frontiers | Frontiers in Science

TYPE Frontiers in Science Editorial
PUBLISHED 16 December 2025
DOI 10.3389/fsci.2025.1760821

OPEN ACCESS

EDITED BY
Frontiers in Science Editorial Office,
Frontiers Media SA, Switzerland

*CORRESPONDENCE
R. Stanley Williams
rstanleywilliams@tamu.edu

RECEIVED 04 December 2025
ACCEPTED 05 December 2025
PUBLISHED 16 December 2025

CITATION
Williams RS. Why does artificial intelligence need active memory to succeed? *Front Sci* (2025) 3:1760821.
doi: 10.3389/fsci.2025.1760821

Why does artificial intelligence need active memory to succeed?

R. Stanley Williams[1,2]*

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, United States, [2]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, United States

KEYWORDS

artificial neural networks, spiking neural networks, bottleneck, compute-in-memory, analog, stochastic

An Editorial on the Frontiers in Science Lead Article
Breaking the memory wall: next-generation artificial intelligence hardware

Key points

- The super-exponential growth of data harvested from human input and physical measurements is exceeding our ability to build and power infrastructure to communicate, store and analyze, making our present artificial intelligence trajectory economically unsustainable.
- A major limitation is the mismatch between the enormous parallel computing needs of artificial neural networks and the traditional computing paradigm that separates independently optimized compute and memory, which has led to the huge latency and energy inefficiencies known as the memory wall.
- There are numerous hardware research proposals in the areas of compute-in-memory and tighter integration of sensors with computing networks that indicate improvements of three orders of magnitude or more in speed and power consumption are possible, dramatically democratizing AI.
- These approaches will need to coordinate research in materials, devices, circuits, architectures and algorithms for a total AI solution with active memory, rather than simply expose a new bottleneck somewhere else in the system, as in Amdahl's Law, to realize their potential.

Artificial intelligence (AI) may be racing ahead, but it is running straight into a wall—the *memory wall*. The faster AI models grow, the more energy and time they waste simply moving data between memory and computation. In their lead article in Frontiers in Science, Kaushik Roy and colleagues confront this problem head-on, providing a cross section of frontier next-generation hardware research that can overcome this bottleneck through active compute-in-memory (CIM) and brain-inspired architectures (1).

Meanwhile, the amount of data being collected worldwide is increasing at a super-exponential rate (2), growing much faster than we can communicate through the cloud or store in data centers, let alone analyse effectively. AI is seen as the only way to make sense of all this data and transform it into useful and valuable information, but data center construction and energy costs are skyrocketing, and significant doubts are mounting about the economic sustainability of the current AI model.

Are we trying to move too fast, or are we trapped in a hype cycle? Where should AI computation occur? In behemoth cloud data factories or on the "edge" of the internet, where data are collected and used? At the present, training AI models demands massive datasets and highly precise calculations that require centralization. Once a particular type of AI has been trained, however, it should be possible to transport its parameters to an edge system to perform inference, which requires less precision and can even benefit from a degree of stochastic behaviour using analogue representations of numbers. A general class known as artificial neural networks (ANNs) has been force-fit into today's computer and data architectures but may ultimately prove inadequate for the kinds of cognition AI is promised to deliver. A more brain-like genre that emulates the way neurons communicate and store information is a spiking neural network (SNN), which is well suited for operating on real-time data inputs in edge environments, though the technology remains in its early stages.

The intellectual challenges faced by those trying to turn AI into a reliable technology are enormous, but no matter how brilliant the algorithms and the software to implement them, AI will not be useful if it remains too slow and expensive to use (3). This is why any long-term effort to advance AI must dismantle the memory wall. For the past sixty years, computing has relied on the Boolean logic paradigm, built on silicon-based material systems that use the physical interaction between voltages and charge. This has enabled chip designers and foundries to iterate and optimize this one platform at an astounding rate, known as Moore's Law. Memory, however, is a very different beast, and it has evolved far more slowly. This mismatch is the origin of the memory wall, a widening chasm between computation and data storage that is growing rapidly with time. There is no universal memory, but rather many diverse types based on a wide range of different materials and physical interactions. Modern computing systems choreograph data through a hierarchy of multiple memories—from vast, inexpensive storage drives to fast but power-hungry caches near the processor. Every calculation requires shuttling data across this hierarchy, consuming significant time and energy. This fundamental separation of memory and compute, known as the von Neumann bottleneck, is an unavoidable consequence of the fact that the technologies used to build them are inherently incompatible.

