frontiers | Frontiers in Computational Neuroscience

TYPE Perspective
PUBLISHED 24 July 2024
DOI 10.3389/fncom.2024.1388166

OPEN ACCESS

EDITED BY
Ying Tan,
Peking University, China

REVIEWED BY
Yueheng Lan,
Beijing University of Posts and Telecommunications (BUPT), China
Ryo Karakida,
National Institute of Advanced Industrial Science and Technology (AIST), Japan

*CORRESPONDENCE
Haiping Huang
✉ huanghp7@mail.sysu.edu.cn

RECEIVED 19 February 2024
ACCEPTED 09 July 2024
PUBLISHED 24 July 2024

CITATION
Huang H (2024) Eight challenges in developing theory of intelligence.
Front. Comput. Neurosci. 18:1388166.
doi: 10.3389/fncom.2024.1388166

Eight challenges in developing theory of intelligence

Haiping Huang*

PMI Lab, School of Physics, Sun Yat-sen University, Guangzhou, China

A good theory of mathematical beauty is more practical than any current observation, as new predictions about physical reality can be self-consistently verified. This belief applies to the current status of understanding deep neural networks including large language models and even the biological intelligence. Toy models provide a metaphor of physical reality, allowing mathematically formulating the reality (i.e., the so-called theory), which can be updated as more conjectures are justified or refuted. One does not need to present all details in a model, but rather, more abstract models are constructed, as complex systems such as the brains or deep networks have many sloppy dimensions but much less stiff dimensions that strongly impact macroscopic observables. This type of bottom-up mechanistic modeling is still promising in the modern era of understanding the natural or artificial intelligence. Here, we shed light on eight challenges in developing theory of intelligence following this theoretical paradigm. Theses challenges are representation learning, generalization, adversarial robustness, continual learning, causal learning, internal model of the brain, next-token prediction, and the mechanics of subjective experience.

KEYWORDS

generalization, continual learning, adversarial robustness, brain dynamics, large language model (LLM), consciousness, statistical physics, artificial intelligence

1 Introduction

Brain is one of the most challenging subjects to understand. The brain is complex with many levels of temporal and spatial complexities (Gerstner et al., 2014), allowing for coarse-grained descriptions at different levels, especially in theoretical studies. More abstract models lose the ability to generate predictions on low-level details but bring the conceptual benefits of explaining precisely how the system works, and the mathematical description may be universal, independent of details (or sloppy variables) (Levenstein et al., 2023). One seminal example is the Hopfield model (Hopfield, 1982), where the mechanism underlying the associative memory observed in the brain was precisely isolated (Amit et al., 1987; Griniasty et al., 1993). There is a resurgence of research interests in Hopfield networks in recent years due to the large language models (Krotov and Hopfield, 2020; Ramsauer et al., 2020).

In Marr's viewpoint (Marr, 1982), understanding a neural system can be divided into three levels: computation (which task the brain solves), algorithms (how the brain solves the task, i.e., information processing level), and implementation (neural circuit level). Following the first two levels, researchers designed artificial neural networks to solve challenging real-world problems, such as powerful deep learning (LeCun et al., 2015; Schmidhuber, 2015). However, biological details are also being incorporated into models of neural networks (Abbott et al., 2016; Marblestone et al., 2016; Richards et al., 2019; Lillicrap et al., 2020) and even used to design new learning rules (Schmidgall et al., 2023).

Indeed, neuroscience studies of biological mechanisms of perception, cognition, memory, and action have already provided a variety of fruitful insights inspiring the empirical or scientific studies of artificial neural networks, which, in turn, inspire the neuroscience researchers to design mechanistic models to understand the brain (Yamins and DiCarlo, 2016; Hassabis et al., 2017; Saxe et al., 2020). Therefore, it is promising to integrate physics, statistics, computer science, psychology, neuroscience, and engineering to reveal inner working of deep (biological) networks and intelligence with testable predictions (Ma et al., 2022), rather than using a black box (e.g., deep artificial neural networks) to understand the other black boxes (e.g, the brain or mind). In fact, the artificial intelligence may follow different principles from the natural intelligence, but both can inspire each other, which may lead to the establishment of a coherent mathematical physics foundation for either artificial intelligence or biological intelligence.

The goal of providing a unified framework for neural computation is very challenging and even impossible. Due to re-boosted interests in neural networks, there appears a lot of important yet unsolved scientific questions. We shall detail these challenging questions below[1] and provide our personal viewpoints toward a statistical mechanics theory, solving these fundamental questions, based on the first principles in physics. These open scientific questions toward theory of intelligence are presented in Figure 1.

2 Challenge I—Representation learning

Given raw data (or input–output pairs in supervised learning), one can ask what a good representation is and how the meaningful representation is achieved in deep neural networks. We have not yet satisfied answers for these questions. A promising argument is that entangled manifolds at earlier layers of a deep hierarchy are gradually disentangled into linearly separable features at output layers (DiCarlo and Cox, 2007; Bengio et al., 2013; Brahma et al., 2016; Huang, 2018; Cohen et al., 2020). This manifold separation perspective is also promising in system neuroscience studies of associative learning by separating overlapping patterns of neural activities (Cayco-Gajic and Silver, 2019). However, an analytic theory of the manifold transformation is still lacking, prohibiting us from a full understanding of which key network parameters control the geometry of manifold and how learning reshapes the manifold. For example, the correlation among synapses (e.g., arising during learning) will attenuate the decorrelation process along the network depth but encourage dimension reduction compared with their orthogonal counterparts (Huang, 2018; Zhou and Huang, 2021). This result is derived by using mean-field approximation and coincides with empirical observations (Zhou and Huang, 2021). In addition, there may exist other biological plausible factors such as normalization, attention, and homeostatic control impacting the manifold transformation (Turrigiano and Nelson, 2004; Reynolds and Heeger, 2009), which can be incorporated into a toy model in future to test the manifold transformation hypothesis.

