COMPUTING CONSCIOUSNESS

Bert Timmermans, consciousness, cognition & computation group Université Libre de Bruxelles
(bert.timmermans@ulb.ac.be)
Axel Cleeremans, consciousness, cognition & computation group Université Libre de Bruxelles
(axcleer@ulb.ac.be)

What is it like to be human? That, of course, is in a sense the question that Turing asked 55 years ago, offering the Turing Test as a way of addressing the question without really answering it. Turing, indeed, suggested that we should attribute human-like intelligence to any machine that exhibits behaviour indistinguishable from that of a human being. But what about the question of human experience? What about a human who consciously feels? Here, we introduce ways in which we might start thinking about building such an artificially conscious agent. We describe the basic mechanisms that such an agent should possess, based on what is currently known about the neural underpinnings of consciousness.

Inject Mind Here: From Intelligence to Consciousness.
Looking at the evolution of artificial creatures in film, one cannot but acknowledge an evolution from mindless to mindful malevolence. Whereas Golem, Frankenstein's creature, the body snatchers and Romero's undead seemingly lack any real mind, making them into monsters, virtual legends such as *2001*'s HAL or *Demon Seed*'s Proteus are actually scary *because* of their mind. Without lingering on the philosophical debates on whether a certain type of mind can exist independent of its specific embodiment or whether any creature can understand a consciousness that is not like his own (recall Lem's *Solaris*), the thing that makes HAL and Proteus so human is not so much their ability to think as their possessing something resembling human consciousness. The point is that, whereas consciousness may or may not be required for an artificial agent to think, it is an essential element in creating anything resembling a human-like thinker that would pass the Turing Test.

The difference between a conscious and a non-conscious agent may seem futile or trivial, but it is neither. For a non-conscious agent, seeing the colour "red" is mere pattern-matching and recognition, as dictated by an algorithm. However, a conscious agent doesn't simply process and identify "red"; instead there is the *subjective experience* of "seeing red". Philosopher Thomas Nagel put the finger on this problem of subjective experience in a seminal paper entitled "what is it like to be a bat?" For conscious agents, there is something it is *like* to experience something. Therefore, the crucial question when we look at neural or computational mechanisms for consciousness, is *why* there exists such subjective experience for us (this has been coined by David Chalmers as the "hard problem" of consciousness). What does it do (if it does anything at all)? Is it simply a consequence of having a sufficiently complex system? Or is consciousness possible in any system? This problem of "explaining" consciousness can be approached in very different ways that we cannot cover here, but we will survey some of these approaches as they are being developed in the emerging field of "machine consciousness".

Mapping The Unknown: Consciousness Phenomenology and Neurology. A formidable challenge lies at the very core of the endeavour to build conscious machines, for consciousness, by its very nature, is a completely private phenomenon and therefore in principle inaccessible to objective measurement. As the philosopher David Chalmers famously pointed out, we do not have a "consciousness-meter" that we can point to people's brain to figure out how conscious they are or what they are currently aware of. Thus, empirical research on consciousness, which is now booming, has essentially sought to develop an approach whereby objective and subjective measures of cognition can be obtained at the same time. For instance, one may ask people to make decisions in a cognitive task (say, identifying a barely visible stimulus) and to simultaneously report on their experience of the stimuli (e.g. "I saw the stimulus" vs. "I am just guessing"). By correlating objective performance and subjective reports, one can assess the extent to which processing involves or requires awareness. One can further look for brain regions that activate differently under different conditions of awareness. This approach has been called the "quest for the Neural Correlates of Consciousness" (NCC). While many different brain areas have been suggested as plausible candidates for the NCC, it is now clear that consciousness depends not merely on activity in one region, but rather on complex interactions that engage the entire cortex. Characterizing the mechanisms that underlie these interactions is the focus of what one could call a "quest for the Computational Correlates of Consciousness" (CCC), that is, computational principles that differentiate between information processing with and without consciousness.

Computational Correlates of Consciousness
What conditions must a mental representation satisfy in order for it to reach consciousness? What are the computational consequences of a representation reaching consciousness? Do conscious and unconscious states influence processing differently? What is the computational utility

of consciousness? As any question in AI, such questions can be approached in different ways: In terms of abstract properties of computations (e.g., holistic, analytical, controlled, self-organising processing); or in a manner that takes direct inspiration from the way in which processing occurs in the brain (e.g., feed-forward versus recurrent processing in neural networks). Most existing theories of consciousness that are rooted in computational considerations can be classified along two dimensions: A *processing vs. representational* dimension, and a *specialised vs. non-specialised* dimension. The first dimension opposes theories that emphasize the involvement of specific processes (e.g., global synchrony) in generating conscious experience to theories that emphasize specific aspects of representations (e.g., their stability in time). The second dimension opposes theories that assume that consciousness depends on the involvement of specific networks and structures in the brain (e.g., the frontal cortex) to theories that assume that conscious representations may occur anywhere in the brain. In the following section, we examine several existing proposals for such CCC, based on what is currently empirically known on NCC, a necessary constraint to anyone who wants to model human consciousness.

