PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-qmm2-7uwu **Comments Due:** May 28, 2025 **Submission Type:** API

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0234
Comment on FR Doc # 2025-07332

Submitter Information

Name: Peter Thomas

General Comment

Please see the attached white paper (5 pages + references), "Energy Efficient Technologies for Next-Generation Artificial Intelligence."

Attachments

Energy-Efficient-Technologies-for-Next-Generation-Artificial-Intelligence

Energy Efficient Technologies for Next-Generation Artificial Intelligence

Docket ID No. NSF-2025-OGC-0001

Hillel J. Chiel, Case Western Reserve University
Jay Coggan, NeuroLinx Research Institute
Gourav Datta, Case Western Reserve University
Jean-Marc Fellous, University of California San Diego
Roger D. Quinn, Case Western Reserve University
Peter J. Thomas, Case Western Reserve University

Since the advent of widely accessible AI tools, AI technology has been in high demand by businesses, individuals, and academic researchers. Technology companies have been building AI infrastructure at a rapid pace, and these facilities have been consuming vast and growing resources, particularly electricity and water, with significant real and projected climate impacts. There is a need for government sponsored research to support long time horizon efforts to develop **energy efficient computing capabilities** to support the continued growth of the nation's AI infrastructure in a sustainable fashion. Such efficiency is required at both the hardware and software levels. *Where can the industry turn for examples of ultra low power, energy efficient computing?* We argue that **neurobiological principles** offer rich and under-exploited sources of inspiration for energy efficient computing, and that new partnerships between industry and academia should be developed.

It is time to clarify what we want from AI systems and what AI actually means. Large Language Models (LLMs) are remarkable tools in their early stages but arguably not intelligent systems by any useful definition and certainly not the path to Artificial General Intelligence (AGI). LLMs are structurally very limited, simply being rapid statistical sampling and prediction programs that operate on input data. The well-known hallucinations result from several fatal flaws including bad or incomplete data and inadequate algorithms. This approach will not scale, they are already hitting the wall of exponential energy use for very incremental gains in function. The costs of AI are already "obscene" as commented by the New Yorker[1], elaborated by the MIT Tech Review[2], and data centers are taking ever more from the grid (de Vries, 2023). This is a poor return on investment. The AI progress envisioned by the "Agent 4" superintelligence in the popular sci-fi speculation based on AI-2027 (https://ai-2027.com/) is very likely physically impossible with current chip architectures on energy consumption grounds alone. If we want energy efficient AGI, we should study how nature does it: biological brains are by far the most energy efficient computing devices in the universe, using only about 20 W of power, which is more than 1 million fold less than the world's largest supercomputer, a machine that is considered not nearly adequate to approximate human intelligence. It is possible that we won't know exactly how efficient the brain actually is in units of Watts/Flop until we establish what the biological equivalent is, the BioFlop, as coined by Stiefel and Coggan (2023).

[1] https://www.newyorker.com/news/daily-comment/the-obscene-energy-demands-of-ai
[2] https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/

In this same paper, the authors also introduced a measure of AI efficiency, the ERASI equation. This calculation has initially shown that the cost of mimicking the human brain (based on what we now know about its computations) would be orders of magnitude higher than the entire annual US energy output. As a benchmark, the authors used estimates of energy use from the now shuttered Blue Brain Project, a Swiss government initiative to simulate an entire mouse brain based on a biologically realistic reconstruction of brain circuits. This initiative has birthed a very different approach to AI than LLMs, one that attempts to discover learning rules that are more biological (see also https://www.inait.ai). For example, one of the many faults of LLMs is underestimating how single cells process information. There is a growing theoretical base for conceptualizing this level of computation (e.g., Rieke et al., 1999; Poirazi and Mel, 2001; Agüera y Arcas et al., 2003; Coggan et al., 2020; Lillicrap et al., 2020; Coggan et al., 2022). In short, research on biological brains cannot be separated from overall initiatives in AI research. "Neuromorphic engineering", a popular concept informing AI computing architectures, is just a form of biomimicry; we have much more to learn from biological solutions. It cannot be overemphasized that basic neurobiology research will lead to more powerful and efficient AI.

Leaner Training Algorithms Should Save Energy. One of the major differences between natural (human/biological) and artificial intelligence is the ability to 1) learn with few examples and 2) learn continuously with minimal teaching or supervision. Both of these well documented features save time and energy. Due to availability and low cost of computing resources, current AI systems do not typically attempt to address, and industry has not yet shown a motivation to develop them. Government-sponsored research specifically targeted at continuous learning and small training set techniques and theories could be a significant drive. Alternative algorithms to those commonly used in LLMs today could also be tested for efficiency (e.g., Grossberg, 2020).

