Scientific Research Publishing

Open Journal of Philosophy, 2025, 15(4), 1032-1047
https://www.scirp.org/journal/ojpp
ISSN Online: 2163-9442
ISSN Print: 2163-9434

The Cognitive Mechanism and Limitations of Deep Learning from the Perspective of the Three-Level Cognitive Theory

Jianwei Sun

School of Philosophy, Beijing Normal University, Beijing, China
Email: sjwneu@icloud.com

How to cite this paper: Sun, J. W. (2025). The Cognitive Mechanism and Limitations of Deep Learning from the Perspective of the Three-Level Cognitive Theory. *Open Journal of Philosophy, 15,* 1032-1047.
https://doi.org/10.4236/ojpp.2025.154062

Received: October 30, 2025
Accepted: November 18, 2025
Published: November 21, 2025

Abstract

Building on the Three-Level Theory of Cognition, this paper examines the architecture and foundational principles of deep learning in order to clarify its specific cognitive mechanisms and the central epistemological challenges it poses. The cognitive efficacy of deep learning derives from the synergistic interaction of three distinct levels: computational theory defines its objectives, algorithms specify the methods for achieving them, and hardware serves as the physical substrate for algorithmic implementation. Together, these levels jointly drive the advancement of deep learning. Meanwhile, the current challenges faced by deep learning, namely the lack of interpretability and theoretical understanding, also constitute the core epistemological issues in machine learning. By identifying these constraints at each hierarchical level, this paper highlights where future research must focus to deepen our understanding of deep learning's cognitive capabilities and to resolve its foundational epistemic problems.

Keywords

Three-Level Cognitive Theory, Deep Learning, Epistemology, Black-Box Dilemma, Crisis of Understanding

1. Introduction

The machine cognition revolution driven by Deep Learning (DL) has exerted a broad impact on cognitive science, philosophy, technological ethics, and other fields, thereby sparking discussions on the cognitive mechanisms of deep learning. Although deep learning has achieved breakthrough progress in numerous domains such as visual recognition and natural language processing, with some in-

dicators approaching or even surpassing human levels, the issues of insufficient interpretability and lack of theoretical understanding in deep learning have become increasingly prominent. To a certain extent, these problems have restricted the further development of artificial intelligence, especially in key fields involving ethics, morality, and life safety. Addressing these two intertwined problems first requires a systematic exploration of the cognitive mechanisms of machine learning. Traditional cognitive science mostly adopts "information processing" or "symbolism" as frameworks, and has fully explained the mechanisms of deep learning from the perspective of symbolic representation and computation (Fodor & Pylyshyn, 1988). The Three-Level Cognitive Theory further provides us with a hierarchical and interpretable perspective. This paper investigates the cognitive mechanisms of deep learning and the developmental limits it encounters by examining three interrelated strata: computation, algorithm, and implementation.

Structure of the Paper:

The discussion proceeds in four parts:

1) Review the current state of deep learning, highlighting its major achievements and persistent challenges.

2) Introduce the Three-Level Theory of Cognition and apply this framework to analyze the cognitive mechanisms underlying deep learning.

3) Examine the core epistemological questions in machine learning from the perspectives of computation, algorithm, and implementation.

4) Summarize the cognitive effectiveness and limitations of deep learning under the Three-Level Cognitive Theory.

By integrating insights from cognitive theory with contemporary deep learning research, the paper aims to advance a more transparent and theoretically grounded understanding of machine cognition, thereby informing both scientific inquiry and the responsible deployment of AI technologies.

2. Achievements and Difficulties of Deep Learning

2.1. Deep Learning Achievements

The breakthrough progress of deep learning originated from the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The deep neural network model (AlexNet), developed by Geoffrey Hinton and his two students, Alex Krizhevsky and Ilya Sutskever, won the championship with an overwhelming advantage in this competition. This achievement marked a qualitative leap in machine vision capabilities: it not only completely reversed people's perceptions of the practicality of deep learning but also drove the explosive growth of deep learning. Over the past decade, deep learning has penetrated almost all fields of information technology, enabling breakthrough progress in artificial intelligence research across numerous domains. Today, machines can reach or even surpass human-level performance in areas such as image recognition, natural language processing, reinforcement learning, and scientific discovery. **Table 1** lists the achievements and performance of deep learning in various fields.

Table 1. Achievements and performance of deep learning in various fields.

