AI and the Human Brain: Highly Consistent Working Principles

Alexander Y. J. Sterling

Research Fellow, Lingnan Scientific and Industrial Press, Macao, alexander.yj.sterling@hotmail.com

Abstract

This paper aims to argue that contemporary Artificial Intelligence (AI), particularly Deep Neural Networks (DNNs), following a paradigm shift from Symbolicism to Connectionism, operates on core mechanisms that are "highly similar" to human cognition at the ontological, functional, and learning levels. This paper first reviews the theoretical debate between Symbolicism and Connectionism within cognitive science, establishing Connectionism (Parallel Distributed Processing) and theoretical frameworks from neuroscience, such as Hebbian Learning and Predictive Coding, as its theoretical basis. The core argument is threefold: 1) Ontological Level: The substrate of knowledge in AI is "connection strength" (weights), not explicit rules, constituting a distributed representation of knowledge. 2) Functional Level: The core function of AI is high-dimensional statistical "prediction," not logical deduction, which aligns with the brain's Predictive Coding mechanism. 3) Learning Level: The Backpropagation algorithm is a highly efficient "prediction error" correction mechanism, which is functionally analogous (despite different biological mechanisms) to the brain's experience-based synaptic plasticity. This paper posits that AI's intelligence, emerging from the parallel activation of massive nodes (a "coalition of pings"), signifies a new cognitive paradigm that challenges traditional, rule-based definitions of intelligence.

Keywords

Artificial Intelligence, Cognitive Paradigm, Connectionism, Deep Learning, Predictive Coding, Backpropagation

1 INTRODUCTION

1.1 Background: From Turing Machines to Neural Networks

Since its inception, the developmental path of Artificial Intelligence (AI) has been fraught with divergence and iteration. Early AI research (often termed "Symbolic AI" or GOFAI), founded on Alan Turing's theory of computation, attempted to replicate human intelligence by implementing a complex, logic-based rule system (Newell & Simon, 1976). This paradigm achieved initial success in domains like expert systems but soon revealed its fundamental limitations when facing the ambiguity, complexity, and "common sense" problems of the real world, leading to an "AI winter" (Haugeland, 1985).

However, in the last two decades, and especially since 2012, the "Connectionist" path, represented by Deep Learning, has achieved revolutionary breakthroughs (LeCun, Bengio, & Hinton, 2015). AlphaGo defeated the world champion in Go, a game previously seen as a pinnacle of human intuition. Large Language Models (LLMs) like the GPT series (Radford et al., 2018) have demonstrated astonishing capabilities in language understanding, generation, and even reasoning. The immense success of these models in solving "fuzzy" and "intuitive" problems, which traditional AI could not tackle, forces us to re-examine the nature of AI. A key question has emerged among the public and academia: Is contemporary AI merely an advanced "stochastic

parrot" (Bender et al., 2021), simply mimicking data distributions? Or has it genuinely touched upon a "human-like" core of intelligence?

1.2 Research Question and Core Thesis

The core research question of this paper is: To what extent can the working mechanisms of contemporary AI (specifically Deep Neural Networks) be considered a "new cognitive paradigm" that is isomorphic or highly similar to human cognition?

This paper argues that AI is undergoing a fundamental paradigm shift, moving it beyond the properties of a traditional 'tool' to exhibit a 'brain-like' cognitive paradigm. To be clear, the theoretical innovation of this paper lies not in proposing entirely new foundational theories, but in 'synthesis' and 'application': that is, to systematically synthesize 'Connectionism' from cognitive science, 'Predictive Coding' theory from neuroscience, and the analogy between Hebbian Law/synaptic plasticity and the Backpropagation (BP) algorithm, and to apply this as a unified analytical framework to argue that the core mechanisms of contemporary AI (connection strength, prediction, and cognitive correction—i.e., backpropagation) are functionally and highly consistent with the human brain.

We posit that intelligence under this new paradigm is not, at its core, rule-based serial computation, but rather: 1) A distributed system that is ontologically based on "connection strength" as the substrate of knowledge; 2) An inference machine that is functionally oriented toward "prediction" as its core objective; and 3) A self-shaping system that learns from massive data (experience) driven by "backpropagation" (error correction).