Sparse Representations May Be Energy Efficient. Sparse representations that improve reconstruction from noisy data have been intensively studied in applied mathematics (Calvetti et al 2019, 2020, 2024). The connection between sparse representation and energy efficiency has been investigated in neurobiology (Hu et al 2012, Sacramento et al 2015, Moosavi et al 2024). In addition to saving energy, sparse representations could also render AI systems more robust to adversarial attacks and catastrophic failures.

Multi-Resolution Representations May Be Energy Efficient. Not all information needs to be represented at the same resolution. The visual system has evolved multi-scale coding (different receptive field sizes within cortical areas (e.g. cortical area V1) and across areas (e.g. cortical areas V1-V4-IT). Spatial navigation in large environments uses multi-scale hippocampal place fields (Harland et al 2021). This multi-scale representation has been shown to have many computational advantages, including efficiency and fast attentional shifting. Most AI systems do not make use of multi-scale representations.

Memory Changes with Time and Use, Yielding More Efficient Representations. Unlike computers, we (humans and animals alike) store information neither "forever" nor exactly as it was acquired. We do not store pixel-level images, or second-by-second sequences of episodic memories. We have an ability to abstract, simplify or ignore our inputs as they come in, and to

re-shape our memories as a function of how or how often we use them, or as a function of their intrinsic “importance”. These features have undoubtedly evolved (at a cost for reliability) to address our limited capacities to perceive and memorize, saving energy and time. AI systems do not implement the (useful) mechanisms of forgetting, or memory consolidation (as during sleep). Industry considers these features as deleterious and artifactual, to be avoided and corrected to achieve reliable and precise recall. There is no incentive to explore the extent to which a trade-off can or should be achieved between full recall and precision, and efficiency of representations.

Offline and Offsite Processing Save Time and Energy. Our brain processes information during sleep, with minimal energy expenditure. It is also capable of dynamically allocating resources (e.g. attention) to specific tasks or sensory pathways in a context-dependent manner. AI is essentially stimulus driven (prompt, get an answer), and is active when humans are active (peak energy expenditure). One could imagine world-wide AI architectures that allocate AI tasks to regions of the world where energy is the cheapest/lowest (e.g. prompt during the day in the USA, get an immediate AI answer/processing at night, in India). One could also imagine AI systems that compile answers, restructure knowledge, and improve their database during ‘off line’ hours (when humans sleep). Not all AI computations need be human-stimulus driven.

Sparse Bursting Activity is Common in Motor Control / Central Pattern Generator Systems.
In neuro-motor systems, it is inefficient to simultaneously tense opposing muscles, or to activate muscles that are not needed for a given movement. Thus it is no surprise that patterns of activity in biological motor control systems exhibit sparse bursting patterns, in which a small number of neurons are active at any one time. Such sparse activity patterns are mediated by inhibition, in which the activity of one cell suppresses the activities of others. Similarly, inhibition plays a key role in winner-take-all (WTA) algorithms that select one out of many possible answers (Maass 2000). If a unit consumes more energy when it is “active” than when it is “silent” then WTA dynamics lends itself to energy-efficient implementation. And inhibition underlies stable heteroclinic channels (SHCs) which have been proposed as a dynamical architecture supporting sparse, functionally effective activation patterns in motor systems (Horchler et al 2015, Rouse and Daltorio 2021, Mengers et al 2025). Government-supported initiatives to study SHCs, WTA architectures, and other mechanisms for inhibition-dominated connectivity in AI systems may lead to novel energy-efficient solutions.

Neuromodulation Suppresses the Activity of Subnetworks that are not Needed for a Given Activity. In motor control systems, metabolic resources are redirected to muscle systems that are actively being used. Switching muscles “on” and “off” is regulated by neuromodulation. Similarly, activity in neural subsystems is also regulated by neuromodulation. Specializing circuits for specific computational tasks and then powering down those circuits when those tasks are not needed may recapitulate biological-like (biomorphic) efficiencies. More generally, the extent to which neuromodulation in general (e.g. via serotonin, norepineprhine, dopamine, and others) has evolved to improve the efficiency of neural computation (rather than perform neural computation *per se*) is understudied and deserves attention (Castrillon et al, 2023; Yu et

al, 2025). Implementing neuromodulatory principles in “classic” AI models might significantly improve their performance, and consequently, their energy expenditure.