Technical Field	Deep Learning Models	Achievements and Performance
Visual Recognition	ResNet (Residual Learning-Based Network)	1. Achieved superhuman-level performance in the ImageNet Challenge, reducing the top 5 image recognition error rate to below 3%. 2. Reached or exceeded expert diagnostic accuracy in fine-grained classification (e.g., birds, plants) and medical imaging (lung CT, breast X-ray).
Natural Language Processing	ChatGPT, DeepSeek	1. Capable of generating coherent and logically consistent long-form text. 2. Efficiently accomplish tasks such as machine translation, text summarization, and dialogue. 3. Some metrics approach or even surpass human levels.
Reinforcement Learning & Adversarial Games	AlphaGo, AlphaZero	1. Mastered the rules of complex games (e.g., Go, chess) through self-play, relying solely on computing power and strategy search. 2. Demonstrated end-to-end learning capabilities from raw perception to high-level strategy.
Scientific Discovery & Engineering Design	AlphaFold, GNoME, MatterGen	1. AlphaFold predicts the 3D structure of proteins with experimental-level accuracy, significantly accelerating the progress of new drug development. 2. GNoME and MatterGen provide new search methods for the inverse design of material microstructures.

These achievements indicate that deep learning has evolved into a powerful tool capable of abstract reasoning, creative generation, and even interdisciplinary scientific discovery.

2.2. Difficulties Faced by Deep Learning

Despite the remarkable achievements made by deep learning, it still confronts two fundamental challenges: the issue of model interpretability (Black-Box Dilemma) and the lack of theoretical understanding (Crisis of Understanding).

1) The Black-Box Dilemma

The Black-Box Dilemma refers to a phenomenon where the internal cognitive processes and logical reasoning paths of an intelligent system are difficult for humans to clearly understand, trace, and explain when the system makes decisions or generates results. It is analogous to a "closed box": we can only observe the inputs (e.g., data, instructions) and outputs (e.g., judgments, answers), but cannot gain insight into the internal operational mechanisms—specifically, which exact rules, principles, or features the system relies on to derive its results. The root causes of the Black-Box Dilemma are partly due to the mismatch between the complexity of the model structure and human cognitive capabilities, and partly due to the model's

training data and optimization objectives.

The black-box nature of the system directly leads to significant implications in terms of safety, reliability, and legality. For instance, in high-risk fields such as medical diagnosis, autonomous driving, and financial credit scoring, if a model produces errors or biases, the lack of an interpretable internal mechanism makes it difficult for researchers to quickly identify the root cause of the failure, resulting in a sharp increase in error-correction costs. Additionally, regulatory authorities struggle to determine whether decisions comply with fairness and legal requirements. The resulting crisis of trust will further hinder the development of technology (von Eschenbach, 2021).

2) The Crisis of Understanding

The Crisis of Understanding refers to the fact that although intelligent systems have achieved performance exceeding that of humans in tasks such as image recognition, language generation, and game competition, we cannot confirm whether these systems truly "understand" the content they process—including concepts, causal relationships, or semantic structures within it. Developers of deep learning no longer need to manually encode knowledge; instead, they allow neural networks to automatically extract features from massive amounts of data. The high accuracy of models is often built on capturing statistical correlations in massive datasets, rather than on abstract, interpretable internal representations of the world.

Consequently, when models encounter distribution shifts, adversarial examples, or require cross-task transfer, their behaviors become unpredictable and fragile, and may even produce results that severely conflict with human common sense. Each layer in a deep learning network learns an abstract representation, and the disconnect between these representations (which are difficult to describe in human language) and real-world semantics, causality, and common sense is the cause of the Crisis of Understanding. Therefore, the essence of the Crisis of Understanding lies in the fact that machines lack the ability for in-depth cognition of "semantics, logic, common sense, and context," and cannot establish connections between "symbols and real-world meanings" in the same way humans do.

The Crisis of Understanding is not only a technical bottleneck but also a fundamental obstacle to trust, regulation, and interdisciplinary integration. In terms of safety and reliability, the lack of understanding at the causal and conceptual levels makes systems prone to failure when facing unseen scenarios. In terms of ethics and legality, if the decision-making basis of a model is unexplainable, it becomes difficult to audit for biases, discrimination, or attribution of responsibility. In terms of scientific progress, researchers struggle to reverse-infer the essence of intelligence from the behaviors of existing models, thereby hindering the interdisciplinary development of artificial intelligence with cognitive science, neuroscience, and other disciplines.

