure	High. An error rate for labelled scenes has a different denominator from a collision rate per journey.

Unit 3 — Can a reader reconstruct one turn without rebuilding the apparatus?

Article data. Figure 4c shows a physical car navigating a 3 m × 2 m course. Supplementary Video 2 records the run and was accelerated fiftyfold for presentation. The authors also release input generators, labelled distance vectors, aggregate water-wave scores and arrays of simulated spin-wave features. These materials let a reader check the classification rules and parts of the numerical analysis.

Reading. To reconstruct a particular turn, one would need its six distance readings, the corresponding reference and wave images, the classifier state and the command returned to the vehicle, all matched in time. We could not identify such a record for the filmed run in the archived release or the listed supplements. A trajectory supplies evidence of where the car went. Several combinations of distances and wave patterns could have produced the same turn.

Critical conclusion. A reader without the apparatus can examine the proposed mechanism, check the archived inputs and follow the recorded behaviour. The public record leaves the cause of an individual manoeuvre

underdetermined. This limit concerns retrospective verification of the run. It does not imply that the apparatus is intrinsically opaque.

Claim type	Independent reconstruction of an observed event
Exposure	High. Investigating an unexpected turn requires a record of the signals present at the time.

Unit 4 — Which operating conditions have been demonstrated?

Article data. The physical car stopped periodically to take its readings and communicated with a reservoir off the vehicle. Event-triggered control was demonstrated with the virtual car. A further experiment used micromagnetic simulation to evaluate a one-micrometre magnetic disc. The supplement's energy estimate for spin waves covers input excitation and excludes the readout and peripheral circuitry.

Reading. Each result answers a different engineering question. The physical run establishes that the sensing and command loop can close through a water reservoir. The virtual experiment shows how changes in the sensor signal can trigger computation. The simulation concerns magnetic dynamics and the information available to a readout. The 150 ms image sampling time does not measure the time from an obstacle entering the physical vehicle's path to the response of its wheels.

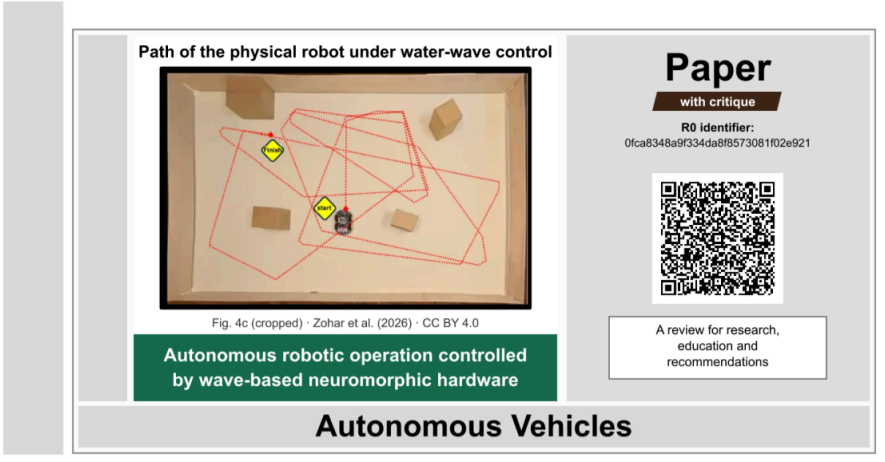
Critical conclusion. A practical vehicle would need measurements of response time, collisions and whole-system energy over repeated journeys under stated operating conditions. The authors explicitly identify fabrication and readout as further steps in assessing a magnetic device.

Claim type	Transfer from a bench experiment to an integrated vehicle
Exposure	High. Both the physical boundary of the energy estimate and the distinction between measured and simulated operation affect deployment decisions.

Overall assessment

The strongest result is the closed control loop. Measured distances altered a physical wave field, a trained readout interpreted an image of that field, and the vehicle responded. By reporting how the image was formed and sampled, the authors give other laboratories specific observations against which to test the mechanism. Their archived data also make it possible to identify scene configurations that a headline recognition rate leaves out.

