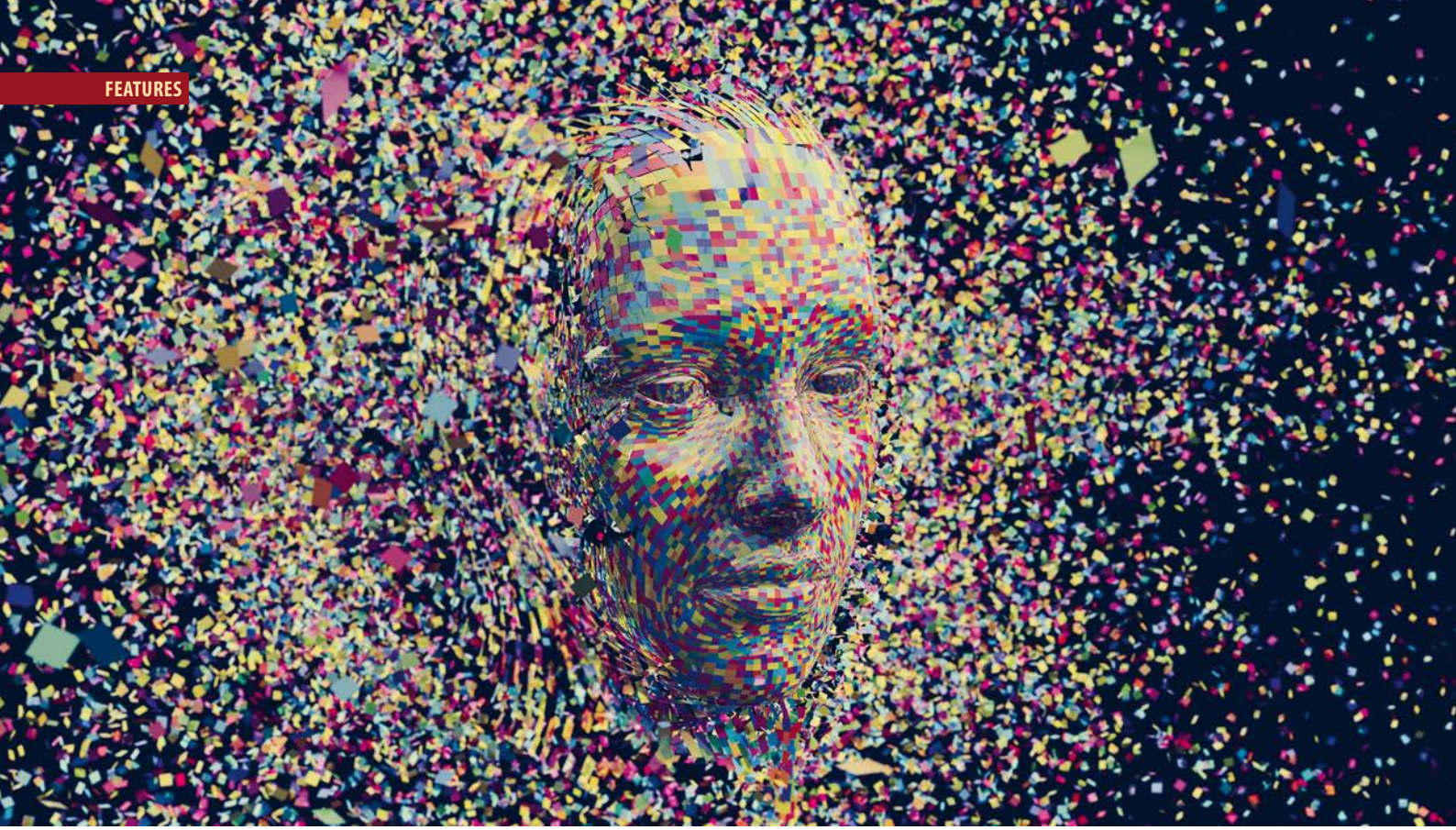


ARTIFICIAL INTELLIGENCE AND ITS LIMITS[1]

Marc Mézard, École normale supérieure, PSL University – DOI: https://doi.org/10.1051/epn/2018505

By now, everybody should know that the recent progress of Artificial Intelligence (AI) is about to have a dramatic impact on many sectors of human activity. In the last ten years, we have seen spectacular breakthroughs on applications of AI, and much more is to come, but should we speak of "intelligence"?