Another argument from information theoretic viewpoints demonstrates that the input information is maximally compressed into a hidden representation, whose task-related information should be maximally retrieved at the output layers, according to the information bottleneck theory (Achille and Soatto, 2017; Shwartz-Ziv and Tishby, 2017). In this sense, an optimal representation must be invariant to nuisance variability, and their components must be maximally independent, which may be related to causal factors (latent causes) explaining the sensory inputs (see the following fifth challenge). In a physics language, a coarse-grained (or more abstract) representation is formed in deeper layers compared with the fine-grained representation in shallower layers. How microscopic interactions among synapses determine this representation transformation remains elusive and thus deserves future studies; a few recent studies started to address the clustering structure in the deep hierarchy (Li and Huang, 2020, 2023; Alemanno et al., 2023; Xie et al., 2024). To conclude, the bottom-up mechanistic modeling would be fruitful in dissecting mechanisms of representation transformation.

3 Challenge II—Generalization

Studying any neural learning system must consider three ingredients: data, network, and algorithm (or DNA of neural learning). The generalization ability refers to the computational performance that the network is able to implement the rule in unseen examples. Therefore, intelligence can be considered to some extent as the ability of generalization, especially given very few examples for learning. Therefore, the generalization is also a hot topic in current studies of deep learning. Traditional statistical learning theory claims that over-fitting effects should be strong when the number of examples is much less than the number of parameters to learn, which could not explain the current success of deep learning. A promising perspective is to study the causal connection between the loss landscape and the generalization properties (Huang and Kabashima, 2014; Baldassi et al., 2016; Spigler et al., 2019; Zou and Huang, 2021). For a single layered perceptron, a statistical mechanics theory can be systematically derived and revealed a discontinuous transition from poor to perfect generalization (Gyorgyi, 1990; Sompolinsky et al., 1990). In contrast to the classical bias-variance trade-off (U-shaped curve of the test error vs. increasing model complexity) (Mehta et al., 2019), the modern deep learning achieves the state-of-the-art performance in the over-parameterized regime (Belkin et al., 2019; Spigler et al., 2019), a regime of the number of parameters much larger than the training data size. However, how to provide an analytic argument about the over-fitting effects vs. different parameterization regimes (e.g., under-, over-, and super-parameterization) for this empirical observation becomes a non-trivial task (Adlam and Pennington, 2020). A recent study of one-hidden-layer networks shows that the first transition

1 Most of them were roughly provided in the book of statistical mechanics of neural networks (Huang, 2022). Here we give a significantly expanded version.

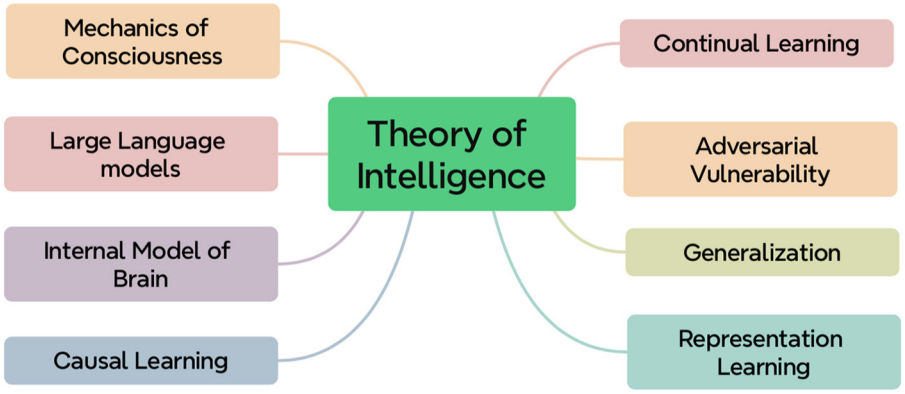

FIGURE 1
Schematic illustration of eight open challenging problems toward the theory of intelligence.

occurs at the interpolation point, where perfect fitting becomes possible. This transition reflects the properties of hard-to-sample typical solutions. Increasing the model complexity, the second transition occurs with the discontinuous appearance of atypical solutions. They are wide minima of good generalization properties. This second transition sets an upper bound for the effectiveness of learning algorithms (Baldassi et al., 2022). This statistical mechanics analysis focuses on the average case (average of all realizations of data, network, and algorithm) rather than the worst case. The worst case determines the computational complexity category, while the average case explains us the universal properties of learning, and the statistical mechanics links the computational hardness to a few order parameters in physics (Huang, 2022), and these previous studies show strong evidence (Huang and Kabashima, 2014; Baldassi et al., 2016, 2022; Spigler et al., 2019; Hou and Huang, 2020).

For an infinitely wide neural network, there exists a lazy learning regime, where the overparameterized neural networks can be well approximated by a linear model corresponding to a first-order Taylor expansion around the initialization, and the complex learning dynamics is simply training a kernel machine (Belkin, 2021). However, in a practical training, the dynamics is prone to escape the lazy regime, which has no satisfied theory yet. Therefore, clarifying which of lazy-learning (or neural tangent kernel limit) and feature-learning (or mean-field limit) may explain the success of deep supervised learning remains open and challenging (Jacot et al., 2018; Bartlett et al., 2021; Fang et al., 2021). The mean-field limit can be studied in the field theoretic framework, characterizing how the solution of learning deviates from the initialization through a systematic perturbation of the action in the framework (Segadlo et al., 2022). Another related challenge is out-of-distribution generalization, which can also be studied using statistical mechanics, e.g., in a recent study, a kernel regression was analyzed (Canatar et al., 2021). In addition, the field theoretic method is also promising to write the learning problem of out-of-distribution prediction into propagating correlations and responses (Segadlo et al., 2022).

4 Challenge III—Adversarial vulnerability

Adversarial examples are defined by those inputs with human-imperceptible modifications, leading to unexpected errors in a deep learning decision making system. The test accuracy drops as the perturbation grows; the perturbation can either rely on the trained network or be an independent noise (Szegedy et al., 2014; Goodfellow et al., 2015; Jiang et al., 2021). The current deep learning is argued to learn predictive yet non-robust features in the data (Geirhos et al., 2020). This adversarial vulnerability of deep neural networks poses a significant challenge in the practical applications of both real-world problems and AI4S (artificial intelligence for science) studies. Adversarial training remains the most effective solution to the problem (Madry et al., 2018), in contrast to human learning. However, the training sacrifices the standard discrimination. A recent study applied the physics principle that the hidden representation is clustered-like replica symmetry breaking in the spin glass theory (Mézard et al., 1987), which leads to contrastive learning that is local and adversarial robust, resolving the trade-off between standard accuracy and adversarial robustness (Xie et al., 2024). Furthermore, the adversarial robustness can be theoretically explained in terms of a cluster separation distance. In physics, systems with a huge number of degrees of freedom are able to be captured by a low-dimensional macroscopic description, such as Ising ferromagnetic model. Explaining the layered computation in terms of geometry may finally help to crack the mysterious property of the networks' susceptibility to adversarial examples (Bortolussi and Sanguinetti, 2018; Gilmer et al., 2018; Li and Huang, 2023). Although some recent efforts were devoted to this direction (Bortolussi and Sanguinetti, 2018; Kenway, 2018), more exciting results are expected in near future studies.