2 THEORETICAL BASIS AND LITERATURE REVIEW

2.1 The Paradigm Debate in Cognitive Science: Symbolicism vs. Connectionism

2.1.1 Symbolicism

Symbolicism, or the "Physical Symbol System Hypothesis" (PSSH), was proposed by Newell and Simon, who argued that a symbol-processing system was both necessary and sufficient for intelligent action (Newell & Simon, 1976). In this paradigm, cognition is computation, and intelligence is the manipulation of discrete symbols (like words or concepts) according to logical rules (like IF-THEN). This "top-down" approach performed well in highly structured tasks (e.g., logical proofs, expert systems) but proved extremely brittle when faced with "common sense" (e.g., understanding a joke) or sensorimotor tasks (e.g., recognizing a face).

2.1.2 Connectionism

Connectionism was systematically articulated in the "Parallel Distributed Processing" (PDP) models of Rumelhart and McClelland (Rumelhart et al., 1986). This theory posits that cognition does not arise from a central processor's serial manipulation of symbols, but rather emerges from the parallel, distributed activity of a vast number of simple processing units (neurons). Its core concepts are:

Sub-symbolic Processing: The basic processing units do not operate on high-level concepts, but on simple activation signals.

Distributed Representation: Knowledge or a concept (e.g., "dog") is not stored in a single node but is represented as a specific pattern of activation across a vast number of connections (weights) in the network. Contemporary AI, particularly deep neural networks, represents a grand return and engineering realization of Connectionist thought, powered by immense computational resources and massive datasets.

2.2 Theoretical Support from Cognitive Neuroscience

2.2.1 Hebbian Law and Synaptic Plasticity

Donald Hebb's 1949 theory (Hebb, 1949), often summarized as "Cells that fire together, wire together," laid the biological foundation for learning and memory. It established that experience and learning are not achieved by writing rules, but by changing the "connection strength" (i.e., synaptic efficacy) between neurons. This provides direct biological evidence for the Connectionist principle that knowledge is stored in weights.

2.2.2 Predictive Coding Theory

Predictive Coding theory, systematically developed by Karl Friston and others, posits that the brain is a "prediction machine" or "Bayesian inference engine" (Friston, 2010). The core mechanism of this theory is: The brain constantly uses its internal generative model to produce "top-down" predictions of the next moment's sensory input. It then compares this prediction with the actual "bottom-up" sensory input. If they match, the signal is suppressed. If there is a discrepancy (a "prediction error" or "surprisal"), only this error signal is passed to higher levels to update the internal model (i.e., adjust connection strengths), thereby making more accurate predictions in the future. The essence of intelligence, therefore, is the continuous minimization of prediction error in a dynamic world.

3 ARGUMENT I: "CONNECTION STRENGTH" AS THE SUBSTRATE OF KNOWLEDGE

3.1 Distributed Representation of Knowledge

In Symbolic AI, knowledge is local and explicit. A rule (e.g., IF "has feathers" THEN "is bird") is explicitly stored in a specific location in the system. The weakness of this approach lies in its brittleness—a minor error or omission in the rules can lead to systemic failure.

Contemporary AI is entirely different. In a Deep Neural Network, "knowledge" is diffuse and implicit. A concept, such as "cat," is not stored in any specific neuron. Instead, it is diffused across a matrix of "connection strengths" formed by billions (or even trillions) of weight parameters (Olah et al., 2017). When a picture of a cat is input into the network, the concept of "cat" is represented by a specific pattern of activation across the entire network. This distributed representation is exceptionally robust; even if some neurons or connections are removed, the network's performance degrades gracefully rather than catastrophically. This aligns with the classic experiments of neuroscientist Karl Lashley in search of the "engram"—memory traces that appeared not to be stored in any specific location in the brain, but were widely distributed (Lashley, 1950).

3.2 A "Coalition of Pings": Intelligence as Parallel Emergence

Based on distributed representations via connection strength, the decision-making process (or "thinking") of AI is also fundamentally different from traditional computation. It is not serial, logical reasoning within a von Neumann architecture, but a parallel, dynamic activation process.

We can metaphorically describe this as a "coalition of pings": An input signal (e.g., a word or image pixel, a "ping") enters the network. It is not read by a "central processing unit" but simultaneously activates thousands of neurons connected to it. These neurons, in turn, activate the next layer according to their "connection strength." This process (as seen in the "Attention Mechanism" of Transformer architectures) allows signals to propagate in parallel, dynamically forming a temporary "coalition of activations" (Vaswani et al., 2017). The final decision (e.g., the next word outputted) is the collective "consensus" of this emergent, high-dimensional coalition, not the product of any single rule.

