Cognition 203 (2020) 104365

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Deep learning and cognitive science

Pietro Perconti, Alessio Plebe*

University of Messina, Department of Cognitive Science, v. Concezione 8, 98121 Messina, Italy

ARTICLE INFO

Keywords:
Deep learning
Embodied cognition
Enactive vision
Artificial neural networks
Mental representations

ABSTRACT

In recent years, the family of algorithms collected under the term "deep learning" has revolutionized artificial intelligence, enabling machines to reach human-like performances in many complex cognitive tasks. Although deep learning models are grounded in the connectionist paradigm, their recent advances were basically developed with engineering goals in mind. Despite of their applied focus, deep learning models eventually seem fruitful for cognitive purposes. This can be thought as a kind of biological exaptation, where a physiological structure becomes applicable for a function different from that for which it was selected. In this paper, it will be argued that it is time for cognitive science to seriously come to terms with deep learning, and we try to spell out the reasons why this is the case. First, the path of the evolution of deep learning from the connectionist project is traced, demonstrating the remarkable continuity, and the differences as well. Then, it will be considered how deep learning models can be useful for many cognitive topics, especially those where it has achieved performance comparable to humans, from perception to language. It will be maintained that deep learning poses questions that cognitive sciences should try to answer. One of such questions is the reasons why deep convolutional models that are disembodied, inactive, unaware of context, and static, are by far the closest to the patterns of activation in the brain visual system.

1. Introduction

The family of techniques collected under the name of *deep learning* is responsible for the current *AI Renaissance*, the fast resurgence of artificial intelligence after several decades of slow and unsatisfactory advances. In 2012, the group at the University of Toronto lead by Geoffrey Hinton, the inventor of deep learning, won ImageNet, the most challenging image classification competition. In 2016, the company DeepMind, founded by Demis Hassabis and soon acquired by Google, defeated the world champion of *Go*, the Chinese chessboard game much more complex than chess (Silver et al., 2016). The leading Internet companies were among the first in employing deep learning on a massive scale (Hazelwood et al., 2018) and are also the largest investors in research well over their own applications. Thanks to the vast success, deep learning was featured on the covers of journals such as Science (July 2015), Nature (January 2016), and The Economist (May 2015).

The astonishing success of deep learning was totally unexpected (Plebe & Grasso, 2019). The most surprising aspect is that the technology contains minor improvements from artificial neural networks, a field that was stagnating at the beginning of this century. Hinton himself was one of the protagonists of the invention of the artificial neural networks of the '80s (Hinton, McClelland, & Rumelhart, 1986; Rumelhart, Hinton, & Williams, 1986). We deem one of the most distinctive differences between the first generation of artificial neural networks and the current deep learning enterprise to be related to its focus. The primary motivation for the development of the early neural networks was the study of cognition. Most of the main players in the group that gave birth to the first artificial neural networks in the '80s were psychologists: Geoffrey Hinton, Michael Jordan, James McClelland, and David Rumelhart. In contrast, the majority of modelers in deep learning is totally indifferent to cognitive studies, with few notable exceptions like Yoshua Bengio (see for example Bengio, Courville, & Vincent, 2013; Chevalier-Boisvert et al., 2019; Ke et al., 2019). Even if several of the protagonists of deep learning are the same scientists associated with earlier artificial neural networks – Hinton included – the scope has drastically shifted towards engineering goals.

Let us provide an example. In current cognitive science the proposal of Karl Friston about the fundamental predictive activity of the brain and the related free-energy principle is well known and discussed. At the heart of his proposal there is a formal expression of free energy, derived from Bayesian variational inference (Friston, 2010; Friston & Kiebel, 2009; Friston & Stephan, 2007). On the deep learning side, an important advancement was achieved a few years ago with an architecture known as the "variational autoencoder", introduced

* Corresponding author.
E-mail addresses: perconti@unime.it (P. Perconti), aplebe@unime.it (A. Plebe).

https://doi.org/10.1016/j.cognition.2020.104365
Received 2 November 2019; Received in revised form 31 May 2020; Accepted 3 June 2020

independently by Kingma and Welling (2014) and by Rezende, Mohamed, and Wierstra (2014). Variational autoencoders in deep learning are a precise correlate of Friston's free-energy principle in the brain, and the mathematical formulations are almost the same. Curiously, Kingma & Welling glaringly neglect the connection between their new architecture and its cognitive counterpart, as do Rezende and co-authors. This striking connection is ignored in all the further refinement on the variational autoencoder in the deep learning community, and it is first acknowledged only by Ofner and Stober (2018). We are rather inclined to push the difference in scope between the earlier artificial neural network community engaged in cognitive explorations and deep learning even further. To the extent that modelers withdrew from pursuing cognitive investigations, the design of neural models was allowed much more freedom in adopting mathematical solutions alien to mental processes.

However, we argue that now deep learning, despite this recent tradition, can and should have its say in cognitive science. There is at least one simple reason: the engineering objectives of deep learning have been met with such success that, for the first time, we have artificial models performing complex cognitive tasks at human performance level. The era of toy worlds in which models are restricted to highly simplified versions of cognitive capabilities is over. We now have empirical examples of algorithms solving cognitive tasks at the full scale of complexity.

The resonance of the successes of deep learning has already stirred up reflections and discussions within cognitive science and philosophy (Cichy & Kaiser, 2019; Edelman, 2015; Lake, Ullman, Tenenbaum, & Gershman, 2017; Landgrebe & Smith, 2019; López-Rubio, 2018; Marcus, 2018; Ras, van Gerven, & Haselager, 2018; Schubbach, 2019). However, most of the focus of these reflections is about the chances and limits of deep learning in fulfilling the promises of artificial intelligence, in particular its possibility to reach the most coveted goal, the so-called "general artificial intelligence". These are important themes, but the focus of this paper is different. Our reflections are on how certain empirical achievements of deep learning may illuminate crucial debates in cognitive science. An easy and immediate consideration is that deep learning may lead to a revision of old debates in cognitive science, in which the first generation of neural networks was engaged, such as symbolic/subsymbolic, or innate/acquired knowledge. Some of these issues are already discussed in a few of the works just cited, like in Marcus (2018) and Landgrebe and Smith (2019).

But, in our opinion, the results of deep learning may play a major role in debates that characterize contemporary cognitive science. There is in particular one debate that is shaking the foundations of cognitive science: the rejection of the concepts of computation and representations (Chemero, 2009; Gelder, 1995; Hutto & Myin, 2013). The anti-representationalist and anticomputationalist stances are related to the so-called "4E cognition", i.e., embodied, embedded, enactive, and extended, which encompasses a wide variety of positions, not necessarily committed to antirepresentationalism and anticomputationalism. Embodiment has taken center stage in cognitive science for several decades (Lakoff & Johnson, 1999). Taking the body as the locus of actions, embodiment has naturally implied enactive cognition (Noë, 2004; O'Regan & Noë, 2001), and since the body interacts with its environment, embodiment also contributes to the possibility of reconciling traditional cognitive science with Gibson's ecological psychology (Gibson, 1966, 1979; Heras-Escribano, 2019). The concepts of embodiment, enactivism, embeddedness, have all certainly contributed to fundamental advances in cognitive science; however the assessment of their relative roles in cognition is currently highly controversial (Aizawa, 2015; Goldinger, Papesh, Barnhart, Hansen, & Hout, 2016; Mahon, 2015).

The performances of deep learning are disconcerting from the perspective of all these alternative stances in cognitive science, and we are surprised that it has gone almost unnoticed. Deep neural models are entirely based on representations and computations. Mostly, their main results are achieved with computations that disregard any action, any embodiment, any dynamics, any interaction with the environment. It is certainly necessary to be cautious in drawing conclusions: as mentioned above, we have to keep in mind that deep neural models are not intended as tools for studying cognition, and are not primarily intended to be biologically plausible models. Nevertheless, we believe the achievements of deep learning to be worthy of reflection on the aforementioned debates.

The paper is organized as follows. In Section 2 we will recap the fundamental role of the idea of computation and representation in cognitive science. We will then examine deep learning on the basis of an analysis of the two words of which its name is composed. In Section 3 we start with "learning", because it represents the continuity with the main tenet of connectionism, grounded in empiricism. Then, in Section 4 the concept of deep has its turn in the pivot of the discontinuity between the current neural network models and its precursors. Section 5 will describe the current challenges to the foundations of cognitive science, and Section 6 will discuss the impact of deep learning results on these challenges. We mainly address results in vision, for two reasons. First, vision is a paradigmatic case used in support of 4E cognition, and also the most successful field of application for deep learning. Second, there are several recent studies claiming that the specific deep learning models used for vision have some level of biological plausibility. Finally, in Section 7 we draw some tentative conclusions on how deep learning results may illuminate contemporary debates in cognitive science.

2. The computational turn in psychology

Until the early '90s, the representational computational theory of mind (RCTM) has been the standard in cognitive science (Fodor, 1987, 1998; Pinker, 1997), and computational psychology was considered to be a genuine scientific achievement able to solve the vexata quaestio of the mind-body problem. Nowadays, the landscape of computational psychology is deeply enriched by other perspectives, including predictive coding theories, Karl Friston free energy, and predictive engagement (Allen & Friston, 2018; Friston, 2012; Gallagher & Allen, 2018). According to predictive processing, the brain is to be considered as an inference engine working by means of Bayesian hypotheses. In this account, organisms use predictive models of the world to shape adaptive behaviors. Active inference in the brain takes advantage of internal generative models to predict incoming sensory data aiming at maintaining the best possible balance between the organism and its ecological niche (Constant, Clark, Kirchhoff, & Friston, 2019) (cfr. Section 5).

Classical computational psychology is also challenged by 4E cognition (cfr. Section 5). In particular, radical enactivists argue for the idea that cognitive science can do without mental representation or use a very minimal concept of what a representation is (Gallagher, 2017; Hutto & Myin, 2013). On their side, connectionists use highly distributed representations, not sentence-like representations. While classical computational psychology stored separate blocks of information in syntactically structured representations, connectionist networks process many kinds of information across their units and weights.