Moreover, memory technology itself is in a constant state of flux, with new materials, device structures and physical interactions continually being introduced in the quest of creating a true universal memory that merges some levels of the existing hierarchy. Many of these innovations appear extremely promising at first, but they invariably require decades of research and development that can cost hundreds of millions of dollars, only to falter when an unanticipated problem arises in bringing a product to market, as happened with Optane—an Intel memory and storage technology (4). The challenges and pitfalls of creating yet another new memory have understandably made the tech industry wary, but without continued exploration of new ideas and investment, we simply reach the edge of the chasm, watching it grow wider and more difficult to cross as we cling to our entrenched technologies.

Is there a path forward for AI? The progress in memory technology has been slower and riskier than in logic, leaving computing architectures increasingly lopsided and unsustainable in both energy and cost. This imbalance opens the door to more radical ideas. If we look to the brain, we see that compute, memory and communication are all embodied in a network of neurons. Rather than attempting to cram more memory into a computing chip, can we instead perform most of the computing *inside* an active memory where data reside, without costly movement? This is the promise of compute-in-memory (CIM), which collapses the traditional boundary between logic and storage by letting the same physical elements perform both roles. CIM not only reduces data movement but allows entirely new types of computation, analog, stochastic and event-driven, that mimic aspects of biological brains (5). These architectures can execute matrix operations and learning tasks with drastically lower energy demands. Yet they require new design principles: error-tolerant algorithms, hybrid analog-digital circuits, and co-design across device, architecture, and algorithm layers. There is great promise in this approach, but also many unresolved questions. What type of active memory is best for CIM, digital or analog? How should it be integrated with the traditional memory hierarchy, or does it require an entirely different architecture? How will data flow through these networks? Will we be able to develop general purpose hardware, or will different types of applications need specific implementations of CIM?

The lead article by Kaushik Roy et al. (1) captures a snapshot of the range of research that is presently being pursued to answer the above questions. The broader AI issues are invoked, but to provide a focus and ground their discussion on an edge application, the authors have chosen the example of an autonomous drone—a quintessential edge application—to illustrate the challenges of sensor fusion, data flow, computation and real-time response to a changing and uncertain environment. Domain experts have come together who collaborate closely with each other but maintain unique viewpoints to provide potential answers for CIM and sensor components, as well as system-level integration. This level of horizontal communication in the research stage is historically unusual but necessary for ideas to be vetted and new breakthroughs to propagate rapidly through the community. The careful attention to detail in the figures of this paper deserves special attention, in that they provide a visceral illustration and unify the concepts presented by the authors, linking material science, circuit design, and intelligent behaviour to provide a coherent whole.

The road ahead will not be linear, nor will any single technology provide a complete solution. The convergence of memory and computation represents a profound technological shift; a move from machines that merely process information to ones that *embody* it. Rapid progress will require simultaneous advances in materials, devices, circuits, architectures and algorithms, with continual communication within the extended community to avoid the creation of hyper-optimized components that are not compatible with a system. Whether AI can truly succeed may depend not on how much it knows, but on how it remembers.

Statements

Author contributions

RSW: Writing – original draft, Writing – review & editing, Conceptualization.

Funding

The author declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author is a named inventor on over 230 issued United States patents in the field of this editorial. Most of these patents were obtained during a 23-year career at HP/Hewlett Packard/HPE, from which the author is now retired. The author does not own any of these patents.

Generative AI Statement

The author declared that generative AI was used in the creation of this manuscript. The author used Claude to suggest wording and structure, but in the end did not use any of the recommendations.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Roy K, Kosta A, Sharma T, Negi S, Sharma D, Saxena U, et al. Breaking the memory wall: next-generation artificial intelligence hardware. *Front Sci* (2025) 3:1611658. doi: 10.3389/fsci.2025.1611658

2. Clissa L, Lassnig M, Rinaldi L. How big is Big Data? A comprehensive survey of data production, storage, and streaming in science and industry. *Front Big Data* (2023) 6:271639. doi: 10.3389/fdata.2023.1271639

3. Chui M, Hazan E, Roberts R, Singla A, Smaje K, Sukharevsky A, et al. *The economic potential of generative AI: the next productivity frontier*. McKinsey & Company (2023). Available at: https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

4. Coughlin T. Intel winding down its Optane memory business. *Forbes* (2022). Available at: https://www.forbes.com/sites/tomcoughlin/2022/07/28/intel-winding-down-its-optane-memory-business/

5. Sebastian A, Le Gallo M, Khaddam-Aljameh R, Eleftheriou E. Memory devices and applications for in-memory computing. *Nat Nanotechnol* (2020) 15:529–44. doi: 10.1038/s41565-020-0655-z