A call for intrinsic learning

Andy Kitchen
Cortical Labs
Berlin, Germany
andy@corticallabs.com

Abstract

Current artificial intelligence systems predominantly rely on extrinsic learning mechanisms, with gradient descent and its variants serving as the primary means of model optimization. This approach treats learning as a distinct, external process separate from cognition. However, natural intelligent systems, such as the human brain, display intrinsic learning where learning and cognition are inseparable, integrated processes. We argue for a shift of focus toward intrinsic learning in AI systems, moving away from the heavy reliance on extrinsic optimization. We highlight the limitations of current AI methods, including their extreme sample inefficiency and dependence on vast amounts of human-generated data. By examining the shortcomings of current scaling approaches and proposing alternative pathways, we emphasize that genuine advancements in artificial general intelligence require systems that learn and adapt intrinsically. We encourage renewed attention to AI architectures that embed learning within the dynamics of the system itself, drawing inspiration from natural intelligence to foster more robust, efficient, and adaptive AI.

1 Towards intrinsic learning

The field of artificial intelligence has been dominated by extrinsic learning approaches, where models are trained using optimization algorithms like gradient descent. This method, while effective for specific tasks, fundamentally separates the process of learning from the behavior of the model itself. Learning is treated as a discrete, externally imposed phase, distinct from the model's operation once training is complete. This artificial separation imposes significant limitations, particularly in terms of adaptability, efficiency, and the alignment of AI systems with natural intelligence.

In natural systems, learning is an intrinsic, continuous process, inseparable from cognition. The human brain, for example, does not undergo a distinct "training" phase with an external optimizer; instead, learning occurs as part of an ongoing interaction with the environment. This intrinsic learning is not only more efficient but also more flexible, allowing for rapid adaptation in the face of novel situations. In stark contrast, modern AI systems are rigid, requiring vast amounts of data and computation to achieve limited forms of adaptation. These systems often fail to generalize beyond their training data, highlighting the inefficiency of extrinsic learning approaches.

The sample inefficiency of current AI methods is staggering. State-of-the-art models require trillions of tokens and massive computational resources to achieve their levels of performance on benchmark tasks OpenAI [2024], in comparison to the adaptive intelligence exhibited by humans, who learn from far fewer experiences. The over-reliance on brute force scaling has led to a fixation on larger datasets and more powerful hardware, rather than a reconsideration of the foundational principles of learning itself.

We argue that this reliance on external optimization algorithms has driven AI into a local minimum, stifling progress toward general intelligence. The separation between model and optimizer is not only

1^{st} Workshop on NeuroAI @ 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

artificial but detrimental, preventing the development of systems that can learn and adapt intrinsically as natural systems do. To move forward, AI research must embrace intrinsic learning — where learning is embedded within the dynamics of the system — enabling more robust, adaptive, and efficient models that better reflect the processes of natural intelligence. It is time for a fundamental rethinking of AI architectures, shifting the focus from extrinsic optimization to systems capable of intrinsic, motivated learning.

2 Beyond Hebbianism

Hebbian learning, often summarized by the phrase "cells that fire together, wire together," has long been a foundational principle in neuroscience and has inspired numerious approaches in artificial intelligence. While this form of learning captures an essential aspect of synaptic plasticity in biological brains, the current implementations in AI systems are too simplistic and limited in their scope. In many cases, the application of Hebbian-like mechanisms has failed to deliver the flexibility, generality and adaptability seen in natural intelligence.

A key limitation of most Hebbian-inspired learning algorithms is their reliance on pairwise correlations between neurons or units. This approach, while elegant in its simplicity, lacks the sophistication needed for complex task-solving and robust generalization. Hebbian rules, as implemented in many artificial systems, typically capture only low-order associations, making them prone to overfitting and incapable of handling more abstract, higher-order learning. Furthermore, these simple Hebbian mechanisms often fall short when learning from sparse feedback is required. As a result, they typically perform poorly compared to optimization-based learning methods.

Furthermore, in some cases, Hebbian learning rules are functionally equivalent to gradient descent Xie and Seung [2003]. This equivalence undermines the goal of moving beyond extrinsic learning frameworks, as it suggests that Hebbian approaches, when used in isolation, offer little advantage over existing optimization-based methods. When these systems largely approximate gradient-based learning, they inherit the same weaknesses, such as the need for large amounts of data, without delivering the full benefits of intrinsic learning.