5 Challenge IV—Continual learning

A biological brain is good at adapting the acquired knowledge from similar tasks to domains of new tasks, even though only

a handful examples are available in the new task domain. This type of learning is called continual learning or multi-task learning (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017), an ability to learn many tasks in sequence, while transfer learning refers to the process of exploiting the previously acquired knowledge from a source task, to improve the generalization performance in a target task (Parisi et al., 2019). However, the stable adaptation to changing environments, an essence of lifelong learning, remains a significant challenge for modern artificial intelligence (Parisi et al., 2019). More precisely, neural networks are, in general, poor at the multi-task learning, although impressive progresses have been achieved in recent years. For example, during learning, a diagonal Fisher information term is computed to measure importances of weights (then a rapid change is not allowed for those important weights) for previous tasks (Kirkpatrick et al., 2017). A later refinement was also proposed by allowing synapses, accumulating task-relevant information over time (Zenke et al., 2017). To reduce the catastrophic-forgetting effects, more machine learning techniques were summarized in the review (Parisi et al., 2019). However, we still do not know the exact mechanisms for principally mitigating the catastrophic-forgetting effects, which calls for theoretical studies of deep learning in terms of adaptation to domain-shift training, i.e., connection weights trained in a solution to one task are transformed to benefit learning on a related task.

Using asymptotic analysis, a recent article studying transfer learning identified a phase transition in the quality of the knowledge transfer (Dhifallah and Lu, 2021). This study reveals how the related knowledge contained in a source task can be effectively transferred to boost the performance in a target task. Other recent theoretical studies interpreted the continual learning with a statistical mechanics framework using Franz-Parisi potential (Li et al., 2023) or as an on-line mean-field dynamics of weight updates (Lee et al., 2021). The Franz-Parisi potential is a thermodynamic potential used to study glass transition (Franz and Parisi, 1995). The recent study assumes that the knowledge from the previous task behaves as a reference configuration (Li et al., 2023), where the previously acquired knowledge serves as an anchor for learning new knowledge. This framework also connects to elastic weight consolidation (Kirkpatrick et al., 2017), heuristic weight-uncertainty modulation (Ebrahimi et al., 2020), and neuroscience-inspired metaplasticity (Laborieux et al., 2021), providing a theory-grounded method for the real-world multi-task learning with deep networks.

6 Challenge V—Causal learning

Deep learning is criticized as being nothing but a fancy curve-fitting tool, making a naive association between inputs and outputs. In other words, this tool could not distinguish correlation from causation. What the deep network learns is not a concept but merely a statistical correlation, prohibiting the network from counterfactual inference (a hallmark ability of intelligence). A human-like AI must be good at retrieving causal relationship among feature components in sensory inputs, thereby carving relevant information from a sea of irrelevant noise (Pearl and Mackenzie, 2018; Schölkopf, 2019; Schölkopf et al., 2021). Therefore, understanding cause and effect in deep learning systems is particularly important for the next-generation artificial intelligence. The question whether the current deep learning algorithm is able to do causal reasoning remains open. Hence, how to design a learning system that can infer the effect of an intervention becomes a key to address this question, although it would be very challenging to make deep learning extract causal structure from observations by applying simple physics principles due to both architecture and learning complexities. This challenge is now intimately related to the astonishing performances of large language models (see the following seventh challenge), and the key question is whether the self-attention mechanism is sufficient for capturing the causal relationships in the training data.

7 Challenge VI—Internal model of the brain

The brain is argued to learn to build an internal model of the outside world, which is reflected by spontaneous neural activities as a reservoir for computing (e.g., sampling) (Ringach, 2009). The agreement between spontaneous activity and stimulus-evoked one increases during development, especially for natural stimuli (Berkes et al., 2011), while the spontaneous activity outlines the regime of evoked neural responses (Luczak et al., 2009). The relationship between the spontaneous fluctuation and task-evoked response causes recent interests in studying brain dynamics (Deco et al., 2023). This can be formulated by the fluctuation–dissipation theorem in physics, and the violation can be a measure of deviation from equilibrium, although a non-equilibrium stationary state exists.

In addition, the stimuli were shown to carve a clustered neural space (Huang and Toyoizumi, 2016; Berry and Tkačik G, 2020). Then, an interesting question is what the spontaneous neural space looks like, and how the space dynamically evolves, especially in the adaptation to changing environments. Furthermore, how sensory inputs combined with the ongoing asynchronous cortical activity determine the animal behavior remains open and challenging. If the reward-mediated learning is considered, reinforcement learning was used to build world models of structured environments (Ha and Schmidhuber, 2018). In the reinforcement learning, observations are used to drive actions, which are evaluated based on reward signals the agent receives from the environment after taking the actions. It is thus interesting to reveal which type of internal models the agent establishes through learning from interactions with the environments. This can be connected to aforementioned representation and generalization challenges. Moreover, a recent study showed a connection between the reinforcement learning and statistical physics (Rahme and Adams, 2019), suggesting that a statistical mechanics theory could be potentially established to understand how an optimal policy is found to maximize the long-term accumulated reward, with an additionally potential impact on studying reward-based neural computations in the brain (Neftci and Averbeck, 2019).

Another angle to look at the internal model of the brain is through the lens of neural dynamics (Buonomano and Maass, 2009; Deco et al., 2009; Sussillo and Abbott, 2009; Vyas et al., 2020), which is placed onto a low-dimensional surface, robust to

variations in detailed properties of individual neurons or circuits. The representation of stimuli, tasks, or contexts can be retrieved for deriving experimentally testable hypotheses (Jazayeri and Ostojic, 2021). Although previous theoretical studies were carried out in recurrent rate or spiking activity neural networks (Sompolinsky et al., 1988; Brunel, 2000), a challenging issue remains to address how neural activity and synaptic plasticity interact with each other to yield a low-dimensional internal representation for cognitive functions. The recent development of synaptic plasticity combining connection probability, local synaptic noise, and neural activity can realize a dynamic network in the adaptation to time-dependent inputs (Zou et al., 2023). This study interprets learning as a variational inference problem, making optimal learning under uncertainty possible in a local circuit. Both learning and neural activity are placed on low-dimensional subspaces. Future studies must include more biological plausible factors to test the hypothesis in neurophysiological experiments. Another recent exciting achievement is using dynamical mean-field theory to uncover rich dynamical regimes of coupled neuronal-synaptic dynamics (Clark and Abbott, 2024).