Following this line of reasoning, one could think of ruling out the notion of "representation" itself from the scene of cognitive science. We should, however, not throw the baby out with the bathwater, that is, not rule out the representation itself, because we need a more ecological notion of "representation" than the classical computational one. Gualtiero Piccinini (2008), after having considered the possibility of a kind of computation without representation, claims that cognitive computations, in a wide sense, are "any process whose function is to manipulate medium-independent vehicles according to a rule defined over the vehicles" (Piccinini & Scarantino, 2010, 239). It remains, of course, to explain how this exactly happens in human brains and then if it is possible to replicate it in an artificial machine. Piccinini argues that

neural computation is sui generis: “Typical neural signals are neither continuous signals nor strings of digits”, i.e., they require graded and continuous signals, but consisting in discrete functional elements, like spikes (Piccinini & Bahar, 2013, 453).

Among others, Charles Gallistel is a key figure in making computational psychology as ecological as desired. In his perspective, and in general, in the view of structural representations, which will be further examined in Section 6.3, a given representation is understood as an abstract rule, but physically realized in neural activation patterns, which connects observable behaviors (domain) with a set of environmental inputs (co-domain) (Beck, 2013; Gallistel, 1990b). What is connected is physical, but the rule of the connection is an abstract theoretical construct. This use of the word “representation” is more than minimal, in the sense of Mark Rowlands (2006, 113–14) (see also Wheeler, 2005), according to which an action concept of representation (or, AOR) should be intentional, teleological, compositional, decoupleable from its reference, and able to misrepresent it. Being more minimal than Rowlands' minimal representation, the use of the word “representation” adopted here is immune to its more common criticism (Gallagher, 2008).

This sense of the phrase “mental representation” is available for computational purposes, but it has at the same time a naturalistic counterpart in brain activation patterns and in the environmental set of inputs able to elicit them. Thanks to this double nature, i.e., being both computational and naturalistic in kind, computational psychology is able to provide a basis for both the mathematical advances in machine learning and neuro-computational modeling.

3. The “learning” paradigm

The two most outstanding philosophical traditions within artificial intelligence are probably the rationalist and the empiricist ones. The “learning” label in “deep learning” flags unequivocally that it belongs to the empiricist party. The empiricist community in artificial intelligence obviously connects to the philosophical tradition from Locke and Hume, for whom concepts are the products of experience, and reason gets all its materials from perception. In the brave new world of computing, Alan Turing (1948) was the first to advance the idea that computers can be designed simply by letting them learn by themselves. He envisioned a machine based on distributed interconnected elements, called *B-type unorganized machine*. Turing's neurons were simple NAND gates with two inputs, randomly interconnected, and each NAND input can be connected or disconnected, and thus a learning method can “organize” the machine by modifying the connections. His idea of learning generic algorithms by reinforcing useful links and by cutting useless ones was the most farsighted of this report, anticipating the empiricist approach characteristic of deep learning. Turing made his commitment to empiricism concerning the human mind explicit (Turing, 1948, p.16):

> We believe then that there are large parts of the brain, chiefly in the cortex, whose function is largely indeterminate. […] All of this suggests that the cortex of the infant is an unorganized machine, which can be organized by suitable interfering training.

Turing's employer at the National Physical Laboratory, for which the report was produced, was not as farsighted as Turing. He dismissed the work as a “schoolboy essay” (Copeland, 2004, p.401). Therefore, this report remained hidden for decades, until upheld by Copeland and Proudfoot (1996).

In the meantime, under the influence of the newborn cognitive science, early artificial intelligence favored the rationalist side. The rationalist tradition, that started with Rene’ Descartes and was followed by philosophers such as Leibniz, Spinoza and Kant, is at the heart of the formal languages in logic developed during the last century (Novaes, 2012). The rationalist soul in artificial intelligence drew heavily from formal languages, in building models of reasoning based on symbols and symbol processing (Newell, Shaw, & Simon, 1957, 1959; Newell & Simon, 1972). Although the unorganized learning machine of Turing was still unknown, there were a few attempts to create learning devices inspired by the way neurons learn. Marvin Minsky (1954) designed SNARK (*Stochastic Neural Analog Reinforcement Computer*), the first neural computer, assembling 40 “neurons”, each made with six vacuum tubes and a motor to adjust its connections mechanically.

From about 1960, however, the rivalry between empiricists and rationalists escalated up to the emergence of distinct sociological coteries (Boden, 2008, ch.11–12). The rationalist coalition had gained a certain dominance, prompted by the impressive initial success of symbolic computing within the newborn artificial intelligence, and even Minsky was attracted to its side. In 1958 an ambitious project led by Frank (Rosenblatt, 1958, 1962) threatened to shake the rationalist supremacy. It was the Perceptron project, funded by the US Office of Naval Research and carried out at Cornell Aeronautical Laboratory, resulting in a photoelectric machine with eight “neurons” and connections that can be adjusted according to a learning rule. The project had considerably enthusiastic media coverage, with consequent irritation in the rationalist community of artificial intelligence, culminating in a stinging critique raised by Minsky and Papert (1969). They replicated the perceptron machine, with the purpose of highlighting its limitations. As remarked by Olazaran (1996) in his historical and sociological analysis of the perceptron controversy, “replication of this kind is quite unusual in science, and it occurs only when the claim under discussion is particularly important”. The simplest example of the limitations of the perceptron is the XOR function, the exclusive disjunction, that is not linearly separable and cannot be learned by machine. The attack by Minsky and Papert achieved the desired effect, marginalizing artificial neural network research for decades.

3.1. The succeeding invention of backpropagation

It was late in the ‘80s that artificial neural networks found their way, with the PDP (*Parallel Distributed Processing*) project of Rumelhart and McClelland (1986b). The basic structure of “parallel distributed” machinery is made of simple units organized into distinct layers, with unidirectional connections between each layer and the next one. This structure, known as the *feedforward network*, is preserved in most deep learning models. PDP adheres to a radical empiricist account, with models that learn any possible meaningful function from scratch, just by experience. The success of PDP was largely due to an efficient mathematical rule, known as *backpropagation*, for adapting the connections between units, from examples of the desired function between known input and output. Being $\overrightarrow{w}$ the vector of all learnable parameters in a network, and $\mathcal{L}(\overrightarrow{x}, \overrightarrow{w})$ a measure of the error of the network with parameters $\overrightarrow{w}$ when applied to the sample $\overrightarrow{x}$, the backpropagation updates the parameters iteratively, according to the following formula:

$$\overrightarrow{w_{t+1}} = \overrightarrow{w_t} - \eta \nabla_w \mathcal{L}(\overrightarrow{x_t}, \overrightarrow{w_t}) \tag{1}$$

where t spans over all available samples $\overrightarrow{x_t}$, and η is the *learning rate*.

Backpropagation actually has a longer history. The term was used the first time by Frank Rosenblatt (1962), one of the pioneers of artificial neural networks cited earlier. Rosenblatt attempted to generalize its *perceptron* architecture (Rosenblatt, 1958) with back propagation, based on a single layer, to multiple layers. His attempt was different from Eq. (1) and not especially successful. Paul Werbos (1994), by titling his book *The Roots of Backpropagation*, claimed to be the originator of the backpropagation algorithm, in his PhD thesis at Harvard (Werbos, 1974), even if he didn't use this name. The supervisor of Werbos was Karl Deutsch, one of the leading social scientists of the 20th century, and one of the first in introducing statistical methods and formal analysis in political and social sciences. The novel technique developed by Werbos aimed at testing the Deutsch-Solow model of

national assimilation and political mobilization (Deutsch, 1966) on real data. For this purpose, he used an iterative technique, termed *dynamic feedback* in which derivatives of the error estimates with respect to the parameters were computed.

Both the algorithm of Werbos and the backpropagation formulated by Rumelhart and Hinton have their roots in the gradient methods, intensively developed in the first half of the last century for several engineering applications (Curry, 1944; Levenberg, 1944; Polak, 1971). Therefore, there was a mature mathematical context in the '70s, applied to engineering problems, fertile enough for being adapted to the formulation of artificial learning.

3.2. Neural networks and cognitive science

There is little doubt that learning is the foundation of cognition for the PDP group, and the successful artificial realization of learning by backpropagation would have paved the road for the "Explorations in the Microstructure of Cognition", as the title of the book by Rumelhart and McClelland (1986b) indicates. One of the most widely cited chapters in the PDP book (Rumelhart & McClelland, 1986a) became the standard bearer of empiricism against the rationalist orthodoxy in cognitive science at that time. It presented a model of how children learn the past tense of English verbs, without any predefined rules in mind. It challenged rationalism in two specific fundamental assumptions within cognitive science: nativism and the psychological reality of processing rules. Both assumptions were strongly defended in linguistics by Noam Chomsky (1966, 1968) and in the philosophy of mind by Jerry Fodor (1981, 1983), and assumed by most cognitive scientists at the time. Rationalists certainly did not retreat at all, and their defense was a harsh criticism of the PDP approach (Fodor & Pylyshyn, 1988; Pinker & Prince, 1988). Nevertheless, connectionist models, fueled with backpropagation learning, become widespread in developmental psychology, especially in the psychology of language development (Elman et al., 1996; Gasser & Smith, 1998; Karmiloff-Smith, 1992; Landau, Smith, & Jones, 1988; MacWhinney, 1999; MacWhinney & Leinbach, 1991; Smith, 1999).

It should be noted that the empiricist turn connected to backpropagation learning was not a menace for the computational theory of mind. The PDP paradigm proposed a form of computation, whose difference from others derives precisely from the primacy of learning. In a traditional algorithm the processing steps prescribe the function to be performed between the input and the output. In a backpropagation model there is no predefined function, the processing steps define how a general model will gradually adapt to any possible function by learning.