To address the black-box problem and interpretability challenges of AI, we must fundamentally re-examine learning objectives, model architectures, and their implementation paradigms. Explainable Artificial Intelligence (XAI) provides vis-

ual explanations for model outputs by constructing global or local interpreters and unifying them with feature importance. Causal machine learning leverages structural causal models, invariant risk minimization (IRM), and causal graph inference to reveal the causal relationships between variables while maintaining predictive performance, thereby enabling logical tracing of decision-making processes and identification of intervention potential (Xu, Ge, & Zhang, 2023). XAI offers intuitive and understandable explanations, while causal methods lay a more solid theoretical foundation for reasoning; their combination can significantly enhance the causal consistency and credibility of explanations. Therefore, only by enabling models to proactively capture causal relationships during training and form intermediate representations aligned with human cognitive concepts can we potentially maintain strong statistical performance while achieving high interpretability and robust transferable understanding capabilities.

3. Deep Learning under the Three-Level Cognitive Framework

3.1. The Three-Level Cognitive Theory

The hierarchical concept of human cognitive processes generally refers to how a system processes information through a hierarchical structure—from low to high levels and from concrete to abstract—to achieve perception, understanding, reasoning, and decision-making regarding the environment. This hierarchy is not a simple division of steps; instead, each level independently performs specific functions while forming an integrated cognitive capability through dynamic interaction. Its core logic lies in decomposing complex cognitive tasks into collaborative subtasks and reducing cognitive load through hierarchical progression. With the development of neuroscience, cognitive science, and computer science, the theory of cognitive levels has also continued to evolve, resulting in the academic community not yet reaching a complete consensus on the definitions of its three levels. The three levels referred to in this paper are specifically: computational theory, representation and algorithm, and physical implementation.

Simon and Newell compared human cognitive activities with the cognitive processes of computers. He argued that human cognitive activities could be studied from three levels or approaches: The highest level consists of thinking activities, which express goals, desires, or beliefs. The middle level involves primary information processing, where symbolic representations implement methods to achieve the goals of the upper level (e.g., through state spaces, heuristic search, or goal analysis). And the lowest level encompasses physiological processes, i.e., the activities of the central nervous system, neurons, and the brain. By analogy, the highest level of a computer corresponds to its program, while the middle and lowest levels correspond to computer language and computer hardware, respectively (Simon & Newell, 1971).

Marr proposed that a complete understanding of an information-processing device requires examining it at different levels:

1) The top level is computational theory, which focuses on the goals of compu-

tation, the rationale for its validity, and the logical basis underlying the strategies used to perform the computation.

2) The middle level involves representation and algorithm, which addresses how to implement the computational theory in practice. It defines the representational form of input and output information (i.e., how data is encoded) and the specific algorithms for converting input to output (i.e., operational steps). This level is directly linked to the selection of input-output representation methods and the design of algorithms that support information conversion.

3) The lowest level concerns computer architecture and physical implementation details, focusing on how algorithms and representations are realized at the physical level, i.e., how abstract algorithms and representations are transformed into executable practical operations through physical carriers such as hardware and circuits (Marr, 1982).

Pylyshyn argued that explaining cognitive behavior requires attention to three distinct levels of a system: the nature of the functional architecture or mechanism, the nature of symbolic structures, and their semantic content. In cognitive science, these three natures are fundamental characteristics of the computational view of the mind. Pylyshyn's perspective echoes Marr's Three-Level Theory of computation-algorithm-implementation while providing further analysis from the dimension of attention-index-language (Pylyshyn, 1986).

3.2. Cognitive Analysis of Deep Learning

Deep learning is an end-to-end mapping that eliminates the need for manually designed features or formalized rule definitions. It acquires model parameters through data-driven methods and replicates the processes executed within the mechanism being imitated—specifically, the input and output encoding of the model aligns with that of the imitated mechanism. Pylyshyn argued that this input-output alignment indicates the model is formally equivalent to the processes occurring in the imitated mechanism: the model and the mechanism it imitates compute the same function, making them equivalent at the computational level.

Advancements in deep learning have strengthened the concept of "cognition as computation." However, unlike the discrete symbolic representations of computationalism, deep learning relies on distributed representations acquired through learning. The fundamental architecture of deep learning is the Deep Feedforward Network, also known as the Multilayer Perceptron (MLP), a critical foundational model architecture in deep learning. Its core is a hierarchically connected chain-like network topology, consisting of an input layer, one or more hidden layers, and an output layer. Neurons between layers are connected via weight matrices, with information transmitted unidirectionally from input to output.