Brain states can be considered as an ensemble of dynamical attractors (von der Malsburg, 2018). The key challenge is how learning shapes the stable attractor landscape. One can interpret the learning as a Bayesian inference in an unsupervised way but not the autoregressive manner (see the next section). The learning can then be driven by synaptic weight symmetry breaking (Hou et al., 2019; Hou and Huang, 2020), separating two phases of recognizing the network itself and the rule hidden in sensory inputs. It is very interesting to observe if this picture still holds in recurrent learning supporting neural trajectories on dynamical attractors, and even predictive learning minimizes a free energy of belief and synaptic weights (the belief leads to error neurons) (Jiang and Rao, 2024). New methods must be developed based on the recently proposed quasi-potential method to study non-equilibrium steady neural dynamics (Qiu and Huang, 2024) or dynamical mean-field theory for learning (Zou and Huang, 2024).

8 Challenge VII—Large language models

The impressive problem-solving capabilities of Chat-GPT, where GPT is a shorthand of generative pretrained transformer, are driving the fourth industrial revolution. The Chat-GPT is based on large language models (LLMs) (OpenAI, 2023), which represent linguistic information as vectors in high-dimensional state space, as trained with a large text corpus in an autoregressive way [in analogy to the hypothesis that the brain is a prediction machine (Clark, 2013)], resulting in a complex statistical model of how the tokens in the training data correlate (Vaswani et al., 2017). The computational model thus shows strong formal linguistic competence (Mahowald et al., 2024). The LLM is also a few-shot or zero-shot learner (Brown et al., 2020; Kojima et al., 2022), i.e., the language model can perform a wide range of computationally challenging tasks with prompting alone [e..g, chain-of-thought prompting (Wei et al., 2022)]. Remarkably, the LLMs display a qualitative leap in capability as the model complexity and sample complexity are both scaled up (Kaplan et al., 2020), akin to phase transitions in thermodynamic systems.

In contrast to the formal linguistic competence, the functional linguistic competence is argued to be weak (Mahowald et al., 2024). This raises a fundamental question what the nature of intelligence is or whether a single next-token context conditional prediction is a standard model of artificial general intelligence (Gerven, 2017; Lake et al., 2017; Sejnowski, 2023). Human's reasoning capabilities in real-world problems rely on non-linguistic information, e.g., it is unpredictable when a creative idea for a scientist would come to a challenging problem at hand, which relies on reasoning about the implications along a sequence of thought. In a biological implementation, the language modules are separated from the other modules involving high-level cognition (Mahowald et al., 2024). The LLM explains hierarchical correlations in word pieces in the training corpora rather than hidden casual dependencies. In other words, the neural network has not constructed a mental model of the world, which requires heterogeneous modular networks, thereby unlike humans. Therefore, the LLM does not know what it generates (as a generative model). Even if some key patterns of statistical regularities are absent in the training data, the model can generate perfect texts in terms of syntax. However, the texts may be different from the truth. Knowing what they know is a crucial hallmark of intelligent systems (Gerven, 2017). In this sense, the inner workings of the LLM are largely opaque, requiring a great effort to mathematically formulate the formal linguistic competence and further identify key elements that must be included to develop a robust model of the world. Mechanisms behind the currently observed false-positive such as hallucination (Chomsky et al., 2023) could then be revealed, which may be related to interpolation between modes of token distributions. A recent study interpreting the attention used in transformer-based LLM as a generalized Potts model in physics seems inspiring (Rende et al., 2024), i.e., tokens as Potts spin vectors.

Most importantly, we currently do not have any knowledge on how to build an additional network that is able to connect performance with awareness (Cleeremans, 2014), which is linked to what makes us conscious (see the last challenge). Following Marr's framework, both computational and neural correlates of consciousness remain unknown (Crick and Koch, 2003; Blum and Blum, 2022; Dwarakanath et al., 2023). A current physical way is to consider a Lyapunov function governing complex neural computation underlying LLMs (Krotov and Hopfield, 2020; Ramsauer et al., 2020). In this way, the Lyapunov function perspective will open the door of many degrees of freedom to control how information is distilled via not only the self-attention but also other potential gating mechanisms, based on the dynamical system theory.

9 Challenge VIII—Theory of consciousness

One of the most controversial questions is the origin of consciousness—whether the consciousness is an emergent behavior of highly heterogeneous and modular brain circuits with various carefully designed regions [e.g., $\sim 10^{14}$ connections for the human

brain and many functionally specific modular structures, such as the prefrontal cortex, hippocampus, and cerebellum (Harris and Shepherd, 2015; Luo, 2021)]. The subjectivity of the conscious experience is in contradiction with the objectivity of a scientific explanation. According to Damasio's model (Damasio, 2001), the ability to identify one-self in the world and its relationship with the world is considered to be a central characteristic of conscious state. Whether a machine algorithm can achieve the self-awareness remains elusive. The self-monitoring ability [or meta-cognition (Dehaene et al., 2017)] may endow the machine (such as LLMs) to know what they generate. It may be important to clarify how the model of one-self is related to the internal model of the brain [e.g., through recurrent or predictive processing (Storm et al., 2024)]. For example, Karl Friston argued that the conscious processing can be interpreted as a statistical inference problem of inferring causes of sensory observations. Therefore, minimizing the surprise (negative log probability of an event) may lead to self-consciousness (Friston, 2018), which is consistent with the hypothesis that the brain is a prediction machine (Clark, 2013; Gerven, 2017).