Moreover, the PDP paradigm shares an affection for the idea of mental representations with the classical computational theory of mind. The format of representations in the PDP framework is different from classical symbolic representation, again reflecting the principle of learning. PDP representations form gradually, are never stable, and their modifications depend on experience. PDP representations found a good agreement with cognitive theories of mental concepts (Horgan & Tienson, 1989; Rosch, 1973).

To sum up, neural networks of the PDP style, precursors of deep learning, have been the most radical form of empiricism in artificial intelligence. Their success derives from the ingenious invention of backpropagation, which will also have a heavy legacy in deep models. Neural networks of the PDP generation have had an important interaction with cognitive sciences, and have preserved a fundamental role in the ideas of computation and representations.

4. From shallow to "deep"

Probably the most prominent representatives of deep learning would not disapprove of the metaphorical usage of the adjective "deep" as intellectually profound, capable of entering far into a subject. But its meaning is actually technical and rather trivial, it refers to the number of "hidden" layers in a feed-forward artificial neural network. Any feed-forward network should include an input layer, where data is read, and an output layer where results are produced. Hinton called the other layers in between "hidden", inspired by the use of this adjective in the hidden Markov models (Anderson & Rosenfeld, 2000, p. 379). Neural models can learn increasingly complex functions by augmenting the number of units, this way, however, the number of parameters to optimize in a fully connected network increases as well, and learning becomes more difficult.

Units can be added using two different options: by increasing the number of units in the existing layers, or by adding new layers. The design of efficient artificial neural network models, 40 years ago as well as today, was based more on heuristics then on theoretical assumptions (Plebe & Grasso, 2019), one such heuristic was that following the second option for augmenting the number of units is a bad idea.

In fact, it was often observed that increasing the number of units by adding layers was much less efficient than increasing the width of a single hidden layer. For example, de Villers and Barnard (1992), based on this sort of observation, concluded this way:

> We have found no difference in the optimal performance of three- and four-layered networks […] four layer networks are more prone to the local minima problem during training […] The above points lead us to conclude that there seems to be no reason to use four layer networks in preference to three layers nets in all but the most esoteric applications.

4.1. Hinton again

The "deep" addition to PDP style of the feedforward network was just a revision of this long assumed dogma of no more than one hidden layer. The point is that the difficulty in training four layer networks, i.e. with two hidden layers, was not due to an intrinsic advantage of having a wide single hidden layer, with respect to many smaller layers. It was the standard backpropagation learning algorithm that worked quite well with models with one hidden layer – now called "shallow" – and lost efficiency with more hidden layers. Hinton and Salakhutdinov (2006) succeeded in training a model with four hidden layers by inventing a novel learning strategy, called the *Deep Belief Network*. The "belief" was borrowed from the Belief Networks (Pearl, 1986), popular in expert systems, which Hinton appreciated. These networks are unrelated to neurons, the nodes are symbolic, often propositional, and the connecting arcs express conditional probabilities.

But Pearl's Belief Networks do not learn, while the core of the Deep Belief Network is, once again, in the learning strategy. It derives from a neural architecture, called *Boltzmann Machines* (Aarts & Korst, 1989), in which neurons have binary values that can change stochastically, with probability given by the contributions of the other connected neurons. Boltzmann Machines adapt their connections in an unsupervised way, with a sort of energy minimization. This is the reason for the dedication to the great Austrian physicist. The clever trick of Hinton was to take two adjacent layers in a feedforward network, and train them as Boltzmann Machines. The procedure starts with the input and the first hidden layer, so that it is possible to use the inputs of the dataset to train the unsupervised Boltzmann Machine model. Then, this model is used to generate a new dataset, just by processing all the inputs. This new set is used to train the next couple of layers. This procedure is a sort of pre-training that gives a first shape to all the connections in the network, to be further refined by ordinary backpropagation using both the inputs and the known outputs of the dataset.

This first success aroused interest for artificial neural networks from a state of relative lethargy. As a result, many of the old neural models and ideas that have been around for decades have been revived, and made fresh and deep (Schmidhuber, 2015). Examples are the *Neocognitron* by Fukushima (1980) for vision, LSTM (*Long Short-Term Memory*) units (Hochreiter & Schmidhuber, 1997) for natural language

processing, the *Reinforcement Learning* framework (Barto & Sutton, 1982) for action selection tasks.

4.2. Backpropagation again

As odd as it may seem, even the good old backpropagation has resurged again, thus dispensing modelers from using the rather complex strategy devised by Hinton with the Deep Belief Network. It was found that a few modifications made backpropagation efficient with deep models almost as well as with shallow models. The main modification is in the following equation:

$$\overrightarrow{w_{t+1}} = \overrightarrow{w_t} - \eta \nabla_w \frac{1}{M} \sum_{i}^{M} \mathcal{L}(\overrightarrow{x_i}, \overrightarrow{w_t}) \tag{2}$$

where instead of computing the gradients over a single sample t, a stochastic estimation is made over a random subset of size M of the entire dataset, and at each iteration step t a different subset, with the same size, is sampled. Despite strong similarity between Eqs. (1) and (2) the term "backpropagation" is now out of fashion. Techniques related with Eq. (2) are referred to as *stochastic gradient descent*.

This change in name gives credit to a different mathematical context, that of stochastic approximation, established by Robbins and Monro (1951). The idea is to solve the equation $f(\overrightarrow{w}) = \overrightarrow{a}$ for a vector $\overrightarrow{w}$, in the case when the function f is not observable, using samples of an auxiliary random function $g(\overrightarrow{w})$ such that $E[g(\overrightarrow{w})] = f(\overrightarrow{w})$. The solution is obtained by the following iterative equation:

$$\overrightarrow{w_{t+1}} = \overrightarrow{w_t} - \frac{\alpha}{t}(g(\overrightarrow{x_t}) - \overrightarrow{a}). \tag{3}$$

Stochastic approximation was mostly developed in engineering domains, and has turned into an ample mathematical discipline (Benveniste, Metivier, & Priouret, 1990; Kushner & Clark, 1978). This mathematical domain provided a fertile context for developing more and more efficient variations of learning techniques for deep neural networks (Bottou & LeCun, 2004; Kingma & Ba, 2014; Schmidt, Roux, & Bach, 2017).

In summary, deep learning preserves the main philosophy of radical empiricism in all its variants and applications. Its chances of functioning depend entirely on learning from experience. There is, however, a fundamental difference in aims between the first generation of artificial neural networks and deep neural models. The former was motivated primarily by "Explorations in the Microstructure of Cognition" (Rumelhart & McClelland, 1986b), as discussed in Section 3. On the contrary, the development of deep neural models is mainly driven by applications, therefore the prevailing part of the deep learning community has far less ambition or interest in exploring cognition.

5. Computationalism in decline

The popularity of the Parallel Distributed Processing (PDP) approach has been based both on the idea that brains actually compute in a parallel and distributed way and on the technical results achieved by the artificial neural networks. In other words, ecology and efficiency were the main reasons why neural networks appeared as such a promising El Dorado. Predictive coding models promise to explain higher level cognitive phenomena, like binocular rivalry (Churchland, 1996). As above mentioned, however, before the success of the deep learning networks, the efficiency of the early artificial neural networks was low. On the other side, the ecological advantages of neuro-computationalism were not stopped while the efficiency of neural networks was not satisfactory (Plebe & De La Cruz, 2016). 4E cognition introduces itself as a more radical option, able to change the framework within the phenomenon of knowledge that should be understood. And – a particularly interesting effect for cognitive science – this would happen in a way which seems incompatible with some key assumptions of computational psychology.

Even the ecological advantages of PDP have not seemed yet enough in the eyes of the theorists who are enthusiasts of 4E cognition. They sincerely appreciate that PDP computational architectures are somehow bio-inspired and intended to be plausible from an ecological point of view. But, the very idea of a static scene in which a given subject represents something in front of him by means of a computational device inside his head remains as something to be rejected. The preferred alternative is a dynamic scenario in which the subject is not engaged in representing the world in front of him, but to interact with it through an action which is oriented to a goal. "Action", in fact, is the new magic word in 4E cognition, instead of "representation". Today we could say that cognitive science is characterized by a sort of "Faustian turn", inspired by the celebrated statement in Goethe's Faust: "Am Anfang war die Tat", "In the beginning was the Action", instead of the Bible's "Word". Preferring the vocabulary of action to that of language is a typical trend in embodied cognitive science. Even the typical examples used in papers and talks reveal this shift. Instead of people representing an object which lies in front of them through linguistic resources, we now have somebody who typically tries to grasp an object located in their perceptual scene.

According to Baggs and Chemero (2018), the future of embodied cognitive science lies in unifying the two major trends, that is, radical enactivism and ecological psychology. In their words: "Both ecological psychology and enactivism reject the idea that cognition is defined by the computational manipulation of mental representations; both also focus on self-organization and nonlinear dynamical explanation. They differ in emphasis, however. Ecological psychology focuses on the nature of the environment that animals perceive and act in; enactivism focuses on the organism as an agent. Combining the two would seem to provide a complete picture of cognition: an enactive story of agency, and an ecological story of the environment to which the agent is coupled". The die is cast. Instead of the "New Synthesis" of evolutionary psychology, that is, the combination of the RCTM and evolutionary biology, proposed by Steven Pinker when computational psychology seemed at its peak, Baggs and Chemero argue that "an enactive story of agency" would be able to provide the new "complete picture of cognition", made up by the combination of radical enactivism and ecological psychology.

This new scenario is a playground for the theoretical ambitions of the 4E approach, as can be appreciated, for instance, in the attempt to extend the "Faustian turn" into traditional "static" fields of investigation, like the machinery underlying the functioning of abstract concepts and words (Borghi et al., 2019; Borghi & Binkofski, 2014). Deep learning networks are facing the same kind of challenge. Like the case of vision, which will be addressed in the next section as a revenge for computationalism, consciousness too is a field of investigation which has recently been stagnant. After having appreciated the hard problem of the qualitative side of consciousness, cognitive science is still in trouble with an ecological and – at the same time – computational account of what it means to be conscious. The explanatory gap between the functional vocabulary, and the corresponding capacity to design computational architectures, and the phenomenology of consciousness is still an epistemological puzzle. Analogously, the 4E cognition framework is a promising theoretical perspective, but it is focused on low level perception-action mechanisms. It is less encouraging for modeling high level cognitive processes. For these reasons, it is possible to consider the recent advances in deep learning networks like a sort of "revenge" for computationalism, insofar as it allow us to successfully deal with the relatively stagnant field of investigation from the computational point of view, like consciousness and vision.

