Journal of
Computational and Theoretical Nanoscience
Vol. 8, 1–6, 2011

Reading What Machines "Think": A Challenge for Nanotechnology

Stefano Prezioso[1,*], Danilo Croce[2], and Fabio Massimo Zanzotto[2]

[1] *Dipartimento di Fisica, Università dell'Aquila, gc-LNGS INFN, Via Vetoio, 10-67100 L'Aquila, Italy*
[2] *Dipartimento di Informatica, Sistemi e Produzione Università di Roma "Tor ergata," Via del Politecnico 1-00133 Roma, Italy*

Determining what machines "think" can be considered a well-set problem. Computational systems have recently become so complex to motivate a parallelism with the human brain. Such a parallelism may represent a test-bed for brain-imaging as brain interpretative models can be tested on a much simpler case. We performed a virtual observation of a computational machine: the machine activity has been observed using a software program that snapshots the machine memory. Images of the memory activation states have been produced with bit resolutions. Building on these results, we are interested in what can be physically reading the machines' "thoughts" and what can be its technological implications. This is a peculiar challenging task of nanotechnology, the elementary information unit 1/0 (the bit) nowadays corresponding on chip to a physical elementary unit of nanometric dimensions. To capture activation states in physical memories we need devices that do not interfere with the chip both from a mechanical and an electro-magnetic point of view and have imaging resolutions comparable with the minimum line separation typical of the modern processors. The present work explores a new scientific field that can foster advances in neurosciences and, secondarily, in computer diagnostics.

Keywords: Computer Science, Nanotechnology, Neuroscience, Machine Learning, CPU, Cognitive Process.

RESEARCH ARTICLE

1. INTRODUCTION

When Jack Kilby realized the first integrated circuit in 1958, none imaged that these devices could become as complex as human brains. Yet, the involvement of thermodynamics in the theory of computation was an alert of a growing in complexity. Information processing became a non-negligible thermodynamic problem because of the rising amount of irreversible operations such as AND or ERASE dissipating an energy of at least $k_B T \ln 2$ for each single bit of information lost,[1] where k_B is the Boltzmann constant and T the temperature. The same Von Neumann became aware of the relevance of thermodynamics in computational processes but his speculation about the increase of entropy during computation proved to be false.[2] From the starting point, the rate of information processing that computers were capable to perform doubled every 18 months (Moore's law). The complexity of the Central Processing Units (CPUs) rose enormously.

Initial machines were still far from what we have now and what we can have in the future. Today machines are a product of nanotechnology, where the term "nanotechnology" refers to the ability of working materials with atomic or molecular precision.[3] While the most advanced microelectronics industry is nowadays trying to extend the standard lithographic capabilities beyond the 45 nm node (this value standing for the minimum line separation so far obtained), the realization of first prototype quantum computers has satisfied the ambition of nanotechnology of functionalizing single atoms or molecules.[4] New strategies like the *top-down* approach and the *bottom-up* approach exploiting the self-assembling properties of some bio-molecular materials are going sufficiently mature to promise a short-term extension of the quantum bit concept to macroscopic devices containing much more than the few elements constituting existing quantum computers.[5] Computational machines with macroscopic dimensions and atomic minimum feature size can be referred to as Avogadro-scale computers, i.e., computers acting on a number of bits comparable to the Avogadro's number ($\sim 10^{23}$). Such computational machines are complex thermodynamic systems characterized by a logical and architectural complexity very similar to that distinguishing the human brain.

*Author to whom correspondence should be addressed.