There are currently two major cognitive theories of consciousness. One is the global workspace framework (Dehaene et al., 1998), which relates consciousness to the widespread and sustained propagation of cortical neural activities by demonstrating that consciousness arises from an ignition that leads to global information broadcast among brain regions. This computational functionalism was recently leveraged to discuss possibility of consciousness in non-organic artificial systems (Bengio, 2017; Butlin et al., 2023). The other is the integrated information theory that provides a quantitative characterization of conscious state by integrated information (Tononi, 2004). In this second theory, unconscious states have a low information content, while conscious states have a high information content. The second theory emphasizes the phenomenal properties of consciousness (Albantakis et al., 2023), i.e., the function performed by the brain is not subjective experience. Both theories follow a top-down approach, which is in stark contrast to the statistical mechanics approach following a bottom-up manner building the bridge from microscopic interactions to macroscopic behavior. These hypotheses are still under intensive criticism despite some cognitive experiments they can explain (Koch et al., 2016). We remark that conscious states may be an emergent property of neural activities, lying at a higher level than neural activities. It is currently unknown how to connect these two levels, for which a new statistical mechanics theory is required. An exciting route is to link the spontaneous fluctuation to stimulus-evoked response, and a maximal response is revealed in a recurrent computational model (Qiu and Huang, 2024), which can be thought of as a necessary condition for consciousness, as information-richness of cortical electrodynamics was also observed to be peaked at the edge-of-chaos (dynamics marginal stability) (Toker et al., 2022). This peak thus distinguishes the conscious from unconscious brain states. From an information-theoretic argument, the conscious state may require a diverse range of configurations of interactions between brain networks, which can be linked to the entropy concept in physics (Guevara Erra et al., 2016). The large entropy leads to optimal segregation and integration of information (Zhou et al., 2015).

Taken together, whether the consciousness can be created from an interaction of local dynamics within complex neural substrate is still unsolved (Krauss and Maier, 2020). A statistical mechanics theory, if possible, is always promising in the sense that one can make theoretical predictions from just a few physics parameters (Huang, 2022), which may be possible from a high degree of abstraction, and thus a universal principle could be expected.

10 Conclusion

To sum up, in this viewpoint, we provide some naive thoughts about fundamental important questions related to neural networks, for which building a good theory is different from being completed. The traditional research studies of statistical physics of neural networks bifurcate to two main streams: one is to the engineering side, developing theory-grounded algorithms; and the other is to the neuroscience side, formulating brain computation by mathematical models solved by physics methods. In physics, we have the principle of least action, from which we can deduce the classical mechanics or electrodynamics laws. We are not sure whether in physics of neural networks (and even the brain), there exists general principles that can be expressed in a concise form of mathematics. It is exciting yet challenging to promote the interplay between physics theory and neural computations along these eight open problems discussed in the perspective paper. The advances will undoubtedly lead to a human-interpretable understanding of underlying mechanisms of the artificial intelligent systems, the brain and mind, especially in the era of big experimental data in brain science and rapid progress in AI studies.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

HH: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the National Natural Science Foundation of China for Grant numbers 12122515.

Acknowledgments

The author would like to thank all PMI members for discussions lasting for 5 years. This perspective also benefits from discussions with students during the on-line course of statistical mechanics of neural networks (from September 2022 to June 2023).

The author are also grateful to invited speakers in the INTheory on-line seminar during the COVID-19 pandemic. The author enjoyed a lot of interesting discussions with Adriano Barra, Yan Fyodorov, Sebastian Goldt, Pulin Gong, Moritz Helias, Kamesh Krishnamurthy, Yi Ma, Alexander van Meegen, Remi Monasson, Srdjan Ostojic, and Riccardo Zecchina.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abbott, L. F., DePasquale, B., and Memmesheimer, R.-M. (2016). Building functional networks of spiking model neurons. *Nat. Neurosci.* 19, 350–355. doi: 10.1038/nn.4241

Achille, A., and Soatto, S. (2017) A separation principle for control in the age of deep learning. *arXiv* [Preprint]. arXiv:1711.03321. doi: 10.48550/arXiv.1711.03321

Adlam, B., and Pennington, J. (2020). "The neural tangent kernel in high dimensions: triple descent and a multi-scale theory of generalization," in *ICML 2020: 37th International Conference on Machine Learning* (PMLR), 119, 74–84. Available online at: http://proceedings.mlr.press/v119/adlam20a/adlam20a.pdf

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., et al. (2023). Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. *PLoS Comput. Biol.* 19, 1–45. doi: 10.1371/journal.pcbi.1011465

Alemanno, F., Aquaro, M., Kanter, I., Barra, A., and Agliari, E. (2023) Supervised hebbian learning. *Europhys. Lett.* 141:11001. doi: 10.1209/0295-5075/aca55f

Amit, D. J., Gutfreund, H., and Sompolinsky, H. (1987). Statistical mechanics of neural networks near saturation. *Ann. Phys.* 173, 30–67. doi: 10.1016/0003-4916(87)90092-3

Baldassi, C., Borgs, C., Chayes, J. T., Ingrosso, A., Lucibello, C., Saglietti, L., et al. (2016). Unreasonable effectiveness of learning neural networks: from accessible states and robust ensembles to basic algorithmic schemes. *Proc. Natl Acad. Sci. U.S.A.* 113, E7655–E7662. doi: 10.1073/pnas.1608103113

Baldassi, C., Lauditi, C., Malatesta, E. M., Pacelli, R., Perugini, G., Zecchina, R., et al. (2022). Learning through atypical phase transitions in overparameterized neural networks. *Phys. Rev. E*, 106:014116. doi: 10.1103/PhysRevE.106.014116

Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *arXiv* [Preprint]. arXiv:2103.09177. doi: 10.48550/arXiv.2103.09177

Belkin, M. (2021). Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *arXiv* [Preprint]. arXiv:2105.14368. doi: 10.48550/arXiv.2105.14368

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA.* 116, 15849–15854. doi: 10.1073/pnas.1903070116

Bengio, Y. (2017). The consciousness prior. *arXiv* [Preprint]. arXiv:1709.08568. doi: 10.48550/arXiv.1709.08568

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intellig.* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Berkes, P., Orban, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* 331:83. doi: 10.1126/science.1195870

Berry, M. J. Tkačik G. (2020). Clustering of neural activity: a design principle for population codes. *Front. Comput. Neurosci.* 14:20. doi: 10.3389/fncom.2020.00020

Blum, L., and Blum, M. (2022). A theory of consciousness from a theoretical computer science perspective: insights from the conscious turing machine. *Proc. Natl. Acad. Sci. USA.* 119:e2115934119. doi: 10.1073/pnas.2115934119

Bortolussi, L., and Sanguinetti, G. (2018). Intrinsic geometric vulnerability of high-dimensional artificial intelligence. *arXiv* [Preprint]. arXiv:1811.03571. doi: 10.48550/arXiv.1811.03571

Brahma, P. P., Wu, D., and She, Y. (2016). Why deep learning works: a manifold disentanglement perspective. *IEEE Trans. Neural Netw. Learn. Syst.* 27, 1997–2008. doi: 10.1109/TNNLS.2015.2496947

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (Eds.) (2020). *Advances in Neural Information Processing Systems, Volume 33*. Red Hook, NY: Curran Associates, Inc, 1877-1901.