For a field investigation, the decisive addition would be a synchronised record of each manoeuvre. It would allow an investigator to ask whether a missed obstacle first disappeared during sensing, during wave encoding, at classification or when the command reached the vehicle. That question remains open even when the intended behaviour is visible in a successful run.



Full-resolution cover (PNG). The red dotted line traces the physical car's route around wooden obstacles. The authors reconstructed the route from Supplementary Video 2. Cropped from Figure 4c of *Autonomous robotic operation controlled by wave-based neuromorphic hardware*, by Zohar et al., *Nature Communications* 17, 10677 (2026). Source article, CC BY 4.0.

Paper page:
https://www.nature.com/articles/s41467-026-77661-3

Abstract

Equipping robots with a physical, brain-inspired computing unit is a key step towards achieving true autonomy in intelligent machines; however, the practical realization of such neuromorphic hardware remains a major challenge. Because both software- and hardware-based spiking neural network approaches demand substantial training and energy consumption, an architecturally simple neuromorphic platform is required to enable scalable and energy-efficient robotics. Here, we experimentally demonstrate robotic control driven by a neuromorphic hardware based on physical wave interactions. Using a water-based experimental system that provides direct visual access to wave-network dynamics, we establish a wave-based reservoir computing framework that achieves near-perfect accuracy in robotic vehicle obstacle recognition, followed by real-time autonomous motion with obstacle avoidance controlled entirely by this wave-based hardware. Through micromagnetic simulations, we extend this concept to electrically excited spin waves in a magnetic nanodevice operating at gigahertz frequencies, outlining an approach toward solid-state robotic chips based on wave computing. These results position wave-based computation as a scalable and energy-efficient alternative to conventional digital neuromorphic architectures for next-generation intelligent machines.

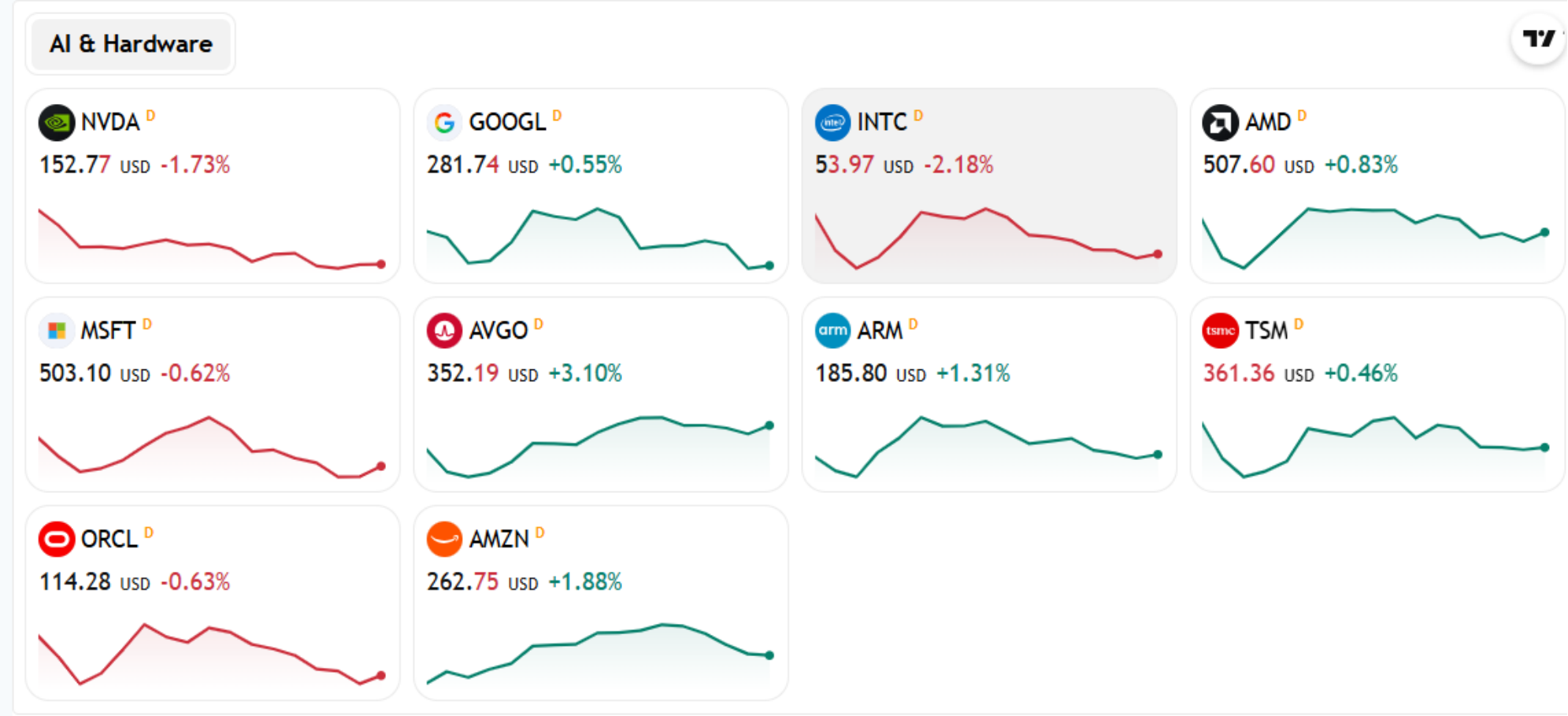
Authors' abstract reproduced from Nature Communications under the Creative Commons Attribution 4.0 International licence.