This theoretical issue is exactly where two senses of the word "phenomenology" become almost the same. On one side, there is phenomenology as the well-known philosophical trend. Some of its basic ideas, like anti-reductionism, seem incompatible with the framework of cognitive science. Other features, like the dynamic account of the

interaction between the subject and the environment, conceived in an action oriented and bodily dependent way, are used as an excellent basis for the "Faustian turn" in cognitive science. On the other side, there is phenomenology used as something to explain what we are talking about when we refer to qualia, "what is it to be like" for other individuals, and the phenomenal aspects of experience. The work of leading figures of 4E cognition, like Shaun Gallagher and Dan Hutto, display both features, being both phenomenologists (in the philosophical sense) and cognitive scientists engaged in the embodied turn.

Deep learning networks, in fact, promise to be useful in this attempt to address high level cognitive processes, like consciousness both in term of accessibility and phenomenology (Mallakin, 2019). The Consciousness Prior Hypothesis by Yoshua Bengio (2017) is a paradigmatic example of this trend. His aim is exactly to extend the achievements of deep learning beyond the field of predictive and unconscious inference over low level inputs: "Instead of making predictions in the sensory (e.g. pixel) space, one can thus make predictions in this high level abstract space, which do not have to be limited to just the next time step but can relate events far away from each other in time" (Bengio, 2017). By combining the attention mechanism, able to extract some pertinent features from the background, and the ability to make predictions about the future – two typical cognitive devices involved in consciousness, Bengio realized that machine learning and deep learning networks are the best computational architectures available to model them. As we will show in the next section, the case of vision could be seen as a revenge for computationalism for the same reason, i.e., being both a stagnant field of investigation after the classical computational achievements and a challenge for a comprehensive account of its phenomenological experience.

6. Pure vision: a renaissance?

For a long time, cognitive vision research has been shaped by the computational approach outlined by the late David Marr (1982). One of the earliest yet more persuasive papers challenging this mainstream wisdom is *A critique of pure vision* by Churchland, Ramachandran, and Sejnowski (1994). The authors clarified that "pure vision" in the form they describe is just a caricature of Marr's position and other vision cognitive scientist, adopted for the convenience of the arguments. The caricatured theory of pure vision conforms to the following three tenets (p.25):

1 *The Visual World*. […] The goal of vision is to create a detailed model of the world in front of the eyes in the brain […]
2 *Hierarchical Processing*. […] At successive stages, the basic processing achievement consists in the extraction of increasingly specific features […]
3 *Dependency Relations*. Higher levels in the processing hierarchy depend on lower levels, but not, in general, vice versa.

It is difficult to find an approach to artificial vision more "shamelessly pure" than contemporary deep learning models, according to this caricatured description of "pure vision". Yet, their performances have left researchers in vision science astonished, breaking the well-established wisdom of an unbridgeable gap between artificial and natural vision. See, for example, VanRullen (2017):

> For decades, perception was considered a unique ability of biological systems, little understood in its inner workings, and virtually impossible to match in artificial systems. But this status quo was upturned in recent years, with dramatic improvements in computer models of perception brought about by 'deep learning' approaches […] For as long as I can remember, we perception scientists have exploited in our papers and grant proposals the lack of human-level artificial perception systems […] But now neural networks […] routinely outperform humans in object recognition tasks […] Our excuse is gone.

Artificial vision as solved by deep learning models is the only sign of mini–singularity (Kurzweil, 2005) we have so far.

6.1. Impure visions

Before describing how "naively pure" deep learning is, let us dwell for a bit on the less naive views that have emerged. Churchland and coworkers characterized a conception alternative to "pure vision" as *interactive vision*, by stressing the interaction of vision with other sensory systems, and its function of guiding actions. A consistent part of their work is a review of compelling neurocognitive evidence of close and complex interactions of the visual system with non-visual signals.

Soon after, Churchland Rao and Ballard (1995) dropped the "inter", launching the concept of *active vision*, in which the connection between vision and activity is essentially in the movements of the eyes. Rao and Ballard fostered the simulation of saccades, in order to improve machine vision. The progressive dismissal of Marr's concept of vision led to the rediscovery of James Gibson (1979), whose idea of direct perception dispensing heavy information processes was strongly criticized by Marr. The ecological approach of Gibson, and his celebrated concept of "affordances" fit well with the new directions in cognitive vision (Nakayama, 1994). Alongside with the "Faustian turn" (see Section 5) in cognitive science, Jeannerod (1994); Jacob and Jeannerod (2003) – among others – have placed great emphasis on the close link between vision and motor representations at the neural level. The visual processing of objects and their attributes is driven by the kind of task the subject is performing, and object affordances are transformed into specific motor representations.

O'Regan and Noë (2001); Noë (2004) push the "impure" vision even further, inventing the label *enactive vision*, where perception is not just a process useful for action, it is a sort of action by itself. While the previous accounts of ecological and active vision still rely on the notion of mental/brain representations, enactive vision raises significant concerns against the need to postulate internal representations. This position was initially moderate, as in the words of Noë (2004, p.22):

> The claim is not that there are no representations in vision. That is a strong claim that most cognitive scientists would reject. The claim rather is that the role of representations in perceptual theory needs to be reconsidered.

With the expansion of the "Faustian turn" in 4E cognition, the claim that there are no representations becomes less heretic, allowing Noë (2010) to reject visual representations without hesitation. Not surprisingly, this position is pursued even more sharply by Myin and Degenaar (2014).

The computer vision community was rather slow and reluctant in abandoning "pure vision", because of the neat advantages of the three tenets of pure vision in the design of software. However, the lesson implied in the various "impure" approaches was that trying to interpret the content of images with a static hierarchy of feature extraction is hopeless. Already Churchland et al. (1994, p.50) reported, as a marginal argument, the difficulties of machine vision systems based on "pure vision" in tasks as easy as reading bar codes. Vision became effective when treated as an interactive process oriented by the goals of the seeing agent. Therefore, attracted by the possibility of a significant gain in performances, several artificial vision systems inspired by embodiment and enactivism were designed during the '90s (Blake & Yuille, 1992). Most of these vision systems were integrated into robots (Viéville, 1997) that offer a physical surrogate of the body, but there have also been examples of systems where "action" is just simulated (Beer, 2003; De Croon, Sprinkhuizen-Kuyper, & Postma, 2009).

Soon, it turned out that the complexity in building embodied/enactive vision systems was not compensated at all by a gain in performance, and such attempts become rare and without any marketable application. The scarce results of these models would not necessarily

threaten the validity of the 4E cognition on which they are based. Put simply, it might be the case that the task of vision is just too difficult for artificial approaches. This way out, however, clashes with the results of the deep models described below.

6.2. How naive is deep learning

We will now qualify our claim that deep learning fits so well with the description of the naive "pure vision" approaches. For this purpose it is useful to dig a bit inside the rapid rise of deep learning in vision. Once again, it was Hinton who first pushed deep models towards unexpected results in vision, but with an architecture that was different from his earlier deep belief networks (see Section 4.1). Hinton and his PhD student Alex Krizhevsky, Sutskever, and Hinton (2012), adapted an old model of the PDP era, called *Neocognitron* (Fukushima, 1980) and transformed it into a deep model. It alternates layers of *S-cell* type units with *C-cell* type units, and those names are evocative of the classification in simple and complex cells by Hubel and Wiesel (1962); Hubel and Wiesel (1968). The S-units act as convolution kernels, while the C-units downsample the images resulting from the convolution by spatial averaging. The crucial difference from conventional convolution in image processing (Bracewell, 2003; Rosenfeld & Kak, 1982) is that the kernels are learned. The first version of the Neocognitron learned by unsupervised self-organization, with a winner-take-all strategy: only the weights of the maximum responding S units, within a certain area, are modified, together with those of neighboring cells. A later version (Fukushima, 1988) used a weak form of supervision: at the beginning of the training the units to be modified in the S-layer were selected manually rather than by winner-take-all. After this first sort of seeding, training proceeded in unsupervised way.

Hinton and his group called the deep version of Neocognitron *Deep Convolutional Neural Network*. Their first model has five layers of convolutions, each with a large number of different kernels, followed by three ordinary neural layers, with a total number of 60 million parameters. The model participated in the ImageNet Large-Scale Visual Recognition Challenge that has been the standard benchmark for large-scale object recognition since 2010 (Russakovsky et al., 2015). The model, now known colloquially as AlexNet, dominated the challenge, dropping the previous error rate from 26.0% down to 16.4%. This first success steered computer vision towards deep models, and many new designs continued to improve performance. The model VGG-16 (Simonyan & Zisserman, 2015), with thirteen convolutional layers and three ordinary layers, and kernels smaller than AlexNet, achieved an error of 7.3% in the 2014 ImageNet challenge, further improved to 6.7% by the Inception (or GoogleNet) model (Szegedy et al., 2015). Several refinements continued to improve performance, even surpassing those of human subjects (Rawat & Wang, 2017).

The first assumption of "pure vision", *The Visual World*, is implicit in the ImageNet benchmark. It is organized according to the hierarchy of nouns in the lexical dictionary WordNet Fellbaum (1998), in which each lexical entry is associated with hundreds of still images. Deep convolutional neural networks like AlexNet meet the other two assumptions of "pure vision" precisely. *Hierarchical Processing*: the convolutional layers are organized in a hierarchical way, with earlier convolutions extracting low level features, which in turn become the input of other convolutions that extract features at progressively higher levels. *Dependency Relations*: the network is strictly feed-forward, with higher levels depending on lower levels, but not vice versa.