The analogy between computational machines and human brain has been always fascinating but nowadays it is becoming more evident. At the beginning, this correlation concerned exclusively theoretical models. It is not hard to imagine that the Von Neumann architecture[6] and the neural-based computing architecture originally introduced by Turing[7] had been inspired by concepts coming from studies on mind and brain. The same was in the opposite direction, from mind and brain to the computer architectures, Cognitive Psychology[8] and the more radical Cognitive Science[9] having used computing machines as a metaphor for defining models of the human mind. Nowadays, at the door of the Avogadro-scale computers era, people wonder if brain and computers will be in some future discriminated only by the kind of elementary physical units they are made of: transistors or neurons. As from now it is sometimes hard to tell where devices end and people begin, as it is in many recent progresses in neuronal medicine.[10]

The common sense, even affirming that a complete identification of computers with the human brain will be never possible, is not impeding to the incoming era of evolved computer architectures to stimulate new ways to look at the fascinating and unexplored parallelism between computers and brains/mind. Recently, in Ref. [11] we have explored a new parallelism between brain and computational machines on the field of neuroimaging. Neuroimaging techniques are used to discover areas related to particular cognitive processes and, also, to induce activation patterns for high-level cognitive processes related to specific semantic categories.[12] These activation patterns can be used to determine what cognitive process a *brain* is performing. This is an extremely fascinating area of research. If successful, the produced analyzers of brain activation images will be able *to read what humans are thinking*.

In Ref. [11] we observed that *computing machines* nowadays perform complex tasks that seem to be "cognitive processes," e.g., manipulating symbols. In a sense, computers can be considered as more "controlled" brains, where these emerging theories on the possibility of reading what brains think can be tested and verified. In this new parallelism between brain and computational machines, we can then address two questions:

(1) how far can we go with neuroimaging in understanding human mind? (*foundational perspective*);

(2) can we understand what computers "think?" (*applicative perspective*).

The *foundational perspective* of the parallelism is extremely important as we can test the foundational hypotheses of neuroimaging studies such as Refs. [12] and [13] in a simpler and more controlled setting. With the parallelism between brains and machines, we can study if finding the correlation between high-level cognitive processes and neuroimages is a feasible task. On the *brain* side (Fig. 1), we have two known variables, i.e., the required cognitive activity and the observed activation pattern, and one unknown variable, i.e., the way the brain is performing the cognitive process. In brain imaging, the aim is to understand and to model the unknown variable. On the *electronic computer* side, there are no unknown variables: the three elements are completely known. This gives a very relevant, simpler, and more controlled testbed. We know exactly how "knowledge" is processed in computers and we know exactly the "cognitive process" we ask machines to do. If we succeed in studying the correlation between the cognitive process and the activation image in the electronic computer side, we can be confident that the same method can be used on the brain side.

The *applicative perspective* is also an extremely interesting and unexplored area of research. Using the ideas developed on the *brain* side of the parallelism (Fig. 1), we can try to apply them to the *electronic computer* side. Can we develop technologies that "read the computers' mind?" This predictive model can have a wide variety of applications, e.g., detecting malicious software, detecting the intentions of hostile computers by looking at their activation patterns, or on-line testing of critical chips. We need specific devices that can capture activation images of computers. We can then study the application of machine learning to induce models that can predict what a computer is doing by analyzing its activation patterns.

This approach is innovative in the panorama of parallelisms between brain and computing machines. It involves