Authors

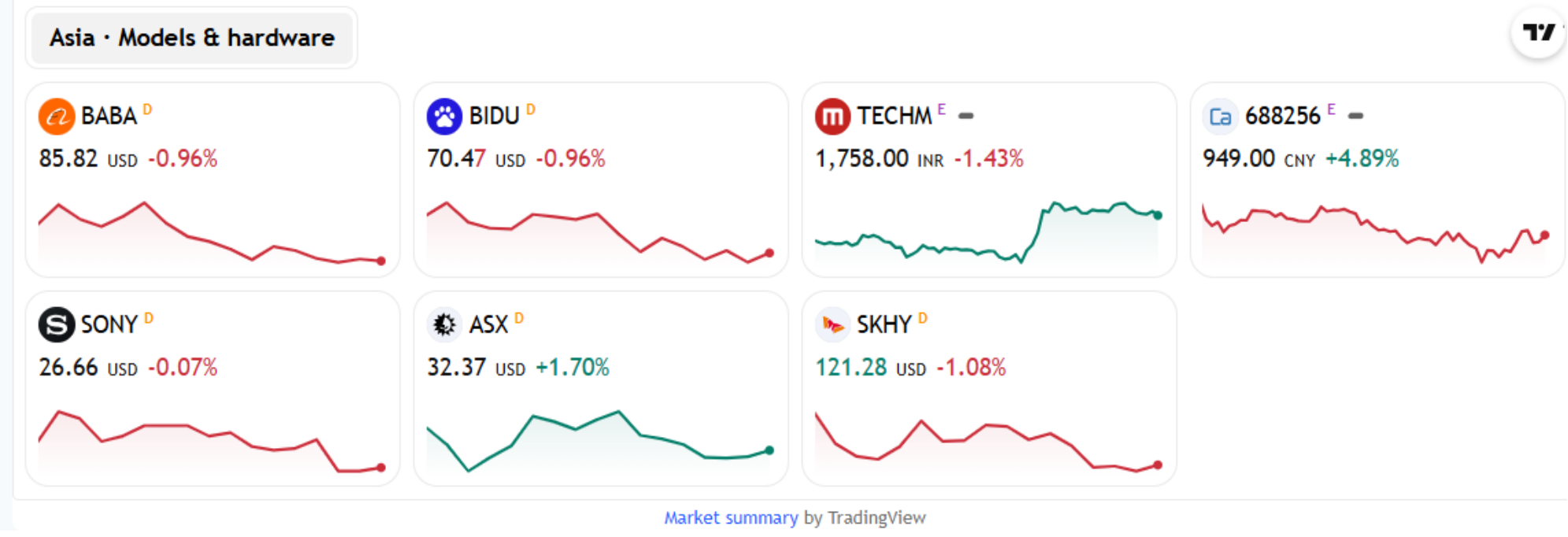
Jacob Zohar, Daniele Pinna, Gerrit van der Laan, Thorsten Hesjedal, C. K. Safeer