Deep convolutional neural networks also ignore all of the many factors involved in natural vision indicated by 4E cognition. These models simply learn, using stochastic gradient descent (see Section 4.2), from examples made by images and the lexical description of the category of objects found in the image. The model is unaware of any contextual information about each image, any conceptual relationship between categories, any information about the poses each object can assume in space, or about the affordances exposed by objects, any information about how objects can change their aspect in time. In summary, the model learns to recognize objects in a fully disembodied and inactive way.

6.3. The plausibility objection

One may raise the objection that using state-of-the-art deep convolutional neural models in discussions about natural vision is misguiding, insofar as these models are engineered for applications, not intended to replicate how natural vision works. Therefore, their results are irrelevant for cognitive science. For sure, just calling units in deep learning models "neurons" does not imply any resemblance with the cells in the brain. Unjustified claims on the link between deep learning and neuroscience are now quite common, as in (Arel, Rose, & Karnowski, 2010, p.13):

> Recent neuroscience findings have provided insight into the principles governing information representation in the mammal brain, leading to new ideas for designing systems that represent information. […] This discovery motivated the emergence of the subfield of deep machine learning, which focuses on computational models for information representation that exhibit similar characteristics to that of the neocortex.

This claim is incorrect from a historical perspective, there is a long-standing line of research on computational models of the neocortex (Plebe, 2018), far distant from deep learning. Claims like that of Arel and co-authors are also unwarranted because not one of the improvements of deep learning over PDP models, reviewed in Section 4 are related to recent (or not recent) neuroscientific findings.

The training regime used in the most successful deep neural models is a further source of implausibility. Models for vision are typically trained with millions of static images, and as many as a thousand images for each category (Russakovsky et al., 2015). This amount of data is probably less than the equivalent experience of an infant, but there is an important difference. Once having acquired a basic knowledge of the visual environment, humans are able to learn new categories of objects and actions with a very small number of examples, and can continue to learn all their lives. These natural forms of learning are difficult to achieve with deep neural models (Parisi, Kemker, Part, Kanan, & Wermter, 2019), and far distant from the standard training regimes used in the top vision models.

Nevertheless, when limiting deep learning to convolutional models for vision, there is a growing evidence of striking analogies between patterns in these models, and patterns of voxels in the brain visual system. One of the first attempts to relate results of deep learning to the visual system was based on the idea of adding another layer at a given level of an artificial network model to predict the space of voxel response, and to train this level on sets of images and corresponding fMRI responses (Güçlü & van Gerven, 2014). Using this method Güçlü and van Gerven (2015) compared a model very similar to AlexNet (Chatfield, Simonyan, Vedaldi, & Zisserman, 2014) with fMRI data. Initially, subjects were presented with 1750 natural images and voxel responses in progressively downstream areas – from V1 up to LO (*Lateral Occipital Complex*) – were recorded. The same images were presented to the model, and the output of the convolutional layers were trained – with a simple linear predictor – to predict voxel patterns. As a result, model responses were predictive of the voxels in the visual cortex above chance, with good prediction accuracy especially in the lower visual areas. This first unexpected result was immediately followed by several other studies, using variants of the same technique (Eickenberg, Gramfort, Varoquaux, & Thirion, 2017; Khan & Tripp, 2017; Tripp, 2017), finding reasonable agreement between features computed by deep learning models and fMRI data.

An alternative method for comparing deep learning models and fMRI responses was offered by the representational similarity analysis, introduced by Kriegeskorte, Mur, and Bandettini (2009); Kriegeskorte

(2009). This method can be applied to any sort of distributed responses to stimuli, computing one minus the correlation between all pairs of stimuli. The resulting matrix is especially informative when the stimuli are grouped by their known categorical similarities. The whole idea is that the responses across the set of stimuli reflect an underlying space in which reciprocal relations correspond to relations between the stimuli. This is exactly the idea of *structural representations*, one of the fundamental concepts in cognitive science (Gallistel, 1990a; O'Brien & Opie, 2004; Plebe & De La Cruz, 2018; Shea, 2014; Swoyer, 1991). Representational similarity analysis is applied by Khaligh-Razavi and Kriegeskorte (2014) in comparing responses in the higher visual cortex, measured with fMRI in humans, and with cell recordings in monkeys, with several artificial models. This study is interesting because it includes, models with more biological plausibility in addition to AlexNet. In particular, it included VisNet (Rolls & Stringer, 2006; Stringer & Rolls, 2002; Stringer, Rolls, & Tromans, 2007; Wallis & Rolls, 1997), a highly biologically plausible model, organized into five layers, where connectivity approximates the sizes of receptive fields in V2, V2, V4, posterior inferior temporal cortex, and the inferior temporal cortex. This network learns by unsupervised self-organization (von der Malsburg, 1973; Willshaw & von der Malsburg, 1976) with synaptic modifications derived from Hebb (1949) rule. Learning includes a specific mechanism called *trace memory*, aimed at accounting for the natural dynamics of vision, where the invariant recognition of objects is learned by seeing them when moving under various different perspectives. The analysis revealed that AlexNet was significantly more similar to the structural representation of the categorical distinction animate/inanimate in the inferior temporal cortex than VisNet.

This impetus of studies on the analogies between deep convolutional models and the visual system has led to a broad discussion in the visual neuroscience community on the relevance of deep learning models for their scientific objective. Positions range from a mostly positive acceptance (Gauthier & Tarr, 2016; VanRullen, 2017) to a cautious interest (Grill-Spector, Weiner, Gomez, Stigliani, & Natu, 2018; Lehky & Tanaka, 2016; Tacchetti, Isik, & Poggio, 2018), down to more skeptical stances (Conway, 2018; Olshausen, 2014; Robinson & Rolls, 2015; Rolls, 2016). However, this ongoing discussion has hitherto eluded our point. Whatever the degree of similarity between activation in deep model layers and in areas of the brain visual system, we have found it astounding that there is a similarity at all, given the diversity of the two systems. There is no doubt that the brain visual system is embodied, enactive, and that it has developed by the continuous interaction with the environment. How is possible then, that the most naively "pure" model, disembodied, inactive, static, unaware of context, is by far the best in predicting patterns of activation in the brain visual system?

One possible answer is that there is a part of the process involved in vision that consists of extracting features at progressive and hierarchical levels, not too far from what Marr had in mind. Curiously, the rationalist tradition was deeply connected to Marr's manifesto (Egan, 1995; Newell, 1980; Pylyshyn, 1984), while the empiricist side was more fond of Marr's critics like Churchland et al. (1994). Now is the deep learning empiricist's turn to (implicitly) reverse Marr's decline caused by 4E cognition. In fact the part of Marr's theory that was especially advantageous for rationalists is his three-level distinction, understood as autonomous levels of description. The autonomy of the top – computational – level permits fully rule-based models to gain explanatory value, in cognition in general, and specifically in vision (Biederman, 1987; Draper, Baek, & Boody, 2004; Ullman, 1996). The rationalist interpretation of Marr's level distinction is controversial (Eliasmith & Kolbeck, 2015; Shagrir & Bechtel, 2018), and most of all irrelevant for deep convolutional neural models. What these models have in common with Marr's theory of vision, and what differentiates them both from 4E cognition, is the series of features collected by Churchland et al. (1994) as "pure vision". Vision in the brain is far distant from the "pure vision" assumption: visual areas certainly collect additional information from other sensorial areas, receive rich top-down feedback, trigger attentive gazing actions, and are dynamically modulated by the ongoing action plan. The suspicion is that the role of all these aspects has been gradually amplified, to the point of discarding the "pure" computational component of the vision process at all.

A possible clue to an explanation of the similarity of "pure" deep models with brain visual areas can be found in the proposal by Joel Norman (2002) of two types of processes in the visual and dorsal visual streams. The latter is seen to work in a manner well described by Gibson's ecological theory, and the ventral stream in a manner that Norman categorizes as "constructivist", in the Helmholtzian tradition (von Helmholtz, 1866). In this dual approach view the 4E instances could be confined to the dorsal stream, and the "pure" deep convolutional networks may well be a modern counterpart of the Helmholtzian constructivist–inferential process. While the independence of the two visual streams may be overstated (Schenk & McIntosh, 2010), the general idea of the parallel pathways for different aspects of visual processing is appealing. Supporting this idea, one should expect that the similarity between deep convolutional models and the brain should be stronger in areas such as V4, the inferior temporal (IT) and the Lateral Occipital Complex (LOC). The results obtained by Yamins et al. (2014) confirm the expectation, but not those recorded by Güçlü and van Gerven (2015) who found that the deep model was more predictive of the brain activity in area V1 then LOC.

In the end, for deep learning to pursue a "shamelessly and naively pure" strategy paid off. Given the various evidence, reviewed here, of similarities between patterns in deep convolutional models and in the visual cortex, one wonders whether there is a degree of "purity" in the human visual system as well.

6.4. Beyond vision: natural language processing

Vision is at the same time the most striking success of deep learning, and the case where its distance from 4E cognition is stunning, as just discussed. There is, however, more than vision. Deep learning is gradually outperforming and replacing alternative approaches in a variety of other fields, even if at a lesser pace than in vision. When deep neural models achieve state of the art performance in tasks highly related to human cognition, it is natural to ask what these models can suggest to cognitive science. This question is even more compelling when the task is the highest cognitive function of humans: language. As we recap in Section 3.2, artificial neural networks made their way into cognitive science with a language model, and they did not do this quietly.

The model of the English past tense formation by Rumelhart and McClelland (1986a) was a pure empiricist example of language learning: it takes a phonological representation of the uninflected form of an English verb as input, and predicts the phonological form of its past tense. In this model there are no explicit rules for determining a past tense morpheme or for deriving the phonological shape of that morpheme, the composition of past tense is just learned from examples. Despite – or possible because of – the harsh rebuttal of this model by Pinker and Prince (1988), artificial neural networks met significant success at the beginning of the '90s in linguistics, especially among developmental linguists. On the other side, for the rationalist component in cognitive science and linguistics the critique of Pinker and Prince was fully successful. One of their most compelling arguments concerned the simple phonological representation used in the model, the so-called Wickelfeature (Wickelgren, 1969), where a central phoneme is related to both the preceding and following ones. As shown by Pinker and Prince, this simplification is scarcely plausible and misses important aspects of the temporal order in phonological forms.