After more than 50 years of existence – with results that did not really match the expectations-, AI has been recently revolutionized, notably through the use of "deep neural networks". Last year, a new algorithm managed to beat the world's best Go player. And now, we can automatically process images, segment them and provide a semantic description of their content. Moreover, voice recognition and automatic translation are progressing rapidly. Most importantly, algorithms are competing with the best professionals at analyzing skin cancer symptoms or detecting specific anomalies in radiology.

It is likely that many aspects of our society -including work organization- will be completely reshaped by these new technologies. Here are few examples among many others:

- Artificial vision and scene analysis are opening the road to autonomous cars and trucks. With 13 million heavy trucks on the EU roads, the impact of these changes in the next decades will be major.
- In the health sector, it is clear that AI will soon change medical practice. Algorithms to assist in medical diagnosis will be developed, starting with very specific problems, before gradually evolving towards more general issues. One can anticipate that the very activity of medical doctors will be deeply transformed.

[1] This text is an expanded version of a tribune that has appeared in EPN49/2

- In many sectors, the monitoring of individual behaviors and the feeding of algorithmic recommendation systems are currently revolutionizing commercial activities at large and relations to clients. They also change the basic rules and principles of personal and professional insurance.
- This surveillance could also be used to control individuals in totalitarian regimes.
- The possibility of autonomous-decision-making robots can open new sectors of commercial activities, but it can also lead to the terrifying perspective of warrior-robots.

Predicting the future is always difficult, and this is a challenge that AI is not yet ready to meet! AI-fanatics predict a radical change in our societies, the end of labor, a much better medicine leading to much longer lifetime, and some of them foresee the emergence of new machine-enhanced "human" beings. Others, the pessimistic ones, predict the end of civilization and the advent of a society in which robots will take power. As for the most cautious colleagues, they already envision major professional changes, and for the first time they consider changes which will not only affect low-wage activities, but mainly intermediate, and sometimes highly specialized jobs, in which repetitive tasks can be well-modelled and taught -such as radiology, law, software development, *etc.*

The rivalry in the leadership of AI high-technology has already started. This leadership is supposed to give the capacity of leading tomorrow's world (and perhaps of dominating it militarily). In the US, DARPA has just launched a $2 billion campaign, "AI next", that aims at exploring "how machines can acquire human-like communication and reasoning capabilities, with the ability to recognize new situations and environments and adapt to them". China has announced its ambition to become the world leader in AI, and invests several billion dollars in a AI technology research center in Beijing. The multinational companies - including the GAFAM[1] and BATX[2]- that currently dominate the data-AI world, are making major investments of tens of billions of dollars.

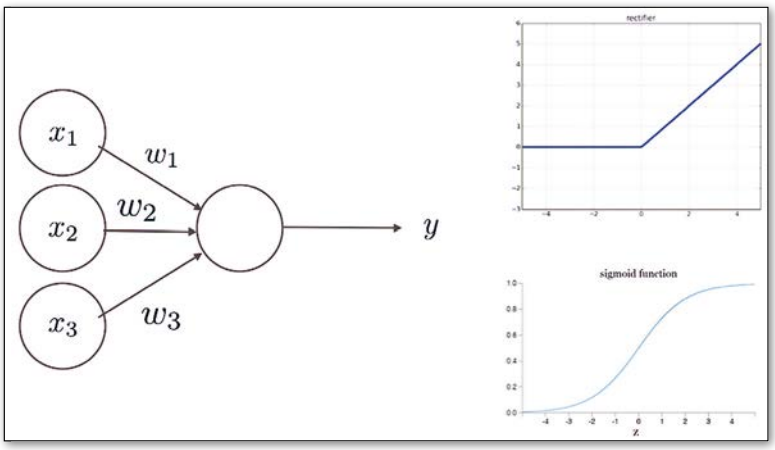
Having recognized the importance of this technological revolution (even if I am unable to predict it in its full depth), I would like to challenge the term "intelligence" in AI. This requires us to have a closer look at the way these new machines operate. Fortunately, it is rather easy.