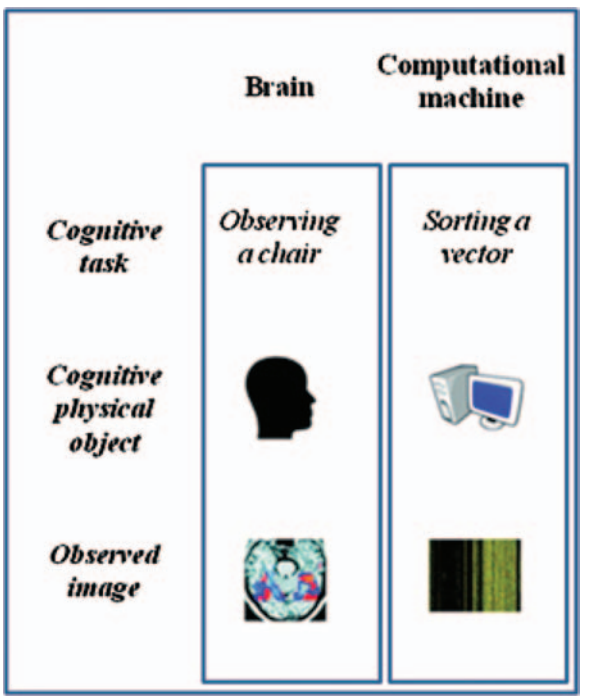

Fig. 1. Reading Brain and Machines with fMRI or fMRI-like techniques. On the brain side, mind state decoders want to guess what brains are thinking observing their activation images. In the proposed parallelism with computing machines, the aim is to build computer state decoders where cognitive activities are class of algorithms and the activation is captured looking at the memory.

at the same time structural aspects related to the nature and topography of the activation areas (in brains or CPUs) and theoretical aspects related to the way of processing the activities (process modeling). The work here presented is at the very early stage. The present communication is to highlight what concepts have been already fixed, what is still obscure, and how this approach is challenging the scientific community.

2. TERMS OF THE CHALLENGE

In the computational machine side of the parallelism, we want to investigate whether it is possible to produce a *predictor* that can determine what the computer is doing, i.e., its cognitive process, observing its activation image. We want to see if we can *read what machines think*. As in neuroimaging, this predictor can be learnt from training data using machine learning techniques. Training data are activation images as well as the performed cognitive processes. Once the predictor has been derived from training data, it can be used to associate "cognitive processes" to new activation images. In the perspective of building predictors for computational machines, we need to determine how we can produce the activation images.

In computational machines, we can assume that "cognitive processes" are represented by software processes. In a Von Neumann architecture, these processes are totally stored in the memory. If we want to pursue the brain-computer parallelism in the neuroimaging perspective, we then need to be able to observe the memory of computational machines and to understand the structure of the processes inside the memory.

Observing the memory to be detected and determining the running processes is extremely complex as we need a specific device able to snapshot the physical status of the memory and we need to understand the organization of the different processes in the memory. Yet, electronic computers have a very nice property. We can *virtually* observe the activity of these machines and the state of the memory using software programs and we can know exactly how processes in memory are organized.

The task we are prefiguring can then be organized in different steps that start from two extremes (Fig. 2). On the right side, we have the *virtual observation* of the memory. Here, we can exploit the possibility that computer memories can be observed using software programs. On the left side, we have the *physical observation* of the *physical memory chip*. In this case, we need a physical device for capturing activation images.

From the *virtual observation* starting point (see Fig. 2), we can easily simulate the *electronic computer* side of our vision without actually having a physical device to observe the activation state of machines. A software program snapshots the memory of the machine. These snapshots can then be used to produce activation images as if they were taken from an external device. The concept of "activation image" slightly differs according to what we want to observe. We have three possibilities of decreasing complexity (Fig. 2). First, we can dump the physical memory, i.e., exactly what is in the memory chip. Second, we can dump the *virtual memory*, i.e., the memory storing all the processes in a well organized and separated way. Finally, we can dump the memory of a single *process*. These three possibilities offer three different sets of images where to analyze the computational machine side of our parallelism. With an increasing difficulty, we can find the predictor that:

(1) given the *process activation images*, determines what cognitive activity the process is doing;
(2) given the *virtual memory activation images*, determines the cognitive activities performed by the machine;
(3) given the *virtual physical memory activation images*, determines all the active cognitive activities in given time points.

The last problem is the most complex and the closest to the challenges we have when *physically* observing the *virtual* chip, where we have also to treat the localization of processes in the real memory. The first problem is the most affordable and, for this reason, it has represented our starting point.

Unfortunately, the virtual approach has some limits. For example, it is practicable till the CPU is accessible. As the access to the CPU is forbidden (e.g., the operating system is protected or system errors are preventing the access to memory), although the machine is still executing activities, the only way to observe the activation states of memory is using an external physical device. Another problem affects the virtual approach, that is more serious because intrinsic to this procedure: the use of a software program snapshotting the machine memory can perturbs the state of memory, because it generates processes that are stored in the memory itself, and the final activation image may be altered.