The function of each layer can be expressed as a mathematical formula: $y = f(x)$, where the input x is derived from the output of the previous layer, and the output y serves as the input for the next layer. Thus, the expression for a multilayer feedforward network takes the form of a chained composition of functions. For

example, a three-layer feedforward network can be written as $y = f^{(3)}(f^{(2)}(f^{(1)}(x)))$, where $f^{(1)}(x)$ represents the function of the first layer, $f^{(2)}(x)$ the second layer, and so on. This chained function composition effectively reflects the "depth" of the model and its computations—hence the name "deep learning".

Cognitive Mechanism of Deep Learning: an examination built on the Three-Level Cognitive Theory. To analyze the cognitive mechanism of deep learning using the Three-Level Cognitive Theory, we examine it across three dimensions: computational theory, representation and algorithm, and physical implementation.

1) Computational Theory

The core value of computational theory lies in its ability to reveal the computational essence underlying cognitive processes such as perception. Its relevance depends on the computational problem itself that needs to be solved, rather than the algorithms used to implement the computation or the specific hardware. Instead of fixating on the existence of effective algorithms or executable hardware, understanding the nature of the problem to be solved is more conducive to identifying appropriate algorithms.

A classic example is the impact of the Recognition by Components Theory (RBC) on the development of computer vision technology (Biederman, 1987). The RBC theory posits that humans achieve pattern recognition by decomposing the structure of complex objects into simple basic geometric components (called "geons") and their spatial relationships, then automatically matching these components to representations stored in memory. Geons include basic shapes such as blocks, cylinders, wedges, and cones. Five spatial relationships—curvature, collinearity, symmetry, parallelism, and contamination—are derived from contrasting edge features that are easily detectable in 2D images. These properties are generally invariant to viewing position and image quality, enabling robust object perception even when the image is projected from a new viewpoint or degraded.

The core logic of the RBC theory—decomposition into basic components, followed by component combination and template matching—not only directly inspired early technologies such as edge detection, part-based models, and view-invariant features but also profoundly influenced the development of feature extraction mechanisms in deep learning. From early algorithms with manually designed features to modern deep learning-based models, the ideas of the RBC theory have permeated core tasks such as object recognition and target detection in various forms, driving the evolution of computer vision from imitating human visual mechanisms to realizing efficient recognition functions.

The significance of computational theory lies in guiding us to transcend the limitations of implementation details and construct cognitive frameworks based on the essence of problems (Goodfellow, Bengio, & Courville, 2016). This enables low-error function approximation for high-dimensional, noise-rich natural inputs—an essential foundation for understanding complex computational processes, whether in perception (natural intelligence) or information processing (ar-

tificial systems).

2) Representation and Algorithm

Representation and algorithms serve as the bridge connecting computational theory to physical implementation. Representation refers to the abstract form in which a system encodes data, knowledge, and task objectives. It is a mechanism that explicates specific types of information from signals, with its core role being to convert complex real-world information into a format processable by machines. For instance, an image can be represented using pixel-based encoding (directly as a sequence of pixel color values) or feature-based encoding (extracting local or global image features to convert raw pixels into more representative descriptor units for image analysis and recognition). The choice of representation format clarifies certain types of information while hiding others—and hidden information becomes harder to retrieve.

Algorithms encompass the steps and rules for processing representations to achieve objectives. Its core role is to convert information embedded in representations into specific outputs (e.g., classification results or action decisions). The efficiency and generalization ability of an algorithm directly determine whether a system can be applied in real-world scenarios.

Representation and algorithm do not exist in isolation; they are interdependent and coevolve. A well-designed representation simplifies algorithm design, while an efficient algorithm unlocks the potential of the representation. This synergy is critical to every technological breakthrough: representation defines what can be processed, and the algorithm determines how it can be processed. The value of representation lies in converting inputs to outputs—effective representations must capture the essence of things. Algorithms, meanwhile, are the driving force behind converting inputs to outputs; without algorithmic breakthroughs, even the best representations cannot be translated into practical output capabilities. A typical example is that before the advent of the backpropagation algorithm, neural networks remained merely theoretical models.

3) Physical Implementation

Finally, representation methods and algorithms must be implemented in physical devices—hardware. As the physical carrier for cognitive implementation, hardware development directly determines the feasibility, efficiency, and application boundaries of algorithms, serving as a core driver for translating theory into practice. From early general-purpose processors to specialized chips, hardware iterations have not only supported increases in algorithm complexity but also reshaped research directions and application scenarios.