The recent breakthrough is based on "machine learning", in which the machine is programmed to learn by itself, from examples. In deep networks, the machine is an artificial neural network, built from millions of elementary units, artificial "neurons", that somewhat mimic the activity of neurons in the brain (figure 1). Each neuron receives information from neurons in the previous layer, performs a simple computation (typically a weighted sum of the inputs, to which a threshold function is then applied) and in turn sends a few bits of information to the next layer. A modern "deep" network with hundreds of layers, analyzing an image, can contain hundreds of millions of adaptable parameters ruling these elementary computations (figure 2). They must be determined through supervised learning. For this we need a large "training set" of examples for which the teacher of the machine knows the answer.

Imagine for example that you want to teach a neural network how to distinguish whether a given picture shows a cat or a dog. Although a few-years old child easily answers this question, this remained for years a main challenge in computer vision, and it was only solved recently, by deep networks. A crucial ingredient is the availability of large database showing cats and dogs pictures. Then you design a layered network, where the input layer contains one neuron per pixel of the image, and the last layer has two neurons, one for "cat", one for "dog". The design of the rest of the network, how many layers, how many neurons in each layer, the nature of the non-linear function used by each neuron, is an art: there is no theory or model guiding the designer, but she uses quite a lot of accumulated experience, of know-how, and of trial-and-error. Experimental evidence indicates that, in practice, learning is easier when you use a 'deep network", one that contains many (tens or hundreds) intermediate layers.

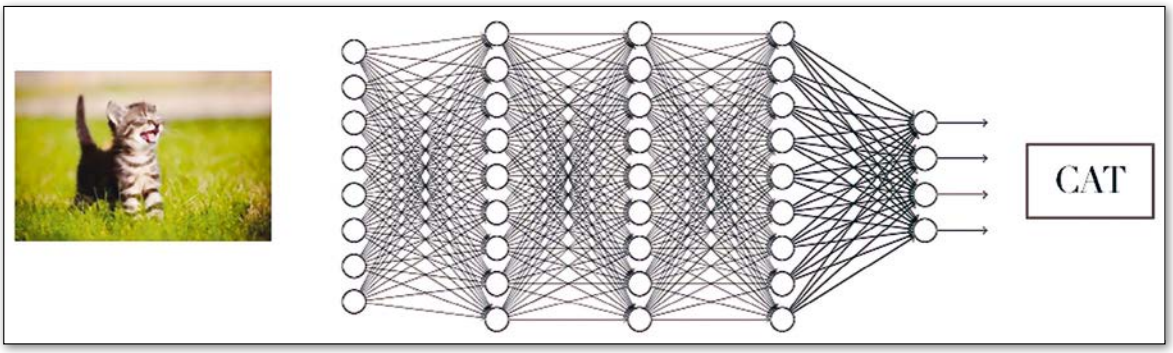
Ideally, one would like a machine that, whenever there is a cat on the image, outputs a 1 on the output "cat" neuron, and the same for dogs. Given a machine, namely a set

▾ **FIG. 1:** Artificial neurons. Left : Given the values x_1, x_2, x_3,... that it receives from other neurons, the neuron computes its output y as follows: it first compute a linear combination of the inputs, $s=w_0+w_1*x_1+w_2*x_2+w_3*x_3+\ldots$, and then outputs $y=f(s)$, where f is a nonlinear function, which can be for instance a linear rectifier (right, top) or a sigmoid function (right, bottom). The parameters w_0, w_1, w_2, w_3,... must be learnt by the machine during the training phase