The *physical observation* starting point (see Fig. 2) is then more rigorous and of more general applicability, but it preludes a much more complex task. A tool allowing the physical observation of a memory state still not exists. It should be a separate module external to the computational machine, it should operate directly on the chip without intermediaries and it should distinguish all the details of the given activation state. Even if based on different concepts and needing different instruments, at the end it should give the same information provided by the virtual observation.

At the moment, we are able to test the overall process of the *electronic computer* side (Fig. 1) only by *virtual observation* of the activation states (see Fig. 2). Using the information provided by this approach, we can study if it is possible to derive a correlation between the images of the activation states and the performed "cognitive processes."

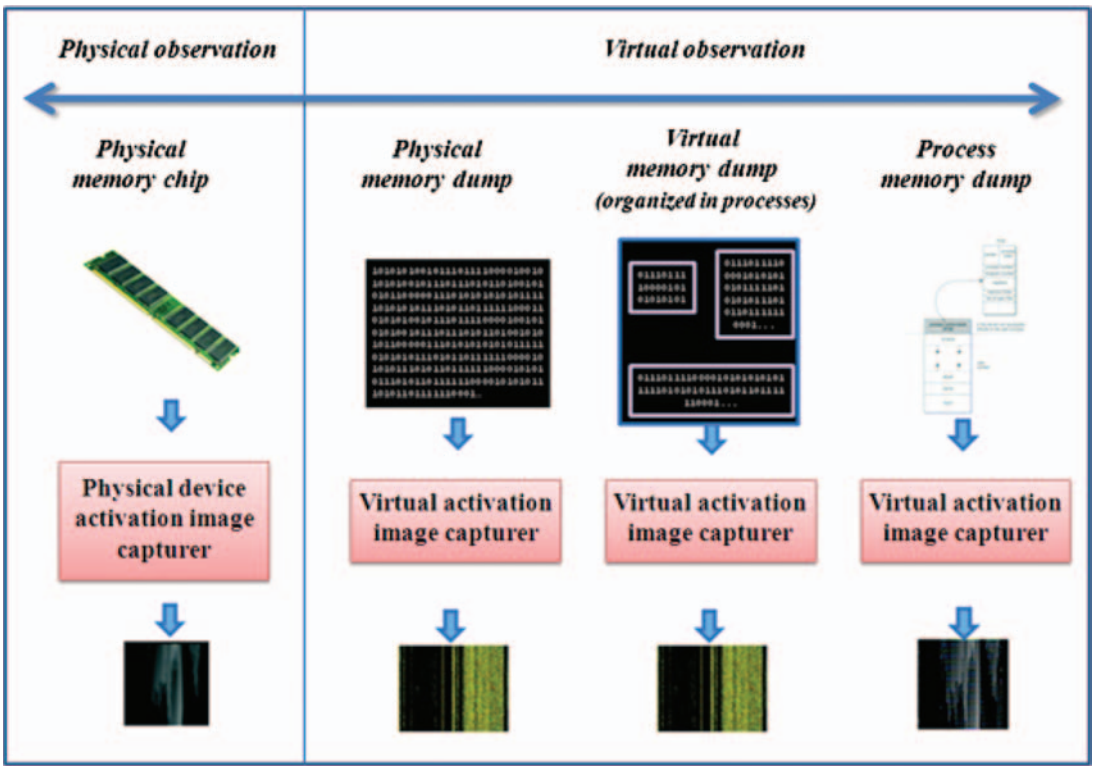

Fig. 2. The overall challenge from Virtual to Physical Observation of Computer Memory. The acquisition of activation images of computer memory can be done in different ways. From right to left in increasing order of complexity: virtual scanning of the memory of a single process, virtual images of the virtual memory (that contains all the ready processes), virtual images of the real memory (that contains only the active processes), and, finally, real scanning of the real memory chip.