Brunel, N. (2000). Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J. Comput. Neurosci.* 8, 183–208. doi: 10.1023/A:1008925309027

Buonomano, D. V., and Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.* 10, 113–125. doi: 10.1038/nrn2558

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv* [Preprint]. arXiv:2308.08708. doi: 10.48550/arXiv.2308.08708

Canatar, A., Bordelon, B., and Pehlevan, C. (2021). "Out-of-distribution generalization in kernel regression," in *Advances in Neural Information Processing Systems, Vol. 34*, eds. M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Red Hook, NY: Curran Associates, Inc.), 12600–12612.

Cayco-Gajic, N. A., and Silver, R. A. (2019). Re-evaluating circuit mechanisms underlying pattern separation. *Neuron* 101, 584–602. doi: 10.1016/j.neuron.2019.01.044

Chomsky, N., Roberts, I., and Watumull, J. (2023). Noam chomsky: the false promise of chatgpt. *The New York Times*, 8.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Clark, D. G., and Abbott, L. F. (2024). Theory of coupled neuronal-synaptic dynamics. Phys. Rev. *X* 14:021001. doi: 10.1103/PhysRevX.14.021001

Cleeremans, A. (2014). Connecting conscious and unconscious processing. *Cogn. Sci.* 38, 1286–1315. doi: 10.1111/cogs.12149

Cohen, U., Chung, S., Lee, D. D., and Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* 11:746. doi: 10.1038/s41467-020-14578-5

Crick, F., and Koch, C. (2003). A framework for consciousness. *Nat. Neurosci.* 6, 119–126. doi: 10.1038/nn0203-119

Damasio, A. (2001). Fundamental feelings. *Nature* 413:781. doi: 10.1038/35101669

Deco, G., Lynn, C. W., Perl, Y. S., and Kringelbach, M. L. (2023). Violations of the fluctuation-dissipation theorem reveal distinct nonequilibrium dynamics of brain states. *Phys. Rev. E* 108:064410. doi: 10.1103/PhysRevE.108.064410

Deco, G., Rolls, E. T., and Romo, R. (2009). Stochastic dynamics as a principle of brain function. *Prog. Neurobiol.* 88, 1–16. doi: 10.1016/j.pneurobio.2009.01.006

Dehaene, S., Kerszberg, M., and Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. USA.* 95, 14529–14534. doi: 10.1073/pnas.95.24.14529

Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? *Science* 358, 486–492. doi: 10.1126/science.aan8871

Dhifallah, O., and Lu, Y. M. (2021). Phase transitions in transfer learning for high-dimensional perceptrons. *Entropy* 23:400. doi: 10.3390/e23040400

DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341. doi: 10.1016/j.tics.2007.06.010

Dwarakanath, A., Kapoor, V., Werner, J., Safavi, S., Fedorov, L. A., Logothetis, N. K., et al. (2023). Bistability of prefrontal states gates access to consciousness. *Neuron* 111, 1666–1683. doi: 10.1016/j.neuron.2023.02.027

Ebrahimi, S., Elhoseiny, M., Darrell, T., and Rohrbach, M. (2020). "Uncertainty-guided continual learning with Bayesian neural networks," in *International Conference on Learning Representations*.

Fang, C., Dong, H., and Zhang, T. (2021). Mathematical models of overparameterized neural networks. *Proc. IEEE* 109, 683–703. doi: 10.1109/JPROC.2020.3048020

Franz, S., and Parisi, G. (1995). Recipes for metastable states in spin glasses. *J. Phys. I* 5, 1401–1415. doi: 10.1051/jp1:1995201

Friston, K. (2018). Am I self-conscious? (or does self-organization entail self-consciousness?). *Front. Psychol.* 9:579. doi: 10.3389/fpsyg.2018.00579

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. (2020). Shortcut learning in deep neural networks. *Nat. Mach. Intell.* 2, 665–673. doi: 10.1038/s42256-020-00257-z

Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge: Cambridge University Press.

Gerven, M. (2017). Computational foundations of natural intelligence. *Front. Comput. Neurosci.* 11:112. doi: 10.3389/fncom.2017.00112

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., et al. (2018). Adversarial spheres. *arXiv* [Preprint]. arXiv:1801.02774. doi: 10.48550/arXiv.1801.02774

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). "Explaining and harnessing adversarial examples," in *ICLR 2015: International Conference on Learning Representations 2015*.

Griniasty, M.,Tsodyks, M. V., and Amit, D. J. (1993). Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Comput.* 5, 1–17. doi: 10.1162/neco.1993.5.1.1

Guevara Erra, R., Mateos, D. M., Wennberg, R., and Perez Velazquez, J. L. (2016). Statistical mechanics of consciousness: maximization of information content of network is associated with conscious awareness. *Phys. Rev. E* 94L52402. doi: 10.1103/PhysRevE.94.052402

Gyorgyi, G. (1990). First-order transition to perfect generalization in a neural network with binary synapses. *Phys. Rev. A* 41, 7097–7100.