AI and hardware markets

Dated snapshots from the front cover - 1 / 2



AI and hardware - US listings
This US panel shows dollar quotations. TSM represents TSMC’s US-listed depositary shares (ADR). These market indicators provide economic context, not a ranking of AI capabilities or investment advice.
Captured 18 Nov 2026, 16:08 UTC
Charts by TradingView



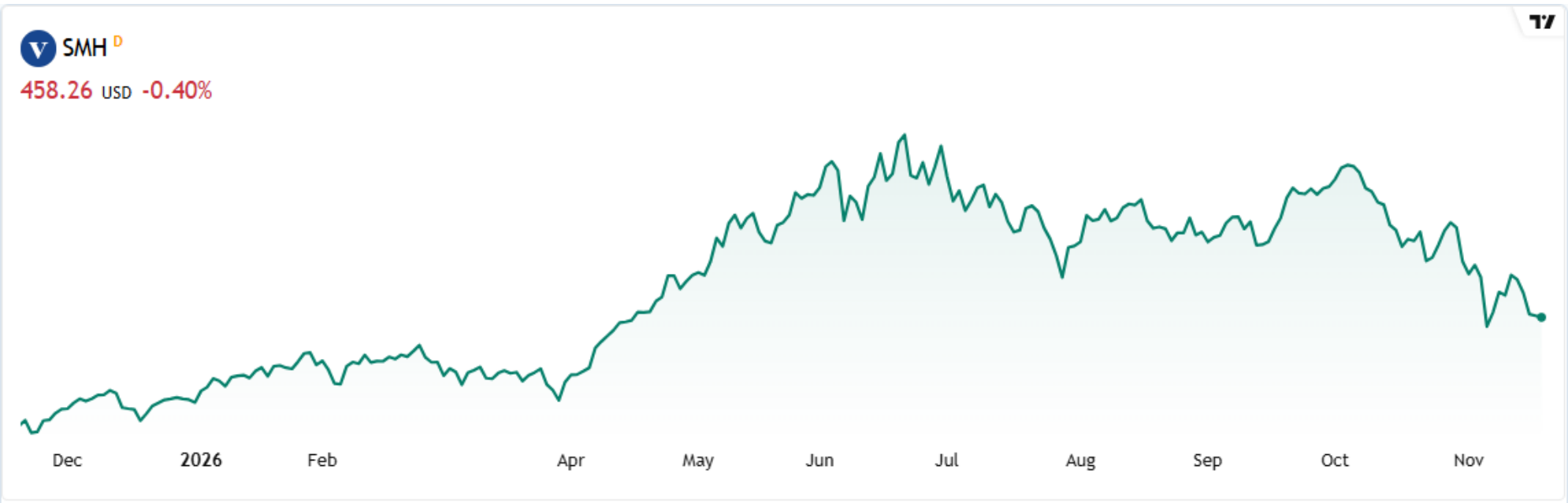
AI markets in Asia
Alibaba and Baidu develop the Qwen and ERNIE models in China. India’s Tech Mahindra works on models for Indian languages. On the hardware side, follow Cambricon’s accelerators, Sony’s vision sensors in Japan, ASE’s chip packaging in Taiwan and SK hynix’s high-bandwidth memory in South Korea. Market data by TradingView. Shanghai (CNY) and Bombay (INR) show end-of-day prices. The US listings are depositary receipts in USD and may be delayed. Each price reflects the entire company.
Captured 18 Nov 2026, 16:08 UTC
Charts by TradingView

Capture time is not the last-trade time. Read the dates, units and market status within each chart.