Just a couple of years after the past tense model, Jeffrey Elman (1990) proposed an alternative artificial neural structure that efficiently the issue of representing temporally ordered information by adding recurrent connections in a feedforward model. Using recurrent networks Elman demonstrated the ability to capture some basic aspects of

syntax from examples, such as noun-verb distinction, and subject-verb agreement (Elman, 1991). A limitation of Elman's recurrent networks was the difficulty in retaining memory of input events lasting more than 4–5 discrete time steps, preventing the processing of complex sentences. This issue was solved by a refinement of the basic recurrent networks called Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), with the addition of multiplicative gate units that learn to open or close access of past signals to recurrent units.

All the work done in the '80s and '90s with artificial neural networks for human language involved small experiments using toy language examples, leaving a fundamental question open: would neural models ever be able to process language at scale? Theoretical speculations oscillate between enthusiastic positive answers and skepticism. Presumably, the ongoing progress in natural language processing based on generative grammar influenced skeptical answers. The lack of linguistic structures like categories and trees, and of the many parameters and constraints of full grammatical systems, seems to form a gap far too wide to be bridged by the simple recurrent neural models.

Now with deep learning, this question is no longer of concern. There is no more need to speculate whether neural models would process aspects of full-fledged human language in a future or not: they already do it. The success of deep learning in language tasks has not been as rapid and surprising as in vision, and its performances are more distant from those of humans than in vision. Nevertheless, deep learning has now replaced approaches based on grammar in almost all applications of natural language processing, including speech recognition (Saon et al., 2017; Veselý, Ghoshal, Burget, & Povey, 2013), language comprehension (Devlin, Chang, Lee, & Toutanova, 2018; Trischler et al., 2016) and translation (Johnson et al., 2017; Vaswani et al., 2017; Zhou, Cao, Wang, Li, & Xu, 2016).

In line with the drop of cognitive interest in the deep learning community, the developers of these successful applications in natural language processing have not cared about vindicating the soundness of the intuitions of Rumelhart, McClelland, and Elman. However, for no more than a couple of years, the impressive breakthroughs of deep learning in natural language processing have started to capture the attention of the linguistic community. So, Rumelhart & McClelland found a late rematch on Pinker & Prince, thanks to Kirov and Cotterell (2018). They reproposed the celebrated neural model of the English present-to-past tense mappings, adopting the modern encoding based on recurrent neural networks. This new model obviates most of Pinker and Prince's criticisms.

Moreover, there seems to be more than one repetition of the never dormant diatribe between rationalists and empiricists in linguistics. There is a growing interest in understanding how neural networks can achieve their language skills, and what kind of linguistic knowledge is embedded in these models. Following this line of investigation, Gulordava, Bojanowski, Grave, Linzen, and Baroni (2018) evaluated the ability of recurrent neural networks to learn English subject-verb agreement over long distance, a task thought to require hierarchical structure of sentences. Above all, they tried to ascertain whether the models were able to rely on the hierarchical structure, regardless of semantics. They did it by recalling – with a hint of irreverence – Chomsky himself, by testing the sentence The colorless green ideas I ate with the chair sleep furiously, for the agreement between ideas and sleep. Their results show how recurrent neural networks acquire deep grammatical competence without any predefined rule.

Today we are witnessing a multiplication of studies aimed at assessing a variety of grammatical competences in modern recurrent neural models. The list includes auxiliary inversion in English yes/no-question formation (Fitz & Chang, 2017), negative polarity item licensing[1] (Jumelet & Hupkes, 2018; Warstadt et al., 2019), and syntactic island constraints on the filler-gap dependency[2] (Wilcox, Levy, & Futrell, 2019; Wilcox, Levy, Morita, & Futrell, 2019).

In a target article Joe Pater (2019) provided a particularly useful overview of this new line of research, and he fostered a fusion between the traditional antagonistic empiricist and rationalist approaches to natural language. Is the rationalist part available to this alliance? Not much, apparently. Berent and Marcus (2019) rejected Pater's invitation, arguing that "either those connectionist models are right, and generative linguistics must be radically revised, or they must be replaced by alternatives that are compatible with the algebraic hypothesis". Corkery, Matusevych, and Goldwater (2019) imitated Pinker and Prince in dissecting the modern version of the English past tense learning model of Kirov and Cotterell (2018), and while acknowledging significant progress with respect to Rumelhart and McClelland's original models, they concluded that "there is still insufficient evidence to claim that neural nets are a good cognitive model for this task."

So the impact of deep learning in cognitive science, in the case of language, is likely to heat up old debates. Still, there is a fundamental difference with respect to the same discussion in the '90s: the pragmatic evidence that today deep learning is the best available computational approach to language. However, the cognitive relevance of deep learning is vulnerable to the same objection that we discussed in the case of vision. Current recurrent neural models are engineered for applications, such as Google Translate, not intended to study how language is processed in humans. Therefore, their results might be irrelevant for cognitive science. Let us recall how in the case of vision, faced with the same objection, a strategy pursued by several scholars has been to find similarities between patterns in deep neural models and patterns in cortical areas. For vision we found a significant body of research on the analogies between deep convolutional models and the human visual system, with several positive results.

Nothing similar can be found for recurrent neural models. Not only is there no attempt of a mapping between components of linguistic neural models and brain areas, there is not even an idea on a correspondence between the basic recurrent units and neural circuitry in the brain. Among the few attempts in this direction, Ponte Costa, Assael, Shillingford, de Freitas, and Vogels (2017) proposed a variation of LSTM in which information is gated through units that are subtractive, therefore akin to inhibitory cells. They showed a possible mapping of this structure onto known canonical excitatory-inhibitory cortical microcircuits. Compared to deep convolutional models for vision, recurrent neural models operate at a more abstract level, far away from sensorial representations, and therefore difficult to map across cortical areas.

The level of abstraction of recurrent neural models not only prevents their mapping onto brain circuits, it also introduces important cognitive differences with respect to the way humans learn language. For example, a standard practice for deep language models initializes the recurrent neural models with a vocabulary of known words and feeds them tokenized corpora during training. This approach is certainly valid for practical purposes but departs from the way humans learn language. One of the major challenges for infant learners is precisely discovering the basic constituents of linguistic structures like words. An important step forward was achieved recently by Hahn and Baroni (2019), who demonstrated the ability of recurrent neural networks to learn from character-level inputs without word boundaries. These networks learn to track word boundaries, and to solve morphological, syntactic and semantic tasks.

In conclusion, even if to a less extent than vision, language also is a field where the breakthroughs of deep learning are so impressive that they deserve more attention in cognitive science.

[1] Negative polarity item licensing is the knowledge of which context in a sentence license the presence of negative polarity items, words such as any.

[2] Syntactic island are positions that locally block the filler-gap dependency, where the filler is a wh-word, like who, and the gap is the empty syntactic position related to the filler.

7. Conclusions

What we have argued for in the above sections is that the advances achieved using deep learning models in cognitive domains are not neutral to cognitive science. We have first attempted to delineate the framework of deep learning. From a purely mathematical point of view it appears in full continuity with artificial neural networks as established in the '80s within the PDP project, and the technical innovations are surprisingly limited. There is, however, a radical difference in the interaction with cognitive science. The PDP group proposed neural models mostly as new tools for exploring cognition, with a radical empiricist perspective of how the mind works. Instead, the deep learning research community is largely driven by application and market motivations, and indifferent to cognitive studies, with a few notable exceptions. We have analyzed the domain of artificial vision in depth, where deep learning models have reached human-like performance, and language processing, where deep learning is the best available computational approach today. Even if these achievements have been eventually a quasi-secondary outcome, given that the primary goals were engineering in kind, they raise fundamental questions to current cognitive science anyway. We have argued that the most pressing questions concern the recent cognitive science journey, away from its earlier computational and representational principles. In principle, deep neural models are not incompatible with the new synthesis sketched here between 4E cognition, including radical enactivism, and ecological psychology. However, it is puzzling that the impressive results of deep learning are achieved disregarding all theoretical indications coming from 4E cognition. The suspicion is that the role of aspects such as embodiment, enaction, dynamic aspects, contextual effects have been gradually amplified in 4E cognition, to the point of neglecting the contribution of basic computational processes. Contrary to the line of argument here, one may still insist on the irrelevance of deep learning models for cognition, imputing the analogies between patterns in the brain and models to mere coincidence. But insisting on this line of reasoning is vulnerable to a sort of "no–miracle argument" (Putnam, 1978, pp.18–19), and, at least within cognitive science, miracles are not supposed to take place.

CRediT authorship contribution statement

Pietro Perconti:Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing.**Alessio Plebe**:Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing.

References

Aarts, E., & Korst, J. (1989). *Simulated annealing and Boltzmann machines.* New York: John Wiley.

Aizawa, K. (2015). What is this cognition that is supposed to be embodied? *Plenum Press, 28*, 755–775.

Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese, 195*, 2459–2482.

Anderson, J. A., & Rosenfeld, E. (Eds.). (2000). *Talking nets: An oral history of neural networks*. Cambridge (MA): MIT Press.

Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning–A new frontier in artificial intelligence research. *IEEE Computational Intelligence Magazine, 5*, 13–18.

Baggs, E., & Chemero, A. (2018). *Radical embodiment in two directions.* (Synthese online).

Barto, A. G., & Sutton, R. S. (1982). Simulation of anticipatory responses in classical conditioning by a neuron–like adaptive element. *The Behavioral and Brain Sciences, 4*, 221–234.

Beck, J. (2013). Why we can't say what animals think. *Philosophical Psychology, 26*, 520–546.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior, 11*, 209–243.

Bengio, Y., 2017. The consciousness prior. CoRR abs/1709.08568.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*, 1798–1828.