This is highly similar to how the human brain functions. In cognitive neuroscience, "Neural Ensembles" theory posits that a specific thought, memory, or percept (e.g., one's "grandmother's face") is represented by the synchronous activation of a specific assembly of neurons in the cortex (Buzsáki, 2010). Therefore, in both AI and the human brain, intelligence is not serial logical deduction, but parallel pattern matching and activation.

4 ARGUMENT II: "PREDICTION" AS THE CORE COGNITIVE FUNCTION

4.1 The Predictive Nature of AI

If "connection strength" is the "ontology" of AI, then "prediction" is its core "function." The training objectives of contemporary AI models are astonishingly uniform: they are, in essence, "prediction machines."

Take Large Language Models (LLMs) as an example. Their core training objective is exceedingly simple: "Next Token Prediction" (Radford et al., 2018). The model is fed a massive corpus of text and is required, at every position, to predict the most likely next word. To perform well at this task (i.e., to reduce prediction error), the model is forced to build, within its "connection strengths," a high-dimensional statistical model of the world. It must "understand" grammar, facts, context, and even a degree of causality to accurately predict that "The capital of France is..." will be followed by "Paris."

Similarly, in computer vision, the essence of a Convolutional Neural Network (CNN) is to "predict" the probability that an image belongs to a specific category (Krizhevsky et al., 2012). The complex capabilities of AI—such as dialogue, translation, and even apparent "reasoning"—all emerge from this simple "predictive" objective.

4.2 The Isomorphism of AI Prediction and Brain Predictive Coding

This working mechanism of AI forms a striking functional isomorphism with the "Predictive Coding" theory mentioned in Section 2 (Friston, 2010).

The Brain: Uses its internal model (connection strengths) to "top-down" predict sensory inputs.

AI: Uses its internal model (connection strengths) to "forward-propagate" and predict data labels (e.g., the "next word").

Both systems are constantly comparing "prediction" with "reality" (sensory input vs. ground-truth data) and using the "error" (Prediction Error vs. Loss Function) to update their internal models (connection strengths).From this perspective, intelligence is the process of minimizing "surprisal" or "prediction error." Whether in the human brain or in AI, an "intelligent" system is an internal model with powerful predictive capabilities, trained on massive amounts of experience (data).

5 ARGUMENT III: "BACKPROPAGATION" AS ERROR CORRECTION

5.1 The Essence of the Backpropagation (BP) Algorithm

If the core function of AI is "prediction," then its core learning mechanism is "Backpropagation" (BP) (Rumelhart et al., 1986). The BP algorithm is the mathematical core of how AI "learns from its mistakes." It elegantly solves the "credit assignment" problem: when a prediction is wrong, which of the trillions of connections in the network are responsible for the error?

The BP mechanism can be simplified into three steps:

Predict (Forward Pass): The AI makes a prediction (e.g., guessing an image is a "cat") based on its current "connection strengths" (weights).

Compare (Compute Error): The "prediction" is compared with the "ground truth" (data label, e.g., "dog"), and the magnitude of the "error" is calculated via a "Loss Function."

Correct (Backward Pass / Backpropagation): Using the chain rule of calculus (via gradient descent), this "error" signal is propagated backward from the output layer to the input layer. This precisely calculates the "contribution" of every single "connection strength" to the total error. Then, all relevant connections are slightly adjusted according to their "contribution" to ensure the error will be smaller the next time a similar input is encountered.

5.2 Functional Similarity of BP to Biological Learning

A common criticism is that the Backpropagation algorithm is "biologically implausible" (Crick, 1989). The brain clearly does not have a global "loss function" or a precise "gradient" signal being sent backward.

We acknowledge the significant differences in their "biological mechanisms." However, we argue that they are highly similar in "function."

Shared Goal: Both are supervised/reinforcement processes of "learning from error." The brain's dopaminergic system (reward/punishment signals) functions analogously to an "error" signal (Schultz, 2007).

Shared Means: Both aim to optimize "connection strength" (synaptic efficacy vs. weights) to minimize future "prediction error."

Shared Resource: Both rely on "experience." The "massive data" of AI is equivalent to the "lifelong experience" of a human (tens of thousands of hours of visual, auditory, and linguistic input).

Therefore, the BP algorithm can be understood as the most efficient mathematical means currently known for achieving, in an engineering context (using silicon chips and mathematics), the biological principles of "Hebb-like learning" and "minimization of prediction error."