Provider credits are retained. Charts and data retain their own rights; the journal licence does not replace them.

AI and hardware markets

Dated snapshots from the front cover - 2 / 2



Semiconductor market
VanEck Semiconductor ETF (SMH) brings together US-listed semiconductor and semiconductor-equipment companies, including international firms. An ETF is an exchange-traded fund: its share price offers a view of a basket of companies rather than a single business. This is the US fund, quoted in USD, not its European UCITS counterpart.
Captured 18 Nov 2026, 16:08 UTC
Charts by TradingView



AI infrastructure - bond market
CoreWeave · USD senior unsecured notes · Fixed annual coupon: 9.625% · Maturity: 15 July 2032 · ISIN: USU2069EAE06 (Reg S). Read the measure: YTM is the annualised yield to maturity implied by price and scheduled payments, not the coupon or a guaranteed return. Price is expressed as a percentage of face value.
Captured 18 Nov 2026, 16:11 UTC
Charts by TradingView

Capture time is not the last-trade time. Read the dates, units and market status within each chart.
Provider credits are retained. Charts and data retain their own rights; the journal licence does not replace them.

Who answers for AI? Lessons from the newsroom and the assembly line

14 de noviembre de 2026

Original publication

Juan Antonio Lloret Egea, Director (BackGround) · AI and Society

Imagine opening the local paper and finding your photograph beneath a report of a theft. You were nowhere near it. Now imagine a letter from your car manufacturer: some welds may be missing, and your car needs an inspection. In the first case, you ring the editor. In the second, you ring the dealer. Nobody sends you to the printing press or asks you to take the matter up with the welding robot. Why should "the AI did it" be an acceptable answer?

As an engineer and a director, I want to know how a failure happened and who owes the person affected an answer. Those questions lead to different places. A machine may explain the failure. Responsibility belongs to the people or organisations whose decisions and duties the evidence establishes. A company, a public authority or a political organisation should not be able to make its own part in a decision disappear by calling it the work of AI.

The newsroom: who put that photograph beside your name?

In our imagined newspaper, the reporter gathered the story, someone selected the picture and wrote the caption, and an editor authorised publication. The publisher put its name on the paper. To find the error, the editor needs the original picture, the caption as submitted and the version that went to print. A perfectly genuine photograph can make a false accusation if it is attached to the wrong story. No detector for synthetic images would catch that mistake.

Now suppose software generated the picture and published it without a final human check. We should not invent a reporter who saw an image that no one reviewed. We can still ask who authorised automatic publication, what checks were required, and who could take the story down when you called. If the complaint is upheld, you are owed a correction that reaches the people who saw the accusation. The publisher can issue it in its own name while protecting a confidential source.

The assembly line: who stands behind the car?

When I buy a car, I do not expect anyone to have built it entirely by hand. Nor do I look for a watermark on the door saying "made by robots and AI". I look for the manufacturer's name. If there is a warning about the bodywork, the useful question is whether *my* car is safe to drive. Knowing which tools were used may help an engineer investigate. The car cannot answer for the defect, and the word "robot" gives the driver no one to call.

Take a joint that is meant to hold part of the body together in a crash. A design engineer specifies where the welds belong. A production team translates the drawing into instructions for the line. Quality staff test whether the parts that leave it meet the specification. If a weld is missing, comparing the drawing, the robot's operating record and the inspected part can reveal where the three diverged. Those people have different duties. Finding a defect does not, by itself, establish that any particular engineer was negligent.

In July 2018, a routine inspection at Subaru's Indiana plant found missing spot welds on a model-year 2019 Ascent, around the upright between the front and rear doors. Subaru traced the fault to the welding robots' programming, corrected the program and used the robots' records to identify the vehicles that needed checking. Its US recall covered **293 potentially affected vehicles**. Subaru undertook to replace any car found to lack the welds. In its closing notice, it reported that all 293 passed inspection, including the nine already in customers' hands. Calling these "293 defective cars" would tell a different and incorrect story. Subaru's defect report to the US safety authority and its final inspection notice.