The performance of deep learning is highly dependent on data volume and model scale, yet training and running large-scale models require enormous computing power. The shift of deep learning from small to large models is driven precisely by improvements in hardware computing power, and increased computing power, in turn, has further promoted the expansion of algorithm scale. From AlexNet to GPT-4, the synergetic evolution of hardware and algorithms has un-

derpinned every major advancement.

4. Core Epistemological Issues of Machine Learning

Within the framework of traditional epistemology, knowledge is regarded as a reliable belief formed through conceptualized reasoning, empirical induction, and theoretical verification. However, the learning mechanism of deep neural networks is a data-driven form of extreme induction. Instead of deriving answers through deductive reasoning, models search for points that minimize empirical risk in the parameter space by statistically fitting massive samples. This process is essentially a form of statistical induction—extracting statistical regularities from observed samples and extrapolating them to unseen scenarios.

From the perspective of cognitive science, this learning method, centered on statistical correlations, stands in stark contrast to the human mechanism of concept formation. Human cognition often relies on interpretable prototypes, hierarchical concept networks, and causal models; these structures enable us to perform concept transfer and reasoning with only a few examples or even in zero-shot scenarios. In contrast, the internal representations of deep neural networks are highly distributed vector spaces, where the activation of a single neuron does not correspond to a specific concept but rather a mixture of several abstract features. Although this form of representation is highly efficient in pattern matching, it lacks composability and causal interpretability emphasized in cognitive science.

Only by establishing closer interdisciplinary connections between computational theory, algorithm design, implementation methods, and philosophical reflection can machine learning be endowed with a more reliable and interpretable cognitive status at the epistemological level. The Three-Level Cognitive Theory provides us with such a perspective to analyze the core issues faced by deep learning in terms of epistemology and cognitive science.

4.1. From the Perspective of the Computational Level

From the perspective of the computational level, the fundamental question raised at the level of computational objectives is: "What exactly do we want machines to learn?" If the objectives themselves fail to align with our knowledge structure of the real world, no matter how sophisticated the algorithms or how powerful the hardware, the output of learning will still be "misaligned knowledge."

Most machine learning systems aim to minimize a certain type of error (e.g., cross-entropy or mean squared error) or maximize cumulative rewards. They are trained to perform optimally on the training data set. The essence of this objective lies in statistical fitting: by continuously adjusting model parameters, the distribution of the model's outputs is made to align as closely as possible with the distribution of observed data.

Since the error function itself only measures "the correctness of predictions" and imposes no constraints on the explanations or structures behind those predictions, the learning process merely captures statistical correlations that reduce

loss—without distinguishing whether these correlations are accidental co-occurrences, biases in data collection, or underlying causal mechanisms. The narrowness of computational objectives is a root cause of the "understanding crisis."

In human cognition, "understanding" means being able to map perceptual inputs to a semantic network composed of causal, functional, conceptual, and contextual information. This network supports cross-contextual reasoning, explanation, and active learning. While deep learning models can achieve remarkable accuracy on specific data distributions, they often only capture statistical correlations rather than causal structures, and lack composable, transferable conceptual hierarchies.

Because the computational objective is to "minimize error rates on the training set", not to construct internal representations that explain the world at the level of human-understandable concepts, these systems exhibit behaviors that seem to reflect understanding yet lack genuine semantic interpretive capabilities. The narrowness of computational objectives manifests in several ways:

First, statistical fitting is extremely sensitive to data. If the training set provides sufficient signals to reduce error, the model will tend to exploit any available correlated features—even if those features have no semantic value. For example, in facial recognition experiments, a network might learn to use background lighting, camera model, or even image compression noise to distinguish between different individuals. While these features are statistically effective, they have nothing to do with humans' essential understanding of the concept of a "human face."

Second, learning without causal mechanisms makes models highly vulnerable to distribution shifts or adversarial perturbations. Without an internal causal graph or structured world model, models can only make inferences within the boundaries of known statistical distributions. Once the statistical properties of inputs deviate from the training distribution, the model's outputs quickly lose reliability—failing rapidly when faced with distribution shifts, adversarial examples, or complex real-world scenarios. This can lead to unforeseen errors in safety-critical contexts and unexplainable biases in ethical audits.