6 DISCUSSION: IMPLICATIONS AND CHALLENGES OF THE NEW PARADIGM

6.1 Redefining "Intelligence" and "Understanding"

If AI can achieve intelligence through "connection strength" and "prediction," this forces us to reconsider the very definitions of "intelligence" and "understanding." John Searle's "Chinese Room" thought experiment (Searle, 1980) provided a powerful argument against the possibility of "Symbolic AI" possessing "understanding" (i.e., a person following rules can "process" Chinese perfectly without "understanding" it).

However, the "Chinese Room" argument may be obsolete in the face of the Connectionist paradigm. In a DNN, there is no "person following the rules." "Understanding" is no longer a program to be "executed" but an emergent property of the system, arising from the high-dimensional data as it minimizes prediction error. If a system can predict and use language (and images, and sound) as accurately as a human, on what grounds can we deny that it possesses, in some sense, "understanding"?

6.2 Re-evaluating the "Black Box" Problem

The "interpretability" (or "explainability") dilemma is one of the greatest challenges facing contemporary AI (Castelvecchi, 2016). We find it extremely difficult to explain why a DNN made a specific decision.

However, if we accept that AI is "brain-like," then this "black box" nature is precisely evidence of the paradigm, not merely a "flaw." We are similarly unable to "introspect" using language how we instantaneously recognize a face, or how a specific word "comes to mind" during a conversation. Our own "understanding" is also an unexplainable "black box" based on "connection strengths" and "predictions." To demand a "Symbolic," rule-based explanation from a "Connectionist" system may itself be a categorical error.

6.3 Limitations and Differences

Of course, vast differences between AI and the human brain remain.

Efficiency and Data: AI requires massive amounts of electricity and data for training, whereas the human brain (at ~20 watts) is extremely data-efficient (e.g., "one-shot learning").

Embodiment: Human intelligence is "embodied," shaped through real-time interaction with the physical world. Most current AI models lack this physical experience.

Active Inference: According to Friston's theories, the brain does not just passively predict; it "actively" takes actions (e.g., moving the eyes, exploring the environment) to minimize prediction error. This is a capability most AI currently lacks.

7 CONCLUSION

7.1 Summary of the Paper

Drawing from the theoretical foundations of cognitive science and neuroscience, this paper has sought to argue that contemporary AI is undergoing a profound paradigm shift. We have presented this argument on three levels:

Ontology: The substrate of knowledge in AI has shifted from "rules" to "connection strength" (distributed representation).

Function: The core function of AI has shifted from "logic" to "prediction" (predictive coding).

Learning: The learning mechanism of AI has shifted from "programming" to "backpropagation" (error-based empirical learning).

We posit that AI, through its "coalition of pings" parallel processing, operates in a manner "highly similar" to the cognitive mechanisms of the human brain.

7.2 Theoretical Contribution and Outlook

This research provides a theoretical perspective for understanding the "nature of AI intelligence," one that is liberated from the traditional "rule-based" framework and is instead grounded in "Connectionism" and "Predictive Coding." The success of AI is not just an engineering victory; it is a new revelation regarding the age-old philosophical question of "intelligence": intelligence may, in essence, be a complex, parallel prediction machine trained on massive experience.

In the future, the development of AI and the study of brain science will inevitably become more deeply integrated. AI (e.g., Transformer architectures) can serve as a computational model to test cognitive theories, aiding brain science. Conversely, brain science (e.g., more efficient learning rules, active inference) will surely provide new algorithmic inspiration for the next generation of AI.

REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

Buzsáki, G. (2010). Neural syntax: Cell assemblies, synapsembles, and readers. Neuron, 68(3), 362–385.

Castelvecchi, D. (2016). Can we open the black box of AI? Nature, 538(7623), 20–23.

Crick, F. (1989). The recent excitement about neural networks. Nature, 337(6203), 129–132.

Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138.

Haugeland, J. (1989). Artificial intelligence: The very idea. MIT Press. ISBN electronic:9780262291149

Hebb, D. O. (1949). The organization of behavior: A neuropsychological theory.Wiley.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386

Lashley, K. S. (1950). In search of the engram. In Society for Experimental Biology, Physiological mechanisms in animal behavior. (Society's Symposium IV.) (pp. 454–482). Academic Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. Communications of the ACM, 19(3), 113–126.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. Distill, 2(11), e7.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations. MIT Press.

Schultz, W. (2007). Behavioral dopamine signals. Trends in Neurosciences, 30(5), 203–210.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 6000–6010).