What is needed is a more powerful and expansive approach to learning that goes beyond the simple low-order interactions captured by Hebbian learning. Biological systems demonstrate rich, multi-level learning processes that adapt dynamically, continuously integrating feedback from complex environments in ways that simple Hebbian rules cannot mimic. To move toward systems that exhibit intrinsic learning, investigators must accept the limitations of current Hebbian approaches and develop models that integrate broader, more dynamic forms of synaptic plasticity and adaptation. This will involve mechanisms that incorporate temporal, contextual, and global feedback processes, enabling a more powerful and flexible form of learning that can adapt to a wide variety of tasks and environments.

3 The limitations of in context learning for LLMs

Large Language Models (LLMs) have demonstrated significant advancements in the realm of artificial intelligence, particularly with their ability to perform "in-context learning." Dong et al. [2024] This form of learning allows models to adapt their responses based on the context provided in a conversation or task, simulating a level of real-time adjustment and flexibility. While in-context learning represents a shift away from the rigid, pre-trained behaviors of earlier models, it still falls short of true intrinsic learning. The underlying mechanics of LLMs remain fundamentally symbolic and static, limiting their adaptability and preventing genuine integration of learning into their operational dynamics.

At its core, in-context learning in LLMs is not learning in the traditional sense. The model does not internalize new knowledge or adjust its dynamics based on new inputs; rather, it generates responses by leveraging pre-trained patterns and statistical correlations learned during massive training phases. The system does not truly adapt to novel information in real time but instead performs a form of pattern-matching within its fixed knowledge base. This symbolic, non-integrated approach means that while LLMs can produce responses that seem contextually relevant, they are constrained by their pre-trained state and lack the ability to evolve with experience.

One of the primary limitations of this approach is its rigid reliance on pre-defined structures. In-context learning simulates flexibility by shifting focus based on input, but the system's behavior is

still driven by static representations learned during its original training. The model operates within the bounds of its existing knowledge, drawing on token-level correlations rather than developing a deeper understanding or adapting its internal architecture to new information. This creates a symbolic façade of learning without the underlying dynamism that characterizes intrinsic learning processes in natural systems.

Moreover, in-context learning lacks the continuous, feedback-driven integration seen in natural intelligence. In humans, learning is a constant process intertwined with cognition, where feedback from actions and environmental interactions directly influences future behavior. LLMs, on the other hand, operate in a distinct separation between training and inference. The model can respond contextually, but once the interaction ends, no information is retained or incorporated into its future decision-making processes. This disconnect between inference and learning fundamentally limits the capacity of LLMs to achieve true intelligence.

4 Chaos and intrinsic learning: a path forward

Modern artificial intelligence systems are typically designed to operate close to linear regimes, carefully tuned to avoid chaos. The reasoning behind this design choice is clear: chaotic dynamics are seen as unpredictable and difficult to control, and thus, AI models rely on smooth, predictable optimization landscapes amenable to gradient descent. However, this aversion to chaos may be preventing the emergence of more adaptive, self-organizing forms of learning, akin to what we observe in natural intelligent systems like the brain.

Contrary to this avoidance of chaos, recent work in neurodynamics suggests that chaotic processes, when harnessed properly, could enhance intrinsic learning. The brain's neurodynamics exhibit properties of self-organized criticality (SOC), where the system operates near the edge of chaos, displaying both stochastic and deterministic behaviors simultaneously Habibollahi et al. [2023], Ovchinnikov and Janusonis [2021], Ovchinnikov [2016]. This criticality is not an undesirable side effect, but a functional feature that allows the brain to balance stability with adaptability, enabling rapid learning from minimal data. In chaotic systems, small perturbations can lead to significant changes in system behavior, a property that could allow AI systems to adapt more flexibly and rapidly to new information.

In contrast, today's AI models avoid this dynamism. They are optimized for smooth behavior, employing gradient-based methods that inherently limit the system's ability to have rich internal dynamics. This reliance on external optimization, such as gradient descent, means that models can only learn when explicitly guided by an optimization algorithm, separating learning from the model's functioning. As a result, AI systems are unable to dynamically adjust their behavior without undergoing a retraining process that often relies on large quantities of data and computational power.

By introducing chaotic dynamics into AI architectures, we may have the missing element to create systems that are capable of powerful intrinsic learning — learning that is continuous and inseparable from the system's operation. The Supersymmetric Theory of Stochastics (STS) suggests that chaotic dynamics can be leveraged in a way that allows for efficient internal adaptation. In STS, noise-induced chaos occurs in a phase where the system remains integrable but is dominated by instanton processes, akin to the neuroavalanches observed in the brain Ovchinnikov and Janusonis [2021]. These chaotic dynamics could enable a form of self-organisation that is not extrinsically imposed but emerges naturally from the system's interactions with its environment.