This was a documented programming fault; the report does not identify it as AI. That is why the comparison is useful. The robot supplied evidence. The manufacturer notified customers and arranged inspections. A related principle appears in European law: the manufacturer answers to the approval authority for conformity of production and must take appropriate corrective action when a vehicle does not conform. That duty does not settle who caused a particular collision. EU vehicle approval regulation, Articles 13–14.

What a watermark can tell a reader

There is one important difference between the photograph and the car. A convincing synthetic photograph can look like eyewitness evidence of something that never happened. Telling a reader how it was made matters. But a label saying "made with AI" does not tell the reader who chose to present it as evidence, whom to contact, or whether the publisher will correct the record. I would ask for an identifiable publisher as well as an honest account of how the picture was made.

EU law itself draws a distinction. Article 50 of the AI Act provides for machine-readable marking of certain synthetic outputs and for disclosure when someone uses a system to make a deepfake. Those are distinct obligations. For text generated or manipulated with AI and published on a matter of public interest, the Act provides an exception where there has been human review or editorial control and a person or organisation bears editorial responsibility. That exception for text is not a blanket exemption for the photograph beside it. AI Act, Article 50; European Commission guidance on scope and dates.

Readers also distinguish between a journalist using a tool and a system producing the news. The survey below asked how comfortable people felt in each situation. It did not ask them to assign legal liability.

Who stands behind the news?

Share of respondents who felt very or somewhat comfortable with each way of producing news. August–September 2026.

Journalist alone**67%**
Mainly a journalist, with AI assistance**45%**
Mainly AI, with human oversight**19%**
AI alone**12%**
Compare with news produced by a journalist alone
Using the journalist-only result (67%) as the reference, reported comfort was 22 percentage points lower when a journalist led with AI assistance, 48 points lower when AI led with human oversight, and 55 points lower for AI alone. These differences compare survey responses; they do not measure the quality of the news.

Read the survey figures as a table

How comfortable are people with different ways of making news?

Who makes the news?	Very or somewhat comfortable
A journalist, entirely	67%
Mainly a journalist, with AI assistance	45%
Mainly AI, with some human oversight	19%
AI, entirely	12%

View the original chart



Reuters Institute, *Generative AI and News Report 2026*. YouGov surveyed approximately 2,000 people per country online in nine countries, 14 August–28 September 2026. Each percentage is a weighted estimate across respondents in the nine countries for one scenario. The four figures do not add up to 100%. They measure comfort, not accuracy, liability or worldwide opinion.

Whose publication was it?

If a user uploads a false video to YouTube or Facebook, the user who posted it and the platform that hosts it have taken different actions. What did the uploader claim? Was the material unlawful? What notice reached the platform, and how did it respond? Under the EU Digital Services Act, hosting services have a conditional exemption for material stored at a user's request. A precise notice may establish knowledge of illegality; how promptly the service acts then matters. The Act does not impose general monitoring of every upload. A false claim is not necessarily unlawful, and neither party is automatically liable for the other's conduct. Digital Services Act, Articles 6, 8 and 16.

Compare that with a business giving advice on its own website. In *Moffatt v. Air Canada* (2024 BCCRT 149, paragraphs 14 and 24–32), a traveller relied on the airline chatbot's incorrect advice about claiming a bereavement fare after travel. British Columbia's Civil Resolution Tribunal awarded damages for negligent misrepresentation. The tribunal examined the airline's duty, the advice, the traveller's reliance and the loss. It did not need to decide what kind of chatbot this was. The important distinction was whose advice the customer had received.

What the reader and the driver are owed

The reader should be able to open tomorrow's paper and find the correction. The driver should be able to give the dealer a vehicle identification number and find out whether the car needs attention. For either answer to mean anything, somebody must have kept the evidence and somebody must have the authority to act on it.