For this purpose, we will extract features from activation images to feed machine learning algorithms. Given a set of training examples, i.e., training activation states, associated with different types of "cognitive activities," the machine learning algorithm can extract a prototypical model of activation for each type of cognitive activity. These models can be used to classify novel activation states, i.e., to recognize the type of cognitive process that the activation state suggests. If classifiers have good performances with respect to a set of testing activation states, we can conclude that the task of reading "machines' thoughts" is reachable using the proposed features.

3. VIRTUAL AND PHYSICAL APPROACH

The virtual approach has been successfully tested in Ref. [11], where we have produced images representing the activation state of a machine performing a particular "cognitive task," e.g., sorting a vector or comparing two strings. We exploit the fact that processes perform "cognitive activities," where a "cognitive activity" is defined as the execution of a program over input data. Processes are completely represented in memory, i.e., both programs and data are stored in memory. Snapshots of the memory associated with target processes can be directly obtained. These snapshots can be used to build images. The procedure for extracting images is reported in Ref. [11] in detail. In that paper we have proved that it is possible both to correlate activation images with "cognitive processes" and, reversely, to determine the class of the "cognitive process" from a given activation image, with an accuracy larger than 80% (in this context, the accuracy is defined as the number of correctly predicted algorithms with respect to all the decisions of the learnt classifier).

From the physical point of view we must consider that the core of the most advanced transistor-based computational machines is a complex processor made of billions of nanometer-scale elementary features. As a product of nanotechnology, these machines require the use of nanotechnological tools to be investigated at the scale of their minimum feature size. In the perspective of realizing a tool that captures images of the activation states of memory measuring physical observables, we must be sure it can replicate on-chip the physical equivalent of the byte-by-byte reading of the memory dump performed by virtual observation. The on-chip byte equivalent is a train of current pulses whose information is stored in the form of accumulated electric charge. The physical size of such information is the size of the tracks where the current flows or the size of the capacitors where the charge is accumulated. Typically, the size of these information

driving/storing features (depending if tracks or capacitors) corresponds to the minimum feature size that characterizes the processor. Given the above mentioned nanometer-scale features in modern processors, the first requirement for a tool devoted to the *physical measure* of the activation states of memory is its suitability of measuring details with nanometric resolution. The second but not less important requirement for such a tool is that it must operate without exerting both a mechanical and an electro-magnetic influence on the chip: this condition is indispensable to avoid perturbations of the activation state during the measure. In fact, physical memory is extremely sensitive to any kind of external stimulus.

The challenge is then for the nanotechnology world, given the nano-scaled resolution of a technique that may be rightly considered the equivalent of the Functional Magnetic Resonance Imaging (fMRI) for Neuroscience. In principle, Physics has all the instruments to win this challenge and many persuasive examples may be reported on this matter. The technique of Atomic Force Microscopy (AFM) is one of the most exemplary cases of detection of signals coming from nanometric features with a negligible incidence on the morphological and electro-magnetic properties of the measured samples. But AFM is not yet a suitable solution to the hard problem of acquiring information simultaneously from the overall chip surface with nanometric resolutions (the number of $10^{-9} - 10^{-8}$ m sized objects over a 10^{-4} m^2 surface is a value between 10^{-14} and 10^{-12}), AFM being a local investigation technique. Completely new investigation techniques will be maybe required to meet this challenge.

Beyond the problems of scanning the physical chip with opportune resolution and without affecting the running process, another big problem concerns the interpretation of the collected images. Images of dumped memory at the initial state (Fig. 3(a)) and when executing the process accomplishing the required task (Fig. 3(b)) have been reported in Ref. [12]. Due to the complexity of the chip architecture, physical measure of the memory chip is expected to be significantly different then virtual observation by memory dumping. In fact, data in a virtual memory can be easily and exactly localized by using software methods (as in Fig. 3), while data in a physical chip are scattered in many chunks located in different places of the chip and there are no deterministic ways to assign physical location of bits to the running logical process.