Moving beyond the linear, gradient-based optimization approaches of today requires embracing the rich potential of chaos. Instead of seeing chaos as a threat to control and predictability, we should view it as a potential pathway to more robust, adaptive, and intrinsically learning AI systems. This shift will enable models that learn as they act, just as natural systems do, operating continuously near the edge of chaos to enhance performance and adaptability.

5 DishBrain and intrinsic learning for AI

DishBrain Kagan et al. [2022] represents a significant example of controlled intrinsic learning in biological neural networks, offering insights directly relevant to the development of intrinsic learning systems. DishBrain demonstrates that neurons, when integrated into a structured environment, exhibit

adaptive goal-directed behavior with minimal external intervention. This real-time, closed-loop interaction between neurons and a simulated game-world highlights the potential for systems to self-organize and learn intrinsically through direct feedback.

The rapid learning observed in DishBrain Habibollahi et al. [2022], driven by active inference via the free energy principle, shows how biological networks can adapt efficiently, even with sparse sensory inputs. This contrasts starkly with the rigid, task-specific optimization seen in modern AI, where learning is often slow and dependent on vast datasets. By embedding learning within the dynamics of the system itself, DishBrain exemplifies a model of intrinsic learning where the boundaries between cognition and adaptation blur. The goal-directed behavior emerging from these systems guides us toward future AI algorithms could operate more like natural intelligence — adapting continuously and intrinsically in response to environmental stimuli.

6 Conclusion

The current reliance on extrinsic learning methods, such as gradient descent, limits adaptability and efficiency compared to natural intelligence which learns intrinsically through continuous interaction with the environment. Hebbian-based approaches, though biologically inspired, remain overly simplistic and sometimes functionally equivalent to gradient descent. Embracing chaos, as seen in neurodynamic models, could lead to more adaptive, self-organizing AI systems by leveraging the unpredictability and flexibility that chaos and criticality offers.

In this context, DishBrain offers a glimpse into the future of intrinsic learning by integrating biological neurons into AI systems. Demonstrating real-time, goal-directed behavior through active inference, DishBrain exemplifies how adaptive intelligence can emerge from minimal external guidance. This approach suggests a pathway for AI systems that learn organically, moving beyond more rigid examples like in-context learning in LLMs.

To advance towards true intrinsic learning, AI must embrace more dynamic, integrated architectures — drawing from biological principles to create systems that adapt continuously and autonomously. This shift would bridge the current wide gap between artificial and natural intelligence, opening new possibilities for robust, flexible, and general intelligence.

References

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Xiong Xie and H. Sebastian Seung. Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural Computation*, 15(2):441–454, Feb 2003. doi: 10.1162/089976603762552988.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024. URL https://arxiv.org/abs/2301.00234.

Forough Habibollahi, Brett J. Kagan, Anthony N. Burkitt, et al. Critical dynamics arise during structured information presentation within embodied in vitro neuronal networks. *Nature Communications*, 14:5287, 2023. doi: 10.1038/s41467-023-41020-3. URL https://doi.org/10.1038/s41467-023-41020-3.

Igor V. Ovchinnikov and Skirmantas Janusonis. Toward an effective theory of neurodynamics: Topological supersymmetry breaking, network coarse-graining, and instanton interaction, 2021. URL https://arxiv.org/abs/2102.03849.

Igor Ovchinnikov. Introduction to supersymmetric theory of stochastics. *Entropy*, 18(4):108, March 2016. ISSN 1099-4300. doi: 10.3390/e18040108. URL http://dx.doi.org/10.3390/e18040108.

Brett J. Kagan, Andy C. Kitchen, Nhi T. Tran, Forough Habibollahi, Moein Khajehnejad, Bradyn J. Parker, Anjali Bhat, Ben Rollo, Adeel Razi, and Karl J. Friston. In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron*, 110(23):3952–3969.e8, 2022. ISSN

0896-6273. doi: 10.1016/j.neuron.2022.09.001. URL https://doi.org/10.1016/j.neuron.2022.09.001.

Forough Habibollahi, Moein Khajehnejad, Anirudh Gaurav, and Brett J. Kagan. Biological neurons vs deep reinforcement learning: Sample efficiency in a simulated game-world. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022. URL https://openreview.net/forum?id=... OpenReview.