Benveniste, A., Metivier, M., & Priouret, P. (1990). *Adaptive algorithms and stochastic approximations.* Berlin: Springer-Verlag.

Berent, I., & Marcus, G. (2019). No integration without structured representations: Response to Pater. *Language, 95*, e75–e86.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Blake, A., & Yuille, A. L. (Eds.). (1992). *Active vision*. Cambridge (MA): MIT Press.

Boden, M. (2008). *Mind as machine: A history of cognitive science.* Oxford (UK): Oxford University Press.

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of Life Reviews, 29*, 120–153.

Borghi, A. M., & Binkofski, F. (2014). *Words as social tools: An embodied view on abstract concepts.* Berlin: Springer-Verlag.

Bottou, L., & LeCun, Y. (2004). Large scale online learning. *Advances in neural information processing systems* (pp. 217–224). .

Bracewell, R. (2003). *Fourier analysis and imaging.* Berlin: Springer-Verlag.

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: Delving deep into convolutional nets. CoRR abs/1405.3531.

Chemero, A. (2009). *Radical embodied cognitive science.* Cambridge (MA): MIT Press.

Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T. H., & Bengio, Y. (2019). BabyAI: A platform to study the sample efficiency of grounded language learning. *International Conference on Learning Representations*.

Chomsky, N. (1966). *Cartesian linguistics: A chapter in the history of rationalist thought.* New York: Harper and Row Pub. Inc.

Chomsky, N. (1968). *Language and mind.* New York: Harcourt, Brace and World second enlarged edition, 1972.

Churchland, P. M. (1996). *The engine of reason, the seat of the soul: A philosophical journey into the brain.* Cambridge (MA): MIT Press.

Churchland, P. S., Ramachandran, V., & Sejnowski, T. (1994). A critique of pure vision. In C. Koch, & J. Davis (Eds.). *Large-scale neuronal theories of the brain.* Cambridge (MA): MIT Press.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23*, 305–317.

Constant, A., Clark, A., Kirchhoff, M., & Friston, K. J. (2019). Extended active inference: Constructing predictive cognition beyond skulls. *Mind and language* in press.

Conway, B. R. (2018). The organization and operation of inferior temporal cortex. *Annual Review of Vision Science, 4*, 19.1–19.22.

Copeland, J. (Ed.). (2004). *The essential Turing – Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life plus the secrets of enigma.* Oxford (UK): Oxford University Press.

Copeland, J., & Proudfoot, D. (1996). On Alan Turing's anticipation of connectionism. *Synthese, 108*, 361–377.

Corkery, M., Matusevych, Y., Goldwater, S., 2019. Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. CoRR abs/1906.01280.

Curry, H. B. (1944). The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics, 2*, 258–261.

De Croon, G. C., Sprinkhuizen-Kuyper, I. G., & Postma, E. (2009). Comparing active vision models. *Image and Vision Computing, 27*, 374–384.

de Villers, J., & Barnard, E. (1992). Backpropagation neural nets with one and two hidden layers. *IEEE Transactions on Neural Networks, 4*, 136–141.

Deutsch, K. W. (1966). *The nerves of government: Models of political communication and control.* New York: Free Press.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805.

Draper, B. A., Baek, K., & Boody, J. (2004). Implementing the expert object recognition pathway. *Machine Vision and Applications, 16*, 27–32.

Edelman, S. (2015). The minority report: Some common assumptions to reconsider in the modelling of the brain and behaviour. *Journal of Experimental & Theoretical Artificial Intelligence, 28*, 751–776.

Egan, F. (1995). Computation and content. *The Philosophical Review, 104*, 181–203.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage, 152*, 184–194.

Eliasmith, C., & Kolbeck, C. (2015). Marr's attacks: On reductionism and vagueness. *Topics in Cognitive Science, 7*, 323–335.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–221.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195–225.

Elman, J. L., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness a connectionist perspective on development.* Cambridge (MA): MIT Press.

Fellbaum, C. (1998). *WordNet.* Malden (MA): Blackwell Publishing.

Fitz, H., & Chang, F. (2017). Meaningful questions: The acquisition of auxiliary inversion in a connectionist model of sentence production. *Cognition, 166*, 225–250.

Fodor, J. (1981). *Representations: Philosofical essay on the foundation of cognitive science.* Cambridge (MA): MIT Press.

Fodor, J. (1983). *Modularity of mind: And essay on faculty psychology.* Cambridge (MA): MIT Press.

Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind.* Cambridge (MA): MIT Press.

Fodor, J. (1998). *Concepts: Where cognitive science went wrong.* Oxford (UK): Oxford University Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71.

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

Friston, K. (2012). A free energy principle for biological systems. *Entropy, 14*, 2100–2121.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free–energy principle. *Philosophical Transactions of the Royal Society B, 364*, 1211–1221.

Friston, K., & Stephan, K. E. (2007). Free–energy and the brain. *Synthese, 159*, 417–458.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193–202.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks, 1*, 119–130.

Gallagher, S. (2008). Are minimal representations still representations? *International Journal of Philosophical Studies, 16*, 351–369.

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind.* Oxford (UK): Oxford University Press.

Gallagher, S., & Allen, M. (2018). Active inference, enactivism and the hermeneutics of social cognition. *Synthese, 195*, 2627–2648.

Gallistel, C. R. (1990a). *The Organization of Learning.* Cambridge (MA): MIT Press.

Gallistel, C. R. (1990b). Representations in animal cognition: An introduction. *Cognition, 37*, 1–22.

Gasser, M., & Smith, L. B. (1998). Learning nouns and adjectives: A connectionist account. *Language & Cognitive Processes, 13*, 269–306.

Gauthier, I., & Tarr, M. J. (2016). Visual object recognition: Do we (finally) know more now than we did? *Annual Review of Vision Science, 2*, 16.1–16.20.

Gelder, T.v. (1995). What might cognition be, if not computation? *Journal of Phylosophy, 91*, 345–381.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston (MA): Houghton Miflin.

Gibson, J. J. (1979). *The ecological approach to perception.* Boston (MA): Houghton Miflin.

Goldinger, S. D., Papesh, M. H., Barnhart, A. S., Hansen, W. A., & Hout, M. C. (2016). The poverty of embodied cognition. *Psychonomic Bulletin & Review, 23*, 171–182.

Grill-Spector, K., Weiner, K. S., Gomez, J., Stigliani, A., & Natu, V. S. (2018). The functional neuroanatomy of face perception: From brain measurements to deep neural networks. *Interface Focus, 8*, 20180013.

Güçlü, U., & van Gerven, M. A. J. (2014). Unsupervised feature learning improves prediction of human brain activity in response to natural images. *PLoS Computational Biology, 10*, 1–16.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience, 35*, 10005–10014.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *Proceedings North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 1195–1205). Association for Computational Linguistics.

Hahn, M., & Baroni, M. (2019). *Tabula* nearly *rasa*: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text. *Transactions of the Association for Computational Linguistics, 7*, 467–484.

Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., ... Wang, X. (2018). Applied machine learning at Facebook: A datacenter infrastructure perspective. *IEEE International Symposium on High Performance Computer Architecture (HPCA)* (pp. 620–629). .

Hebb, D. O. (1949). *The organization of behavior.* New York: John Wiley.

Heras-Escribano, M. (2019). *The philosophy of affordances.* London: Palgrave Macmillan.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). *Distributed representations.* Rumelhart and Mc Clelland (1986b)77–109.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 28*, 504–507.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.

Horgan, T., & Tienson, J. (1989). Representations without rules. *Philosophical Topics, 17*, 147–174.

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology, 160*, 106–154.

Hubel, D., & Wiesel, T. (1968). Receptive fields and functional architecture of mokey striate cortex. *Journal of Physiology, 195*, 215–243.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content.* Cambridge (MA): MIT Press.

Jacob, P., & Jeannerod, M. (2003). *Ways of seeing – The scope and limits of visual cognition.* Oxford (UK): Oxford University Press.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *The Behavioral and Brain Sciences, 17*, 187–245.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics, 5*, 339–351.

Jumelet, J., Hupkes, D., 2018. Do language models understand *anything*? On the ability of LSTMs to understand negative polarity items. CoRR abs/1808.10627.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science.* Cambridge (MA): MIT Press.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Pal, C., Bengio, Y., 2019. Learning neural causal models from unknown interventions. CoRR abs/1906.01280.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology, 10*, e1003915.

Khan, S., Tripp, B. P., 2017. One model to learn them all. CoRR abs/1706.05137.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *Proceedings of International Conference on Learning Representations*.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *Proceedings of International Conference on Learning Representations*.

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics, 6*, 651–666.

Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience, 3*, 363–373.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2009). Representational similarity analysis – Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems* (pp. 1090–1098). .

Kurzweil, R. (2005). *The singularity is near: When humans trascend biology.* New York: Viking.

Kushner, H. J., & Clark, D. (1978). *Stochastic approximation methods for constrained and unconstrained systems.* Berlin: Springer-Verlag.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *The Behavioral and Brain Sciences, 40*, 1–72.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh. The embodied mind and its challenge to Western thought.* New York: Basic Books.

Landau, B., Smith, L. B., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299–321.

Landgrebe, J., & Smith, B. (2019). Making AI meaningful again. *Synthese,* 1–21. https://doi.org/10.1007/s11229-019-02192-y.

Lehky, S. R., & Tanaka, K. (2016). Neural representation for object recognition in inferotemporal cortex. *Current Opinion in Neurobiology, 37*, 23–35.

Levenberg, K. (1944). A method for solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics, 2*, 164–168.

López-Rubio, E. (2018). Computational functionalism for the deep learning era. *Minds and Machines, 28*, 667–688.

MacWhinney, B. (Ed.). (1999). *The emergence of language*(2nd ed.). Mahwah (NJ): Lawrence Erlbaum Associates.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 29*, 121–157.

Mahon, B. Z. (2015). What is embodied about cognition? *Language and Cognitive Neuroscience, 30*, 420–429.