Images of Figures 3(a) and (b) have been reported in Ref. [11] also after blurring (respectively, Figs. 3(c) and (d)). The blurring effect is obtained merging contiguous pixels, each pixel in the smoothed image being a weighted sum of the square $K \times K$ of pixels around the target pixel in the original image (in our case $K = 10$). The intent of blurring images from virtual observation is to show what can be observed in principle using a physical scanning device. According to the choice of K different *blurring* degrees can be obtained, large values of K corresponding to a high blurring degree, small values of K corresponding to a low blurring degree. The lower is the blurring degree the smaller are the details of memory that the scanning device is able to resolve. In terms of K, we are associating the area $K \times K$ of contiguous pixels where the information is merged to the size of the minimum physical detail that the scanning device is able to resolve. In other words, K can be considered a measure of the detection resolution when expressed in pixel, where the single pixel is assumed to correspond to the minimum feature of the physical memory. In terms of spatial resolution, one pixel has the size of the minimum line separation characterizing the integrated circuit of the physical memory. As an example, considering that the typical minimum line separation in modern processors is about 10^{-8} m, the case $K = 10$ reported in Figure 3 may be correlated to a spatial resolution of 0.1 μm (poor to our purposes!).

But the interpretation of blurring is also critical. As it has been presented above, it cannot be considered a good approximation of the errors that a real physical scanner makes. In fact, the blurring method guarantees that pixels used to calculate blur are logically connected while in the physical chip the reading of a bit can be distorted by bits of memory not connected with the measured pixel. Because of this non linear memory storage, the distortion can come from memory bits that belongs to another process and it assumes a different meaning with respect to the graphical rendering of a virtual blurred image.

Actually, the difficulty of interpreting the virtual observation and the discrepancies with the physical observation reveal with particular evidence the affinity between the world of chips and Neurosciences. In the virtual observation bits are assigned to a well defined location and pixels representing bits are logically connected. To the aim of a parallelism with Neurosciences this ideal situation cannot give useful information. On the contrary, what is obtained from physical observation is much more similar to what is obtained by fMRI. As observed before, there are no *deterministic* ways to assign physical location of bits to a specific process, but a *non deterministic* study of the correlation between the real images of the running memory and the activated processes is the same approach of Neurosciences when correlating a stimulated area of the human brain to a specific required task. Furthermore, in our case we take also advantage from studying a simplified case, going from a tridimensional down to a less complex bidimensional system. This consideration is what rightly motivates the approach here presented for a better understanding of human mind, a purpose that involves also the possibility of extending the acquired know-how to the fascinating perspective of understanding what computers "think."