4.2. From the Perspective of the Algorithmic and Representational Level

From the perspective of the algorithmic and representational level, the key question at the algorithmic level is: "What methods should be used to transform objectives into knowledge?" Inductive biases, optimization paths, and learning paradigms determine the form, structure, and reliability of knowledge. If the assumptions of an algorithm do not match the causal structure of the real world, the resulting model can only provide "surface-level knowledge" based on statistical correlations.

The "black box" nature of deep learning models is rooted in the inherent opacity of their algorithmic design. Modern neural networks perform nonlinear mappings in high-dimensional spaces using tens of millions of parameters. The learn-

ing objective is merely to minimize empirical risk, with no requirement that the organization of parameters adheres to any interpretable semantic constraints.

The process by which gradient descent searches for optimal solutions in such a vast parameter space is entirely data-driven: it neither generates explicit rule sets nor provides traceable descriptions of the logic behind each weight update. As a result, the internal state of the model is often only represented in the form of abstract tensors, lacking direct correspondence with human language or symbolic systems.

The ambiguity at the algorithmic level is further amplified in practical applications by the difficulty of interpreting distributed representations. The activation vector of each layer is typically a mixed projection of numerous abstract concepts; the response of a single neuron rarely corresponds to an identifiable feature but instead is cross-coupled with multiple semantic dimensions (Lipton, 2016). When these distributed vectors are layered and nonlinearly transformed in deep networks, the final decision depends on the overall geometric structure of the entire high-dimensional space—rather than on separable, interpretable chains of rules.

This makes it impossible to logically trace back the reasoning for individual input instances. Even if we can locate certain salient regions or gradient peaks, we can only determine "which features were amplified," not "why these features are conceptually decisive." The lack of reversible causal chains directly leads to the interpretability crisis: while the model's outputs are statistically reliable, they cannot be supported by human-acceptable causal or conceptual explanations to justify their decisions.

For example, when a Convolutional Neural Network (CNN) is fed an image of the digit "3," the neurons in the first layer recognize edges, the second layer recognizes strokes, the third layer recognizes the shape of the digit, and the final output is the specific digit. However, humans cannot trace which specific parameters determined this final judgment.

The model's reasoning process is the result of nonlinear computations involving an extremely large number of parameters. The nonlinear combination of these parameters leads to high-dimensional, non-convex mappings that cannot be described by simple formulas—far exceeding the human brain's ability to trace and understand. At the algorithmic level, we can only observe that "parameters have converged numerically," but cannot see "which update allowed the network to capture the upper and lower semicircles of the digit '3'." This makes it impossible to base explanations on the learning process.

The opacity at the algorithmic level manifests as follows: in each layer, the compression, mixing, and randomization of information convert original causal factors into internal states that cannot be directly mapped. This prevents external observers (humans) from directly obtaining answers to "why this happened." Training only preserves parameter values—there is no explicit record of "which strokes were learned." Distributed representations are unreadable: each dimension is a mixture of multiple low-level features, with no one-to-one semantic mapping. Visualiza-

tion can only generate heatmaps, not conceptual labels. The decision-making process is non-retraceable: predictions are numerical functions, lacking intuitive judgment rules. Even if activation maps are saved, we can only determine "which channels were activated," not provide interpretable descriptions such as "it is the digit '3' because two semicircles were detected."

4.3. From the Perspective of the Implementation Level

The training process of deep neural networks relies on backpropagation algorithms and large-scale, dense floating-point operations, which demand enormous energy consumption at the hardware level. Backpropagation requires computing gradients for all network parameters after each forward pass and propagating errors accurately between layers—this translates to hundreds of millions of cumulative multiply-accumulate operations during training. Such operations manifest as highly parallel and continuous computation flows, with power consumption primarily allocated to large-scale matrix multiplications and repeated access to convolution kernels.

By contrast, the biological implementation mechanism of the human brain exhibits distinctly different energy consumption characteristics: Synaptic connections between neurons are extremely sparse; the firing rate of a single neuron ranges from a few hertz to several tens of hertz; information transmission occurs via event-driven spiking; and synaptic plasticity is mainly achieved through local Hebbian rules or postsynaptic potential regulation. The entire system only requires a few tens of watts to sustain complete perceptual, motor, and cognitive functions. This sparse, asynchronous, and approximately analog computing method enables biological neural networks to far outperform current artificial neural networks (based on dense floating-point operations) in energy efficiency.