Ha, D., and Schmidhuber, J. (2018). World models. *arXiv* [Preprint]. arXiv:1803.10122. doi: 10.48550/arXiv.1803.10122

Harris, K., and Shepherd, G. (2015). The neocortical circuit: themes and variations. *Nat. Neurosci.* 18, 170–181. doi: 10.1038/nn.3917

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558. doi: 10.1073/pnas.79.8.2554

Hou, T., and Huang, H. (2020). Statistical physics of unsupervised learning with prior knowledge in neural networks. *Phys. Rev. Lett.* 124:248302. doi: 10.1103/PhysRevLett.124.248302

Hou, T., Wong, K. Y. M., and Huang, H. (2019). Minimal model of permutation symmetry in unsupervised learning. *J. Phys. A: Math. Theor.* 52:414001. doi: 10.1088/1751-8121/ab3f3f

Huang, H. (2018). Mechanisms of dimensionality reduction and decorrelation in deep neural networks. *Phys. Rev. E* 98:062313. doi: 10.1103/PhysRevE.98.062313

Huang, H. (2022). *Statistical Mechanics of Neural Networks*. Singapore: Springer. doi: 10.1007/978-981-16-7570-6

Huang, H., and Kabashima, Y. (2014). Origin of the computational hardness for learning with binary synapses. *Phys. Rev. E* 90:052813. doi: 10.1103/PhysRevE.90.052813

Huang, H., and Toyoizumi, T. (2016). Clustering of neural code words revealed by a first-order phase transition. *Phys. Rev. E* 93:062416. doi: 10.1103/PhysRevE.93.062416

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: convergence and generalization in neural networks. *Adv. Neural Inf. Process. Syst.* 31, 8571–8580. Available online at: https://papers.nips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html

Jazayeri, M., and Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Curr. Opin. Neurobiol.* 70, 113–120. doi: 10.1016/j.conb.2021.08.002

Jiang, L. P., and Rao, R. P. N. (2024). Dynamic predictive coding: a model of hierarchical sequence learning and prediction in the neocortex. *PLoS Comput. Biol.* 20, 1–30. doi: 10.1371/journal.pcbi.1011801

Jiang, Z., Zhou, J., and Huang, H. (2021). Relationship between manifold smoothness and adversarial vulnerability in deep learning with local errors. *Chin. Phys. B* 30:048702. doi: 10.1088/1674-1056/abd68e

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling laws for neural language models. *arXiv* [Preprint]. arXiv:2001.08361. doi: 10.48550/arXiv.2001.08361

Kenway, R. (2018). Vulnerability of deep learning. *arXiv* [Preprint]. arXiv:1803.06111. doi: 10.48550/arXiv.1803.06111

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA*, 114:3521-3526. doi: 10.1073/pnas.1611835114

Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci.* 17, 307–321. doi: 10.1038/nrn.2016.22

Kojima, T., Shane Gu, S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv* [Preprint]. arXiv:2205.11916. doi: 10.48550/arXiv.2205.11916

Krauss, P., and Maier, A. (2020). Will we ever have conscious machines? *Front. Comput. Neurosci.* 14:556544. doi: 10.3389/fncom.2020.556544

Krotov, D., and Hopfield, J. (2020). Large associative memory problem in neurobiology and machine learning. *arXiv* [Preprint]. arXiv:2008.06996. doi: 10.48550/arXiv.2008.06996

Laborieux, A., Ernoult, M., Hirtzlin, T., and Querlioz, D. (2021). Synaptic metaplasticity in binarized neural networks. *Nat. Commun.* 12:2549. doi: 10.1038/s41467-021-22768-y

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci.* 40:e253. doi: 10.1017/S0140525X16001837

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 52, 436–444. doi: 10.1038/nature14539

Lee, S., Goldt, S., and Saxe, A. (2021). "Continual learning in the teacher-student setup: impact of task similarity," in *Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research*, M. Meila, and T. Zhang (PMLR), 6109–6119.

Levenstein, D., Alvarez, V. A., Amarasingham, A., Azab, H., Chen, Z. S., Gerkin, R. C., et al. (2023). On the role of theory and modeling in neuroscience. *J. Neurosci.* 43, 1074–1088. doi: 10.1523/JNEUROSCI.1179-22.2022

Li, C., and Huang, H. (2020). Learning credit assignment. *Phys. Rev. Lett.* 125:178301. doi: 10.1103/PhysRevLett.125.178301

Li, C., and Huang, H. (2023). Emergence of hierarchical modes from deep learning. *Phys. Rev. Res.* 5:L022011. doi: 10.1103/PhysRevResearch.5.L022011

Li, C., Huang, Z., Zou, W., and Huang, H. (2023). Statistical mechanics of continual learning: variational principle and mean-field potential. *Phys. Rev. E* 108:014309. doi: 10.1103/PhysRevE.108.014309

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nat. Rev. Neurosci.* 21, 335–346. doi: 10.1038/s41583-020-0277-3

Luczak, A., Bartho, P., and Harris, K. D. (2009). Spontaneous events outline the realm of possible sensory responses in neocortical populations. *Neuron* 62:413. doi: 10.1016/j.neuron.2009.03.014

Luo, L. (2021). Architectures of neuronal circuits. *Science* 373:eabg7285. doi: 10.1126/science.abg7285

Ma, Y., Tsao, D., and Shum, H.-Y. (2022). On the principles of parsimony and self-consistency for the emergence of intelligence. *Front. Inform. Technol. Electron. Eng.* 23, 1298–1323. doi: 10.1631/FITEE.2200297

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., Fedorenko, E., et al. (2024). Dissociating language and thought in large language models. *Trends Cogn. Sci.* 28, 517–540. doi: 10.1016/j.tics.2024.01.011

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.*10:94. doi: 10.3389/fncom.2016.00094

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

McCloskey, M., and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. *Psychol. Learn. Motiv.* 24, 109–165. doi: 10.1016/S0079-7421(08)60536-8

Mehta, P., Bukov, M., Wang, C.-H., Day, A. G. R., Richardson, C., Fisher, C. K., et al. (2019). A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* 810, 1–124. doi: 10.1016/j.physrep.2019.03.001

Mézard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin Glass Theory and Beyond*. Singapore: *World Scientific*. doi: 10.1142/0271

Neftci, E. O., and Averbeck, B. B. (2019). Reinforcement learning in artificial and biological systems. *Nat. Mach. Intell.* 1, 133–143. doi: 10.1038/s42256-019-0025-4

OpenAI (2023). Gpt-4 technical report. *arXiv* [Preprint]. arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: a review. *Neural Netw.* 113, 54–71. doi: 10.1016/j.neunet.2019.01.012

Pearl, J., and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books.