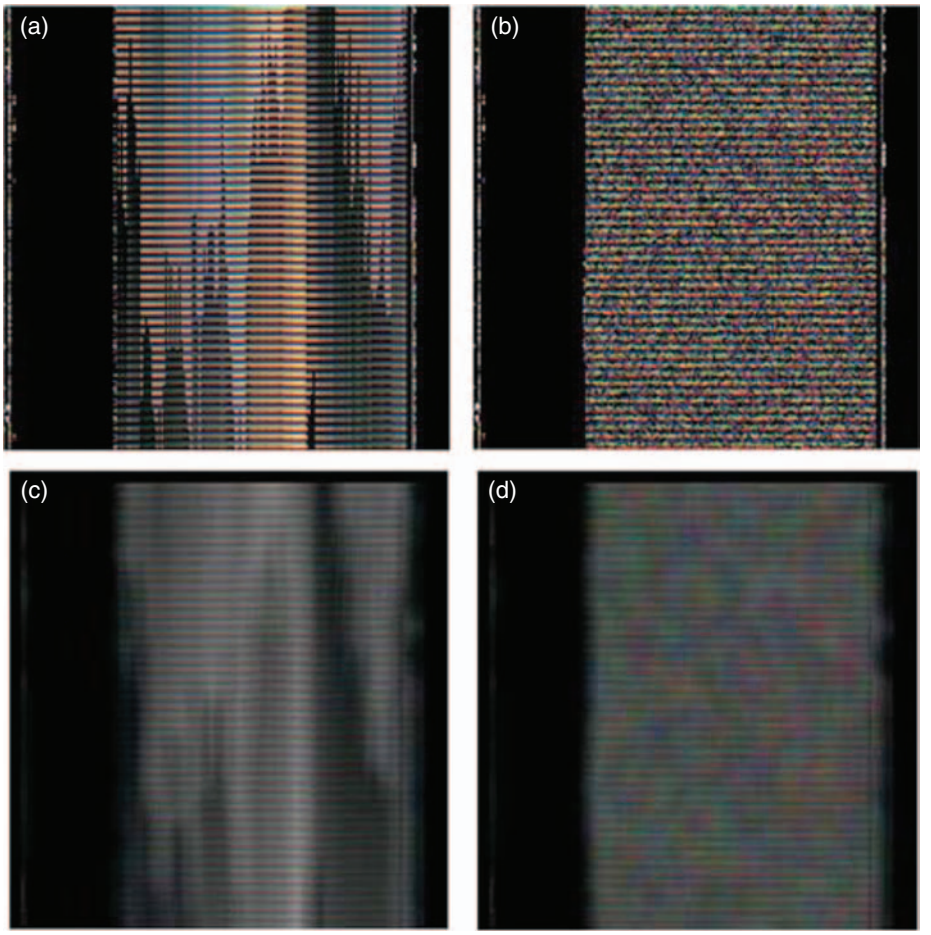

Fig. 3. Images of dumped memory at the initial state (a) and when executing the process accomplishing the required task (b); the same images after blurring (respectively (c) and (d)).

4. CONCLUSIONS

This work shows the opportunity to progress in the comprehension of the brain activity exploring the parallelism with the activity of computational machines. Despite the simplicity of this idea and the excessively large gap in complexity between the two terms of comparison, the possibility to get useful information on the behavior of the human mind from the study of inanimate machines is a fascinating problem. Using firstly a virtual approach, we have provided concrete elements to believe that this parallelism is a practicable investigation method, with fascinating perspectives if it will switch to a physical approach. The next steps are really ambitious but we are fully confident in the scientific and technological background of nanotechnology community, that can easily accept the challenge of realizing a tool that will read what machines "think."

References

1. R. Landauer, IBM *J. Res. Dev.* 5, 183 **(1961)**.
2. J. Neumann, Theory of Self-Reproducing Automata, University of Illinois Press **(1966)**.
3. N. Taniguchi, *Proceeding International Conference Prod.* London, British Society of Precision Engineering **(1974)**, Part II.
4. N. A. Gershenfeld and I. L. Chuang, *Science* 275, 350 **(1997)**.
5. R. Cingolani, R. Rinaldi, G. Maruccio, and A. Biasco, *Physica E* 13, 1229 **(2002)**.
6. J. Neumann, *IEEE Ann. Hist. Comput.* 15, 27 **(1993)**.
7. A. Turing, Machine Intelligence, 5th edn., edited by B. Meltzer and D. Michie, Edinburgh University Press **(1969)**, p. 3.
8. U. Neisser, Cognitive Psychology, Appleton-Century-Crofts **(1967)**.
9. B. Eckardt, What is Cognitive Science? MIT Press **(1993)**.
10. J. Clausen, *Nature* 457, 1080 **(2009)**.
11. F. Zanzotto and D. Croce, *Proc. BI* 2009, Springer-Verlag **(2009)**, p. 159.
12. A. Ishai, L. Ungerleider, A. Martin, J. Schouten, and J. Haxby, *Proc. Natl. Acad. Sci. USA* 96, 9379 **(1999)**.
13. T. Mitchell, S. Shinkareva, A. Carlson, Kai-Min Chang, V. Malave, R. Mason, and M. Just, *Science* 320, 1191 **(2008)**.

Received: 5 December 2010. Accepted: 12 December 2010.