[2] Google, Apple, Facebook, Amazon and Microsoft

[3] Baidu, Alibaba, Tencent, Xiaomi



▲ **FIG. 2:** In a layered network, the input is presented to the left layer. The signal is then propagated to the right, each neuron doing its weighted sum of the inputs it receives, followed but the non-linear function. The output is read in the last layer. Supervised learning is done by presenting a large database of images, each associated with the desired output, and updating the weights in order to get the desired output for each item in the database.

of values of weight parameters, one can define an error, a function of these parameters that measures the distance between the performance of the machine and the ideal one. In this supervised learning process, one optimizes the values of the parameters, using typically a stochastic gradient descent which iteratively improves each of the parameters in the direction which lowers the error. In practice, you have to train the machine using hundreds of thousands of images, with a supervisor telling in each case whether there is a cat or not. Having found the parameters of the machine such that it performs well on the training set, the real issue is the machine's ability to generalize. This is tested on a new dataset, distinct from the one used for training.

This paradigm of supervised learning in neural networks has existed for over 50 years. However, until the field's recent revival, it was not successful on real-size practical applications. Its revival is due to the increase in computing power, to the availability of very large labeled datasets for training (in fact, the development of "big-data" and the progress in machine learning are strongly correlated), and to some clever network-design know-how, pre-processing and training tricks developed in the 2000's.

In spite of its practical success, the scientific understanding of deep networks lags far behind. The learning process is poorly understood. Gradient descent in a complicated 108-dimensional parameter space should typically be trapped in inefficient regions. If one defines the training error -measuring the number of images that are misclassified by the machine- as an "energy function", the learning process amounts to finding the lowest energy configuration -the ground state- of a statistical mechanics problem with 108 variables (the adaptable parameters of the machine). This energy is a complicated function which depends on all the examples presented during training. Such large disordered systems have been much studied in the last four decades, in order to understand glasses. The result is well-known: typically, the glass energy landscape is very rough, with "traps" at all scales. The relaxation dynamics of stochastic gradient descent in a glass is extremely slow, making it very hard to reach the ground state. Yet, in practice, in all problems mentioned above, and many other ones, training in deep networks finds a good-enough set of parameters, producing a machine that can be smarter than us at some tasks. It is therefore an experimental fact that the learning problem in deep networks seems to have a much smoother energy landscape than expected, particularly when the number of layers becomes large. Why is it so? In spite of the many theoretical papers on this issue, proposing as many conjectures and ideas, it is fair to say that this is still a mystery, and I will refrain from presenting here these conjectures (including mine)!

So here we are: after observing many labelled examples, the machine has found the values of millions of parameters, and it does perform well. But what is our understanding of its performance? I claim that, if we know everything, we understand very little. On the one hand, we know everything at the microscopic level: we can observe all the operations that every neuron is doing, and we can read out all the parameters that it uses in order to perform these operations. But this is the same situation as that of an observer who would only see the microscopic structure of a computer: she could list all the transistors, how they are connected to other transistors, *etc.*, but still totally miss the point of what the computer is doing, how the information is stored and transformed.

We are slightly better-off in our understanding of layered network, but not much. We can see experimentally that the information obtained in each layer becomes more and more high level and global when one gets deeper into the network. For instance, in image recognition, the first layers tend to be sensitive to small scale patterns like local edges, and progressing deeper into the network we will find layers that are sensitive to lines and contours, then to specific patterns, eyes maybe, or whiskers, and in the last layer the abstract information comes out: this is a cat! Information is stored collectively in each layer: each neuron separately does not

know anything, it is only by looking collectively at the activity of a large group of neurons that one can see information emerging.