The pre-training process of modern large-scale language models often requires thousands of GPU cores operating for weeks, consuming hundreds of thousands to millions of kilowatt-hours of electricity—equivalent to the daily electricity consumption of a medium-sized city. Energy sustainability has thus become a bottleneck for the scaling of deep learning: As model scales grow exponentially, the demand for computing power and corresponding electricity also rises exponentially, leading to a significant increase in data center carbon emissions. The urgency of global climate change compels us to re-examine this continuous growth in computing power.

From the perspective of implementation pathways, backpropagation itself lacks a direct corresponding mechanism in biology. Although continuous attempts are made to develop new network structures to realize similar gradient transmission in neuromorphic hardware or networks more aligned with brain structures, these methods still rely on the synchronous propagation of global error signals. They struggle to capture the hierarchical self-organization and unsupervised self-supervised learning processes prevalent in the brain. Brain learning is largely achieved through local synaptic plasticity, a reinforced neuromodulator system, and mul-

timodal cross-modal statistical-causal hybrid signals—this mechanism inherently imposes strict constraints on energy consumption. Therefore, the current paradigm of combining backpropagation with dense floating-point computing is neither biologically plausible in terms of energy efficiency nor likely to be feasible in resource-constrained environments.

From the standpoint of implementation materials, data availability directly determines the depth of feature learning and the generalization ability of deep learning models. Large-scale and diverse training datasets can significantly enhance a model's cognitive accuracy and robustness. Conversely, data scarcity or imbalanced distribution may lead to underfitting, diminished transfer performance, and even catastrophic forgetting. More critically, historical inequities and sampling biases embedded within the data can be further amplified by neural networks, resulting in systematic discrimination in predictions. For instance, face recognition technologies often exhibit significantly higher misclassification rates for specific demographic groups. This not only undermines the accuracy of machine cognition but also precipitates a cascade of legal disputes, ethical conflicts, and crises of social trust (Budach, Feuerpfeil, Ihde et al., 2022).

4.4. The Problem of Disconnection between Levels

While the three cognitive levels are governed by distinct principles and each level possesses considerable autonomy, allowing us to describe the laws of each level to the greatest extent without considering how these laws are implemented at lower levels, the emergence of cognitive functions requires coordination across all three levels. Progress confined to a single level is insufficient to achieve complete cognitive capabilities. Coordination among the three levels is of great significance for the development of machine cognition.

The excessive reliance of current artificial intelligence (AI) development on computing power essentially reflects the overshadowing of theoretical rationality by instrumental rationality. The value of philosophical epistemology lies in its reminder that the goal of deep learning is to approach the essence of intelligence, the core of computational theory. The inquiry into the nature of cognition is precisely the "sense of direction" that philosophical epistemology provides for technological progress.

The further development of AI depends first on new computational theories. However, the current situation sees massive resources invested at the physical implementation level, with hardware designs continuously updated to meet the computing power demands of complex models and algorithms. This leads to a cycle: On one hand, hardware iterations constantly cater to the model's hunger for computing power; on the other hand, stimulated by advances in computing power, model scale expands endlessly. Technological development thus requires coordinated alignment between computational theory, algorithms, and physical implementation.

Progress in intelligence has never been a breakthrough confined to a single level,

but rather the result of the co-evolution of computational theory, algorithms, and physical implementation through mutual interaction. Only when cognitive goals, algorithmic goals, and implementation mechanisms achieve conceptual synergy can machine learning potentially bridge the gap from being a "prediction machine" to an "understanding machine"—truly becoming a cognitive tool that aligns with human values and the needs of sustainable development.

5. Conclusion

The Three-Level Cognitive Theory provides a hierarchical and interpretable perspective for investigating the cognitive mechanisms and limitations of deep learning. Within this framework, we systematically examine deep learning from three dimensions: the computational level (the goals to be achieved and abstract functions), the algorithmic level (the processes and representations for realizing these functions), and the implementation level (physical components and resource constraints).

At the computational level, deep learning achieves a statistical mapping goal equivalent to human perceptual tasks, ensuring low error rates on large volumes of natural data. At the algorithmic level, hierarchical distributed representations, gradient-driven adaptive learning, and structural innovations (such as attention and residual mechanisms) enable networks to automatically extract and organize information, forming internal models that support high-level cognitive functions. At the implementation level, powerful parallel hardware, scalable distributed training infrastructures, and mature software ecosystems provide the necessary resources and operability for large-scale learning.