Plasticity Modulation for Adaptive Intelligence

In addition to gating circuit activity for energy efficiency, neuromodulation in biological systems plays a central role in regulating when and where learning occurs. Neuromodulators such as dopamine, acetylcholine, and norepinephrine do not directly compute but instead modulate synaptic plasticity, guiding long-term changes in neural circuits based on behavioral relevance, reward prediction, or uncertainty. This dynamic plasticity control enables animals to learn selectively, preserving stable functions while adapting rapidly to new situations—a capability that current AI systems, including large language models (LLMs), largely lack. Instead, these models rely on static, global learning schedules and costly retraining for adaptation. By integrating neuromodulatory principles, such as plasticity gating conditioned on internal goals or novelty signals, future AI systems could enable targeted, context-aware updates to their internal states. This approach offers a path to learning-to-learn (meta-learning) in large-scale models, improving data efficiency, stability, and the ability to generalize across tasks, while avoiding the overhead of continual full-network retraining.

Neuroscience-Inspired Memory Efficiency

Biological brains achieve remarkable memory efficiency through mechanisms such as sparse coding, synaptic consolidation, and multi-scale memory hierarchies. Unlike large-scale AI models that rely on dense parameter storage and exhaustive training data exposure, the brain encodes information using compact, context-dependent neural activations, often reusing the same networks across tasks via dynamic routing or population coding. Furthermore, systems consolidation, involving transfer from fast-learning hippocampal circuits to slower cortical storage, enables long-term retention without continuous memory access. These strategies contrast with the monolithic memory structures of LLMs, which grow in size and cost with increasing data. Incorporating such biological insights—e.g., sparsity-inducing priors, gated memory consolidation, or attention-modulated storage and retrieval mechanisms—could significantly reduce the memory footprint of large models without sacrificing performance. This approach not only improves energy and storage efficiency, but also supports more flexible, lifelong learning paradigms where memory is treated as a dynamic, structured resource rather than a static archive.

Cortical Traveling Waves and Structured Dynamics in Large-Scale Neural Models

A growing body of neuroscience research points to the central role of spatiotemporal cortical waves, such as alpha, beta, and theta oscillations, in orchestrating perception, attention, and memory in the brain. These waves enable multiplexed signaling and efficient integration across distributed cortical areas by rhythmically modulating neuronal excitability. Yet modern large-scale AI models, including LLMs, operate with fundamentally static or token-synchronous dynamics. Bridging this gap presents a compelling research frontier: introducing traveling-wave-like dynamics and oscillatory gating into LLMs and other large-scale architectures. This approach could involve rhythm-based attention windows, phase-coupled memory activation, and dynamic routing paths that emulate the selective coherence seen in

cortical circuits. Such structured dynamics would not only reduce inference cost and improve temporal coherence, but also pave the way for closed-loop agentic systems, where internal wave states regulate external actions in response to changing sensory or goal contexts. Simulating these wave phenomena in neuromorphic or FPGA-based edge architectures can further support real-time, biologically grounded inference under strict energy constraints.

Novel Mechanisms for Research Partnerships with Industry and/or Academia:
AI can mimic and eventually improve natural, biological intelligence, the functions and efficiencies of which are far from understood. It therefore stands to reason that we need to understand more about biological intelligence, otherwise AI efforts will be searching for solutions in the dark, forced to implement untested strategies at potentially high financial and energy costs. One problem with the current scientific approach to AI is the very limited cross-pollination of ideas between the two fields. On the one hand, most computer scientists and physicists, including those enjoying notoriety in the AI field today, have limited appreciation of neurobiology or psychology. On the other hand, most neurobiologists lack sufficient understanding of computer algorithms, coding and the physics of information processing. Although there have been some efforts to integrate this knowledge, communication between these research communities remains hamstrung, partially due to siloed training experiences.

Solving AI will require establishing new research ecosystems where these two approaches to intelligence are encouraged to flourish synergistically, with a new generation of experts who are well-versed in biology, psychology, physics, engineering and computer science. Such integrated cross-disciplinary training should be reflected in new departments at major universities, new funding for multi-disciplinary training at multiple career stages, as well as initiatives to bridge academic and industry priorities. In contrast to today, 40 years ago, there were no undergraduate neuroscience departments or programs; interested students had to choose between psychology or biology or computer science. It is not unprecedented therefore that the organization of academia should adapt to the new requirements of the society. This process has already started at a few institutions, such as at Rice University[3], which now offer AI majors, but these curricula still fall short, particularly in biology. Given the challenges facing the budget of the US government, the integrated involvement of foundations, non-profits and private industry would allow more efficient use of taxpayer allocated resources by reducing overhead costs, enabling administrative agility and creating a profit-sharing environment that fosters innovation.