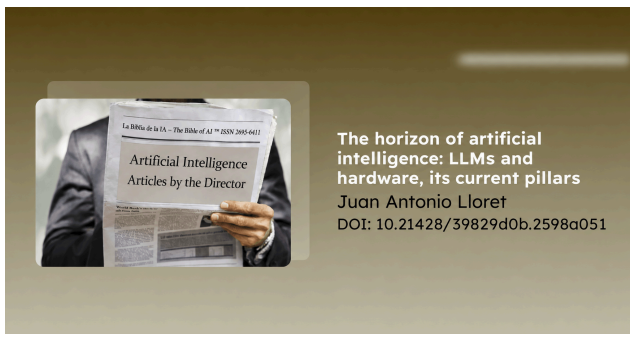
That is the standard I would apply when AI is involved. **Traceability** lets us follow a photograph from its source to the printed page, or a weld from the drawing to the finished car. **Auditability** lets another person check those records and challenge the conclusion. A watermark may tell me how a picture was made. The publisher's or manufacturer's name tells me whom to ask. The real test of responsibility comes when that person or institution has to answer.

The horizon of artificial intelligence: LLMs and hardware, its current pillars

Article 006 · 11 November 2026

© 2026. All rights reserved. | Juan Antonio Lloret Egea | ORCID: 0000-0002-6634-3351 | Instituto Tecnológico Virtual de la Inteligencia Artificial para el Español™ (ITVIA) | IA eñ™ — La Biblia de la IA™ | ISSN 2695-6411 | CC BY-NC-ND 4.0 | Article 006 · 11 November 2026

Spanish original in ITVIA · DOI: 10.21428/39829d0b.2598a051



Competition between accelerators expands AI's possibilities, but capacity contracts and dependence on their tools also shape which paths can be pursued.

Anthropic, the company behind Claude, offers a revealing picture of the race for artificial intelligence. It has committed to spending more than $100 billion on Amazon technologies over ten years [1]; agreed to purchase $30 billion of Azure capacity, powered by NVIDIA [2]; and announced five gigawatts of capacity with Google and Broadcom, within a TPU expansion expected to come online from 2027 [3, 4]. It thus combines Trainium, GPUs and TPUs. These are commitments with different delivery schedules, not capacity already available in full. Securing somewhere to run models has become a decision capable of shaping which models can be built.

This diversification reflects real differences. NVIDIA and AMD graphics processing units (GPUs) support a broad range of parallel computing, from simulation and scientific analysis to artificial intelligence [5, 6]. Google tensor processing units (TPUs) specialise in matrix operations, which are central to neural networks [7]. Generality makes it easier to experiment with varied tasks; specialisation allows computation to be organised around particular workloads. Both families support training and inference, and some GPUs also contain specialised units [7, 8]. The advantage depends on the workload, memory, communications and cost of the complete system. A peak performance figure alone cannot compare the accuracy, response times and actual utilisation of two installations.

NVIDIA's position becomes clearer when we examine CUDA, a computing platform and programming model with libraries and tools accumulated over many years [5]. That investment becomes embedded in programs, teams' experience and their verification procedures. Moving them to another manufacturer can require adaptation, optimisation and fresh testing [6]. This creates a competitive barrier that a TPU's greater efficiency on a particular task does not automatically remove. Even incorporating Rust—a language I strongly recommend—into CUDA, as NVIDIA announced in September, expands the ways to program its GPUs without removing dependence on its platform [9].

Yet that barrier is being challenged. Google is developing TorchTPU to run PyTorch programs on its accelerators with fewer changes [10]. AMD combines its Instinct GPUs with ROCm and has an agreement with OpenAI to deploy six gigawatts over several years [6, 11]. Commercial boundaries are not absolute either: Google's TPU 8tx system, announced on 9 November, incorporates AMD central processors [12]. Competing over the accelerator and collaborating on the rest of the system can be parts of the same business. The cost of switching suppliers is therefore itself a target for innovation, rather than an immutable property of the market.

Large language models (LLMs) feed this pressure. Training them requires processing vast quantities of data; using them sustains ongoing demand. Long conversations, elaborate responses and systems that perform sequences of tasks add computation and memory requirements [12, 13]. Each operation can become cheaper while aggregate consumption rises. The International Energy Agency records efficiency improvements alongside constraints on electricity supply, semiconductor manufacturing and high-bandwidth memory [13]. Pressure consequently reaches manufacturers of memory, networking and electrical equipment. The constraint extends beyond the model to the infrastructure needed to use it.