Qiu, J., and Huang, H. (2024). An optimization-based equilibrium measure describes non-equilibrium steady state dynamics: application to edge of chaos. *arXiv* [Preprint]. arXiv:2401.10009. doi: 10.48550/arXiv.2401.10009

Rahme, J., and Adams, R. P. (2019). A theoretical connection between statistical physics and reinforcement learning. *arXiv* [Preprint]. arXiv:1906.10228. doi: 10.48550/arXiv.1906.10228

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., et al. (2020). Hopfield networks is all you need. *arXiv* [Preprint]. arXiv:2008.02217. doi: 10.48550/arXiv.2008.02217

Rende, R., Gerace, F., Laio, A., and Goldt, S. (2024). Mapping of attention mechanisms to a generalized potts model. *Phys. Rev. Res.* 6:023057. doi: 10.1103/PhysRevResearch.6.023057

Reynolds, J. H., and Heeger, D. J. (2009). The normalization model of attention. *Neuron* 61, 168–185. doi: 10.1016/j.neuron.2009.01.002

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770. doi: 10.1038/s41593-019-0520-2

Ringach, D. L. (2009). Spontaneous and driven cortical activity: implications for computation. *Curr. Opin. Neurobiol.* 19, 439–444. doi: 10.1016/j.conb.2009.07.005

Saxe, A., Nelli, S., and Summerfield, C. (2020). If deep learning is the answer, then what is the question? *Nat. Rev. Neurosci.* 22, 55–67. doi: 10.1038/s41583-020-00395-8

Schmidgall, S., Achterberg, J., Miconi, T., Kirsch, L., Ziaei, R., Hajiseyedrazi, S. P., et al. (2023). Brain-inspired learning in artificial neural networks: a review. *arXiv* [Preprint]. arXiv:2305.11252. doi: 10.48550/arXiv.2305.11252

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y (2021). Toward causal representation learning. *Proc. IEEE* 109, 612–634. doi: 10.1109/JPROC.2021.3058954

Schölkopf, B (2019). Causality for machine learning. *arXiv* [Preprint]. arXiv:1911.10500. doi: 10.48550/arXiv.1911.10500

Segadlo, K., Epping, B., van Meegen, A., and Dahmen, D. Krämer M, Helias M. (2022). Unified field theoretical approach to deep and recurrent neuronal networks. *J. Stat. Mech. Theor. Exp.* 2022:103401. doi: 10.1088/1742-5468/ac8e57

Sejnowski, T. J. (2023). Large language models and the reverse turing test. *Neural Comput.* 35, 309–342. doi: 10.1162/neco_a_01563

Shwartz-Ziv, R., and Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv* [Preprint]. arXiv:1703.00810. doi: 10.48550/arXiv.1703.00810

Sompolinsky, H., Crisanti, A., and Sommers, H. J. (1988). Chaos in random neural networks. *Phys. Rev. Lett.* 61, 259–262. doi: 10.1103/PhysRevLett.61.259

Sompolinsky, H., Tishby, N., and Seung, H. S. (1990). Learning from examples in large neural networks. *Phys. Rev. Lett.* 65, 1683–1686. doi: 10.1103/PhysRevLett.65.1683

Spigler, S., Geiger, M., D'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. (2019). A jamming transition from under- to over-parametrization affects generalization in deep learning. *J. Phys. A Math. Theor.* 52:474001. doi: 10.1088/1751-8121/ab4c8b

Storm, J. F., Christiaan Klink, P., Aru, J., Senn, W., Goebel, R., Pigorini, A., et al. (2024). An integrative, multiscale view on neural theories of consciousness. *Neuron* 112, 1531–1552. doi: 10.1016/j.neuron.2024.02.004

Sussillo, D., and Abbott, L. F. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron* 63, 544–557. doi: 10.1016/j.neuron.2009.07.018

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2014). "Intriguing properties of neural networks," in *ICLR 2014 : International Conference on Learning Representations (ICLR) 2014.*

Toker, D., Pappas, I., Lendner, J. D., Frohlich, J., Mateos, D. M., Muthukumaraswamy, S., et al. (2022). Consciousness is supported by near-critical slow cortical electrodynamics. *Proc. Natl Acad. Sci.* 119:e2024455119. doi: 10.1073/pnas.2024455119

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci.* 5:42. doi: 10.1186/1471-2202-5-42

Turrigiano, G. G., and Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nat. Rev. Neurosci.* 5, 97–107. doi: 10.1038/nrn1327

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (Red Hook, NY: Curran Associates Inc), 6000–6010.

von der Malsburg C (2018). Concerning the neural code. *arXiv* [Preprint]. arXiv:1811.01199. doi: 10.48550/arXiv.1811.01199

Vyas, S., Golub, M. D., Sussillo, D., and Shenoy, K. V. (2020). Computation through neural population dynamics. *Ann. Rev. Neurosci.* 43, 249–275. doi: 10.1146/annurev-neuro-092619-094115

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Xia, F., et al. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv* [Preprint]. arXiv:2201.11903. doi: 10.48550/arXiv.2201.11903

Xie, M., Wang, Y., and Huang, H. (2024). Fermi-bose machine. *arXiv* [Preprint]. arXiv:2404.13631. doi: 10.48550/arXiv.2404.13631

Yamins, D. L. K., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 1, 356–365. doi: 10.1038/nn.4244

Zenke, F., Poole, B., and Ganguli, S. (2017). "Continual learning through synaptic intelligence," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 3987–3995.

Zhou, D. W., Mowrey, D. D., Tang, P., and Xu, Y. (2015). Percolation model of sensory transmission and loss of consciousness under general anesthesia. *Phys. Rev. Lett.* 115:108103. doi: 10.1103/PhysRevLett.115.108103

Zhou, J., and Huang, H (2021). Weakly-correlated synapses promote dimension reduction in deep neural networks. *Phys. Rev. E* 103:012315. doi: 10.1103/PhysRevE.103.012315

Zou, W., and Huang, H. (2021). Data-driven effective model shows a liquid-like deep learning. *Phys. Rev. Res.* 3:033290. doi: 10.1103/PhysRevResearch.3.033290

Zou, W., and Huang, H. (2024). "Introduction to dynamical mean-field theory of randomly connected neural networks with bidirectionally correlated couplings," in *SciPost Phys. Lect. Notes*, 79 (Amsterdam: SciPost Foundation). doi: 10.21468/SciPostPhysLectNotes.79

Zou, W., Li, C., and Huang, H. (2023). Ensemble perspective for understanding temporal credit assignment. *Phys. Rev. E* 107:024307. doi: 10.1103/PhysRevE.107.024307