References:

Agüera y Arcas B, Fairhall AL, Bialek W. (2003) Computation in a single neuron: Hodgkin and Huxley revisited. Neural Comput. Aug;15(8):1715-49. doi: 10.1162/08997660360675017.

Calvetti, Daniela, Erkki Somersalo, and A. Strang. (2019) Hierachical Bayesian models and sparsity: ℓ2-magic. *Inverse Problems* 35.3: 035003.

[3] https://news.rice.edu/news/2025/rice-offer-bachelor-science-artificial-intelligence

Calvetti, D., Pragliola, M., Somersalo, E., & Strang, A. (2020). Sparse reconstructions from few noisy data: analysis of hierarchical Bayesian models with generalized gamma hyperpriors. *Inverse Problems*, *36*(2), 025010.

Calvetti, Daniela, and Erkki Somersalo.(2024) Computationally efficient sampling methods for sparsity promoting hierarchical Bayesian models. *SIAM/ASA Journal on Uncertainty Quantification* 12.2: 524-548.

Gabriel Castrillon et al. (2023)An energy costly architecture of neuromodulators for human brain evolution and cognition. Sci. Adv. 9, eadi7632 .

Coggan JS, Keller D, Markram H, Schürmann F, Magistretti PJ. (2022) Representing stimulus information in an energy metabolism pathway. J Theor Biol. 2022 May 7;540:111090. doi: 10.1016/j.jtbi.2022.111090. Epub 2022 Mar 7.

Coggan JS, Keller D, Markram H, Schürmann F, Magistretti PJ. (2020) Excitation states of metabolic networks predict dose-response fingerprinting and ligand pulse phase signalling. J Theor Biol. 2020 Feb 21;487:110123. doi: 10.1016/j.jtbi.2019.110123. Epub 2019 Dec 19.

Grossberg S (2020) A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action. Front Neurorobot. 2020 Jun 25:14:36. doi: 10.3389/fnbot.2020.00036.eCollection 2020.

Harland B, Contreras M, Souder M, Fellous JM (2021) Dorsal CA1 hippocampal place cells form a multi-scale representation of megaspace. Curr Biol 31:2178-2190 e2176.

Horchler, A. D., Daltorio, K. A., Chiel, H. J., & Quinn, R. D. (2015). Designing responsive pattern generators: stable heteroclinic channel cycles for modeling and control. Bioinspiration & biomimetics, 10(2), 026001.

Hu, T., Genkin, A., & Chklovskii, D. B. (2012). A network of spiking neurons for computing sparse representations in an energy-efficient way. Neural computation, 24(11), 2852-2872.

Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G. (2020) Backpropagation and the brain. Nat Rev Neurosci. 2020 Jun;21(6):335-346. doi: 10.1038/s41583-020-0277-3. Epub 2020 Apr 17.

Maass, W. (2000). On the computational power of winner-take-all. Neural computation, 12(11), 2519-2535.

Mengers, N., Rouse, N., & Daltorio, K. A. (2025). Stable heteroclinic channels for controlling a simulated aquatic serpentine robot in narrow crevices. Frontiers in Electronics, 6, 1507644.

Moosavi, S. A., Pastor, A., Ornelas, A. G., Tring, E., & Ringach, D. L. (2024). Temporal dynamics of energy-efficient coding in mouse primary visual cortex. bioRxiv.

Poirazi P, Mel BW. (2001) Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. Neuron. 2001 Mar;29(3):779-96. doi: 10.1016/s0896-6273(01)00252-5.

Rieke, F., Warland, D., Van Steveninck, R. D. R., & Bialek, W. (1999). Spikes: exploring the neural code. MIT press. https://mitpress.mit.edu/9780262181747/spikes/

Rouse, N. A., & Daltorio, K. A. (2021). Visualization of stable heteroclinic channel-based movement primitives. IEEE Robotics and Automation Letters, 6(2), 2343-2348.

Sacramento, J., Wichert, A., & van Rossum, M. C. (2015). Energy efficient sparse connectivity from imbalanced synaptic plasticity rules. PLOS Computational Biology, 11(6), e1004265.

Stiefel KM, Coggan JS. (2023) The energy challenges of artificial superintelligence. Front Artif Intell. 2023 Oct 24;6:1240653. doi: 10.3389/frai.2023.1240653. eCollection 2023.

de Vries, A (2023) The growing energy footprint of artificial intelligence Joule, Volume 7, Issue 10, 2191 - 2194.

Yu, Z., Wang, Y., Thomas, P. J., & Chiel, H. J. (2025). Tradeoffs in the energetic value of neuromodulation in a closed-loop neuromechanical system. Journal of Theoretical Biology, 112050.