Mallakin, A. (2019). An integration of deep learning and neuroscience for machine consciousness. *Global Journal of Computer Science and Technology, 19*, 1–10.

Marcus, G., 2018. Deep learning: A critical appraisal. CoRR abs/1801.00631.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* Freeman, San Francisco (CA): W. H.

Minsky, M. (1954). Neural nets and the brain-model problem. *Ph.D. thesis*. Princeton University.

Minsky, M., & Papert, S. (1969). *Perceptrons.* Cambridge (MA): MIT Press.

Myin, E., & Degenaar, J. (2014). Enactive vision. In L. Shapiro (Ed.). *The Routledge handbook of embodied cognition* (pp. 90–107). London: Routledge.

Nakayama, K. (1994). James J. Gibson – An appreciation. *Psychological Review, 101*, 329–335.

Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135–183.

Newell, A., Shaw, C., & Simon, H. A. (1957). Empirical explorations of the logic theory machine: A case study in heuristic. *Western Joint Computer Conference Proceedings* (pp. 218–230). New York: ACM.

Newell, A., Shaw, C., & Simon, H. A. (1959). Report on a general problem-solving program. *Scientific report P-1584*. RAND: Corporation, Santa Monica (CA).

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs (NJ): Prentice Hall.

Noë, A. (2004). *Action in perception.* Cambridge (MA): MIT Press.

Noë, A. (2010). Vision without representation. In N. Gangopadhyay, M. Madary, & F. Spicer (Eds.). *Perception, action, and consciousness: Sensorimotor dynamics and two visual systems* (pp. 245–256). Oxford (UK): Oxford University Press.

Norman, J. (2002). Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches. *The Behavioral and Brain Sciences, 25*, 73–144.

Novaes, C. D. (2012). *Formal languages in logic: A philosophical and cognitive analysis.* Cambridge (UK): Cambridge University Press.

O'Brien, G., & Opie, J. (2004). Notes toward a structuralist theory of mental representation. In H. Clapin, P. Staines, & P. Slezak (Eds.). *Representation in mind – New approaches to mental representation.* Amsterdam: Elsevier.

Ofner, A., & Stober, S. (2018). Towards bridging human and artificial cognition: Hybrid variational predictive coding of the physical world, the body and the brain. *Advances in neural information processing systems*.

Olazaran, M. (1996). A sociological study of the official history of the perceptrons controversy. *Social Studies of Science, 26*, 611–659.

Olshausen, B. A. (2014). Perception as an inference problem. In M. S. Gazzaniga (Ed.). *The cognitive neurosciences* (pp. 295–304). (5th ed.). Cambridge (MA): MIT Press.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *The Behavioral and Brain Sciences, 24*, 939–1031.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks, 113*, 54–71.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language, 95*, e41–e74.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence, 29*, 241–288.

Piccinini, G. (2008). Computation without representation. *Philosophical Studies, 137*, 205–241.

Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science, 34*, 453–488.

Piccinini, G., & Scarantino, A. (2010). Computation vs. information processing: Why their difference matters to cognitive science. *Studies in History and Philosophy of Science, 41*, 237–246.
Pinker, S. (1997). *How the mind works.* New York: Norton.
Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*, 73–193.
Plebe, A. (2018). The search of "canonical" explanations for the cerebral cortex. *History and Philosophy of the Life Sciences, 40*, 40–76.
Plebe, A., & De La Cruz, V. M. (2016). *Neurosemantics – Neural processes and the construction of linguistic meaning.* Berlin: Springer.
Plebe, A., & De La Cruz, V. M. (2018). Neural representations beyond "plus X". *Minds and Machines, 28*, 93–117.
Plebe, A., & Grasso, G. (2019). The unbearable shallow understanding of deep learning. *Minds and Machines, 29*, 515–553.
Polak, E. (1971). *Computational methods in optimization: A unified approach.* New York: Academic Press.
Ponte Costa, R., Assael, Y. M., Shillingford, B., de Freitas, N., & Vogels, T. P. (2017). Attention is all you need. *Advances in neural information processing systems* (pp. 272–283). .
Putnam, H. (1978). *Meaning and the moral sciences.* London: Routledge.
Pylyshyn, Z. (1984). *Computation and cognition.* Cambridge (MA): MIT Press.
Rao, R. P., & Ballard, D. H. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence, 78*, 461–505.
Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. van Gerven (Eds.). *Explainable and interpretable models in computer vision and machine learning* (pp. 19–36). Berlin: Springer-Verlag.
Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation, 29*, 2352–2449.
Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In E. P. Xing, & T. Jebara (Eds.). *Proceedings of Machine Learning Research* (pp. 1278–1286). .
Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics, 22*, 400–407.
Robinson, L., & Rolls, E. T. (2015). Invariant visual object recognition: Biologically plausible approaches. *Biological Cybernetics, 109*, 505–535.
Rolls, E. (2016). *Cerebral cortex: Principles of operation.* Oxford (UK): Oxford University Press.
Rolls, E. T., & Stringer, S. M. (2006). Invariant visual object recognition: A model, with lighting invariance. *Journal of Physiology – Paris, 100*, 43–62.
Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.). *Cognitive development and acquisition of language*. New York: Academic Press.
Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organisation in the brain. *Psychological Review, 65*, 386–408.
Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptron and the theory of brain mechanisms.* Washington (DC): Spartan.
Rosenfeld, A., & Kak, A. C. (1982). *Digital picture processing* (2nd ed.). New York: Academic Press.
Rowlands, M. (2006). *Body language.* Cambridge (MA): MIT Press.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536.
Rumelhart, D. E., & McClelland, J. L. (1986a). *On learning the past tenses of English verbs.* Rumelhart and Mc Clelland (1986b)216–271.
Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge (MA): MIT Press.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*, 211–252.
Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., et al. (2017). English conversational telephone speech recognition by humans and machines. *Conference of the International Speech Communication Association* (pp. 132–136). .
Schenk, T., & McIntosh, R. D. (2010). Do we have independent visual streams for perception and action? *Cognitive Neuroscience, 1*, 52–62.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117.
Schmidt, M., Roux, N. L., & Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming, 162*, 83–112.
Schubbach, A. (2019). Judging machines: Philosophical aspects of deep learning. *Synthese,* 1–21. https://doi.org/10.1007/s11229-019-02167-z.
Shagrir, O., & Bechtel, W. (2018). Marr's computational level and delineating phenomena. In D. M. Kaplan (Ed.). *Explanation and integration in mind and brain science* (pp. 190–214). Oxford (UK): Oxford University Press.
Shea, N. (2014). Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society, 114*, 123–144.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*, 484–489.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.
Smith, L. B. (1999). *Children's noun learning: How general learning processes make specialized learning mechanisms.* MacWhinney (1999).
Stringer, S. M., & Rolls, E. T. (2002). Invariant object recognition in the visual system with novel views of 3d objects. *Neural Computation, 14*, 2585–2596.
Stringer, S. M., Rolls, E. T., & Tromans, J. M. (2007). Invariant object recognition with trace learning and multiple stimuli present during training. *Network: Computation in Neural Systems, 18*, 161–187.
Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese, 87*, 449–508.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 1–9). .
Tacchetti, A., Isik, L., & Poggio, T. A. (2018). Invariant recognition shapes neural representations of visual input. *Annual Review of Vision Science, 4*, 403–422.
Tripp, B. P. (2017). Similarities and differences between stimulus tuning in the inferotemporal visual cortex and convolutional networks. *International Joint Conference on Neural Networks* (pp. 3551–3560). .
Trischler, A., Ye, Z., Yuan, X., He, J., Bachman, P., Suleman, K., 2016. A parallel-hierarchical model for machine comprehension on sparse data. CoRR abs/1603.08884.
Turing, A. (1948). Intelligent machinery. Tech. rep., National Physical Laboratory, London. In D. C. Raccolto in Ince (Ed.). *Collected works of A. M. Turing: mechanical intelligence* (pp. 1969). Edinburgh University Press.
Ullman, S. (Ed.). (1996). *High-level vision – Object recognition and visual cognition*. Cambridge (MA): MIT Press.
VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology, 8*, 142.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* (pp. 6000–6010). .
Veselý, K., Ghoshal, A., Burget, L., & Povey, D. (2013). Sequence-discriminative training of deep neural networks. *Conference of the International Speech Communication Association* (pp. 2345–2349). .
Viéville, T. (1997). *A few steps towards 3D active vision.* Berlin: Springer-Verlag.
von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetic, 14*, 85–100.
von Helmholtz, H. (1866). Handbuch der physiologische Optik. Voss, Hamburg, english translation in 1925. *Treatise on physiological optic*. New York: Dover Pub.
Wallis, G., & Rolls, E. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51*, 167–194.
Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., Wang, S.-F., Phang, J., Mohananey, A., Htut, P. M., Jeretic, P., Bowman, S. R., 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. CoRR abs/1909.02597.
Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. *Ph.D. thesis*. Harvard University.
Werbos, P. (1994). *The roots of backpropagation: From ordered derivatives to neural networks.* New York: John Wiley.
Wheeler, M. (2005). *Reconstructing the cognitive world: The next step.* Cambridge (MA): MIT Press.
Wickelgren, W. (1969). Context sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review, 76*, 1–15.
Wilcox, E., Levy, R., & Futrell, R. (2019). Hierarchical representation in neural language models: Suppression and recovery of expectations. *Proceedings BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 181–190). Association for Computational Linguistics.
Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2019). What syntactic structures block dependencies in RNN language models? In A. Goel, C. Seifert, & C. Freksa (Eds.). *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 1199–1205). Cognitive Science Society.
Willshaw, D. J., & von der Malsburg, C. (1976). How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London, B194*, 431–445.
Yamins, D. L. K., Honga, H., Cadieua, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the Natural Academy of Science USA, 23*, 8619–8624.
Zhou, J., Cao, Y., Wang, X., Li, P., & Xu, W. (2016). Deep recurrent models with fast-forward connections for neural machine translation. *Transactions of the Association for Computational Linguistics, 4*, 371–383.