My reading is that the market is selecting part of AI's future before that future can be scientifically assessed. Amazon, NVIDIA and Microsoft have announced investments in Anthropic [1, 2]. Suppliers thus finance the companies buying their capacity, while buyers diversify to secure supply and negotiate terms. These agreements enable advances, but also commit resources for years. We should ask which research directions this structure favours and which become difficult to finance. The ability to contract more computing power does not, by itself, measure the quality of the reasoning obtained.

As director of several publications, I believe we should also investigate artificial intelligence that does not make large-scale training a condition of its existence. The book *Artificial Intelligence of the Vectorial System* develops this alternative from Image-Based Language and its relationship with a new programming language: the Vectorial System's Computing Language [14, 15]. Learning is conceived as incorporating explicit knowledge; reasoning, as verifiable transformations of visual representations, without verbal prediction or opaque weight adjustment. This architecture does not conceptually require a data centre [16]. Its complete implementation and cost on comparable tasks must be demonstrated [15, 16], but it allows us to question whether LLM requirements are universal necessities for AI.

Competition between processors can expand the possibilities of artificial intelligence, provided alternatives remain open. Progress should be judged by the cost of obtaining a valid result, what makes it verifiable and who can reproduce it. If access to capital also determines what deserves to be called intelligence, we will have handed the market a decision that still requires scientific judgement.

Sources

Accessed on 11 November 2026. Corporate announcements document the agreements announced; they are not independent assessments of their implementation or performance.

1. Amazon (2026, April 20). *Amazon and Anthropic expand strategic collaboration*. Corporate announcement.
2. Anthropic (2025, November 18). *Microsoft, NVIDIA, and Anthropic announce strategic partnerships*. Corporate announcement.
3. Anthropic (2026, May 28). *Anthropic raises $65B in Series H funding at $965B post-money valuation*. Corporate announcement.
4. Anthropic (2026, April 6). *Anthropic expands partnership with Google and Broadcom for multiple gigawatts of next-generation compute*. Corporate announcement.
5. NVIDIA (n.d.). *CUDA Platform for Accelerated Computing*. Technical documentation.
6. AMD (n.d.). *AMD ROCm Software*. Technical documentation; sections on scientific applications and adapting CUDA programs.
7. Google Cloud (n.d.). *TPU architecture*. Technical documentation.
8. NVIDIA Developer Blog (2018, September 14). *NVIDIA Turing Architecture In-Depth*. Technical article; section "*Turing Tensor Cores*".
9. NVIDIA Developer Blog (2026, September 8). *Introducing CUDA Rust: Two Tracks for Writing GPU Kernels*. Technical article.
10. Google Developers Blog (2026, April 7). *TorchTPU: Running PyTorch Natively on TPUs at Google Scale*. Technical article.
11. AMD (2025, October 6). *AMD and OpenAI Announce Strategic Partnership to Deploy 6 Gigawatts of AMD GPUs*. Press release.
12. Google Cloud Blog (2026, November 9). *Our newest TPUs: 8tx and 8i Max, built for the agentic era*. Technical system announcement.
13. International Energy Agency (2026). *Key Questions on Energy and AI*. Report; executive summary. Paris: IEA.
14. Lloret Egea, J. A. (2026, November 8). *Inteligencia Artificial del Sistema Vectorial: razonamiento visual, evidencia y trazabilidad*. Version V.05. ITVIA.
15. Lloret Egea, J. A. (2026, November 8). *Tratado Lenguaje Basado en Imágenes*. Version V.13, chapter 19: "Programación visual y Lenguaje de Computación del Sistema Vectorial". ITVIA.
16. Lloret Egea, J. A. (2026, November 8). *Memoria, aprendizaje y realización computacional*. In *Inteligencia Artificial del Sistema Vectorial*, version V.05, chapter 8. ITVIA.