The synergy of these three levels allows deep learning to demonstrate remarkable effectiveness in multimodal tasks such as perception, language processing, reasoning, and control, making it the most empirically supported technical paradigm in contemporary artificial cognitive systems. It is evident that the Three-Level framework helps us move beyond mere empirical reports, clarifying: in what sense deep networks achieve human-like cognitive functions, to what extent they provide operable processing procedures, and under what physical conditions they operate. Precisely this synergy across the three levels has enabled deep learning to achieve unprecedented success in a wide range of task domains, including visual recognition, natural language processing, machine translation, and reinforcement learning.

The Three-Level Cognitive Theory also reveals the cognitive limitations of deep learning. It helps decompose seemingly monolithic problems into constraints across different levels, preventing the conflation of factors from various levels when explaining the cognitive mechanisms of deep networks. This allows for clearer identification of which characteristics stem from limitations in computational theory and goals, which problems arise from structural constraints of the algorithmic model itself, and which biases result from side effects of hardware implementation and training resources.

Only by imposing synchronous constraints across these three levels can we overcome deep learning's current limitations and advance it toward artificial cognitive systems with greater cognitive generality and social acceptability. Examples of such constraints include introducing causal or rule-oriented objective functions at the computational level, by constraining model outputs to satisfy specific logical relationships (such as first-order logic) or align with causal assumptions (e.g., counterfactual reasoning), the learning process can be guided to break through the limitations of pure data-driven approaches. For instance, one approach incorporates temporal logic formulas into neural networks to effectively embed traffic rules within deep learning-based trajectory prediction models (Li, Rosman, Gilitschenski et al., 2020). By incorporating counterfactual information between disentangled concepts into diffusion models, we generate more reliable disentangled cell representations. This approach significantly enhances the interpretability, generalization ability, and controllability of cell data, with performance that outperforms existing black-box representation learning models (Gao, Dong, Shan et al., 2025). Enforcing composable intermediate concepts or attention structures at the algorithmic level; Adopting traceable training logs, low-power hardware, and fair data pipelines at the implementation level. By synchronously integrating mechanisms such as causality, values, multi-objectives, sparse representation interpretation, and energy efficiency constraints across the three levels, deep learning can be made more aligned with the structural, causal, and interpretable characteristics of human cognition.

Acknowledgements

I would like to express my sincere gratitude to Prof. Xue Yonghong and Prof. Dong Chunyu for their valuable guidance and constructive suggestions. I also appreciate the hard work and dedication of all those who participated in the compilation process.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review, 94,* 115-147. https://doi.org/10.1037/0033-295x.94.2.115

Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N. S., Patzlaff, H., Harmouch, H., & Naumann, F. (2022). *The Effects of Data Quality on ML-Model Performance*. https://arxiv.org/abs/2207.14529

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition, 28,* 3-71. https://doi.org/10.1016/0010-0277(88)90031-5

Gao, Y., Dong, K., Shan, C., Li, D., & Liu, Q. (2025). Causal Disentanglement for Single-Cell Representations and Controllable Counterfactual Generation. *Nature Communications, 16,* Article No. 6775. https://doi.org/10.1038/s41467-025-62008-1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

https://www.deeplearningbook.org/lecture_slides.html

Li, X., Rosman, G., Gilitschenski, I., DeCastro, J. A., Vasile, C. I., Karaman, S., & Rus, D. (2020). Differentiable Logic Layer for Rule Guided Trajectory Prediction. In *4th Conference on Robot Learning*. https://tisl.cs.toronto.edu/publication/202011-corl-logic_layer/corl20-logic_layer.pdf

Lipton, Z. C. (2016). The Mythos of Model Interpretability. *Communications of the ACM, 61,* 36-43. https://doi.org/10.1145/3233231

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.

Pylyshyn, Z. W. (1986). *Computation and Cognition: Toward a Foundation for Cognitive Science*. The MIT Press.

Simon, H. A., & Newell, A. (1971). Human Problem Solving: The State of the Theory in 1970. *American Psychologist, 26,* 145-159. https://doi.org/10.1037/h0030806

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology, 34,* 1607-1622. https://doi.org/10.1007/s13347-021-00477-0

Xu, S., Ge, Y., & Zhang, Y. (2023). Causal Explainable AI. In S. Li, & Z. Chu (Eds.), *Machine Learning for Causal Inference* (pp. 137-159). Springer International Publishing. https://doi.org/10.1007/978-3-031-35051-1_7