This phenomenon of emergence is crucial. It is well known in statistical physics: for instance, the notions of pressure or entropy are "emerging" concepts that make sense only in presence of many particles, as the result of a collective behavior. Similarly, it was already understood in the 80's that the storage of information in a neural network is radically different from the one used in standard computers. In a standard computer, if you flip a bit, the information is changed radically. In a neural network, the information is kind of delocalized in the activity of a large number of neurons. A mistake in one of them does not change much the information. Similarly, the values of weight parameters do not need to be fine-tuned: a rough approximation is enough, the final behavior of the machine is robust to small changes.

Emergence is a complicated phenomenon, and at the moment we understand little of how information is processed in smart deep networks. One might ask whether this lack of understanding is actually a problem: after all, if we don't understand the machine that nevertheless functions well, who cares? Actually, it is a problem, because without a clear understanding we are not able to give any guarantee that our smart deep network always performs the task for which it was trained. A particularly nasty case is that of adversary examples. A group of our colleagues has worked with a neural network that was trained at distinguishing a panda from a gibbon, and had excellent performance. Then they picked a picture of a panda, and they were able to change a very small fraction of the pixels in such a way that this tiny alteration, totally invisible to our eye, which fooled the machine: on this slightly altered image, the network answered that our panda was a gibbon. The existence of such "adversarial examples", that can be obtained by automatic learning, may seem anecdotal when your aim is to identify species of apes, but it no longer is when you think that someone could fool a machine that is supposed to identify a "STOP" road sign and action the breaks. The absence of guarantee, linked to our poor understanding of the real processes at work, can be a serious obstacle to many practical applications, and a real nightmare for legal issues.

The second big problem raised by the absence of understanding is much deeper, and relates to the very notion of "intelligence". I will argue that, however smart these machines are, as far as "intelligence" is concerned, they are very limited. They can certainly achieve specific tasks, characterized by simple answers, in a well-defined setup, and in this they can be very useful. But they are far from elaborating a representation of the world, and even further away from any kind of creative reasoning. This raises fundamental questions. Is our brain more than a machine that reacts to inputs and produces outputs (maybe with some degree of stochasticity)? Let me mention here one small aspect of this question, dealing with science. An extreme position has been taken by Chris Anderson, chief Editor of "Wired". In a 2008 paper he declared "the end of the scientific method", arguing that the traditional way of doing science, building models, putting forward hypotheses, testing them and modifying the model, is obsolete because data science, and AI, allow to practice science purely on the basis of correlations, without any need for models and theories. Imagine that we have taken many movies of falling objects, and trained with these movies a deep network that is then able to determine the trajectory of objects of various sizes and shapes thrown in the air, as precisely as the solution of Newton's equations incorporating friction, wind speed *etc.* This network is a nice device, maybe useful for gun-manufacturers, for instance. But it stays very far behind the "model" described by Newton's laws in several key aspects. First of all, it does not capture the generality and the universality of a law: it can never figure out that the same law describes the move of planets around the sun. Secondly, a model or a law has a major virtue, its compactness, which will make it possible to use it as a building block for further developments, by combining it with other laws, equations, and models. This is one of the major ingredients of intelligence applied to the description of the world: it creates a concise, workable and predictive representation of the world, built of elements that can be combined. You start from Newton's law, then work out the approximation for an object at the surface of the earth, combine it with the laws of friction, and there you are. But you can also use it on the moon, or understand its limitations and discover a new theory, relativity…

The spectacular progress in AI is a major technological breakthrough. New machines will be able to make decisions, or, if we implement appropriate controls, to help us making decisions. They will affect our lives, for better or for worse. But they are very far from being intelligent. ■

About the Author

After graduating from École normale supérieure, **Marc Mézard** became a CNRS research associate in 1981 and obtained his PhD in 1984 on spin glass theory. In 2001 he moved to Université Paris Sud as CNRS research director, and became director of the Laboratoire de physique théorique et modèles statistiques, and of the "Labex" Physics, Atoms, Light, Matter. Between 1987 and 2012 he worked as associate professor and then professor at École Polytechnique. Since 2012 he is the director of École normale supérieure in Paris. His main field of research is statistical physics of disordered systems and its use in various branches of science.