The Seven Samurai: Essential Characteristics of CCC. While many existing proposals concerning the CCC are incompatible with each other, most share a number of basic assumptions that roughly fall into two broad categories. First are "fame in the brain" proposals, initially pioneered by the philosopher Daniel Dennett, which assume that consciousness occurs whenever some conditions are fulfilled, such as *Stability and Strength* of representation, which can be viewed as resulting from *Re-entrant processing* and/or from *Synchrony of processing*. Essentially, these proposals assume that the brain is a large dynamical system in which stable, attractor states come in and out of existence as a result of continuously operating global constraint satisfaction processes. The main functional consequence of such states is that the information they convey then becomes *globally available* for further information processing, which in turn enables the brain to form interpretations of the world that are both *integrated and differentiated* (see below for an explanation of these concepts). The other main proposal is the notion that consciousness depends on "higher-order thoughts", that is, on the existence of *meta-representations* that enrich first-order representations.

Basic features

Our brain is a hodgepodge of constant neural firing, where information is passed around countless of times in fractions of a second, but unlike Star Trek's Commander Data, we do not have total access to this content. *Stability* and *strength* of representation are likely to form two basic requirements for neural representations to be available to consciousness. In the case of stability, content becomes conscious as soon as its activation persists over some period of time. Representations acquire stability as a result of relaxation processes as they occur in dynamical systems. An interactive network, for instance, will tend to "settle" in one of a limited number of stable, "attractor" states. In the case of strength, content becomes conscious when its activation passes a certain activation threshold. Strength, in this context, could refer to the number of neurons involved in the representation relative the to the number of neurons involved in competing representations, or to the fact that a self-sustaining coalition of neurons has formed and inhibits other competing coalitions. Both concepts are intertwined, as one can imagine that a representation only acquires strength over time, and so stability would be a prerequisite to gain sufficient strength.

Processes, or the how of it

How do representations become strong and stable in time? Two kinds of processes have been proposed. The first, *re-entrant* or *recurrent* or *feedback* processing, is the process by which, rather than just feeding forward through a set of neurons, activation is passed back to the sending neurons. Neural networks in the brain are massively recurrent, with "downstream" neurons connecting back to the "upstream" neurons from which they receive connections. Recurrent networks have very different computational properties from purely feed-forward networks. In particular, recurrent networks have internal dynamics and can thus settle onto particular attractor states independently of the input, whereas feed-forward networks only become active when their preferred inputs are present. It has been suggested that exactly this difference accounts for consciousness, in that feed-forward sweeps would be unconscious, while recurrent activation would produce conscious content. It is clear how this relates to stability and strength, as in neural networks both are typically acquired through recurrent processing. A potential problem with this view is that we can have recurrence at any level which does not necessarily involve consciousness. Furthermore, it does not explain the so-called "binding problem", which refers to the fact all this information somehow needs to be integrated, since we experience our conscious content as an undivided whole. A second view on processing that attempts to overcome this problem, claims that *synchrony* or *gamma oscillations* of neural firing is a necessary prerequisite for consciousness. It has been observed that when content becomes conscious, this is accompanied by different parts of the brain working in temporal synchrony through high frequency oscillations (neural firings). Apart from solving the binding problem, neural synchrony has the computational consequence that it strengthens and selects specific signals on a more global level than local recurrence, and could explain why at some point content becomes globally available.

Consequences

When content becomes *globally available*, it means that the brain can work with it, that is, it becomes available to all sorts of brain processes (for instance, cognitive control), as through a *global workspace*, which would consist of long-range connections in the cortex. It's important to note that this point of view, and in fact any view endorsing some sort of selection threshold, implicitly assumes that consciousness is an all-or-none phenomenon, because as soon as you have either recurrence, synchrony, or global availability, consciousness is "ignited". Empirical studies, however, have so far remained inconclusive about whether consciousness is dichotomous or gradual. Even when it is gradual, it can be gradual in that it consists of graded representational strength, but also that this impression of a gradual consciousness it is a mere illusion brought about by increasing "on"-switching of "pixels" of information, which, when a stimulus is sufficiently complex (such as the real world), can create the impression of graded consciousness. The main issue here is that from a bottom-up point of view, content at some point, through any or all of the mechanisms described higher and depicted in Figure 1, becomes available to different processes, which is associated in some way with that content becoming conscious.

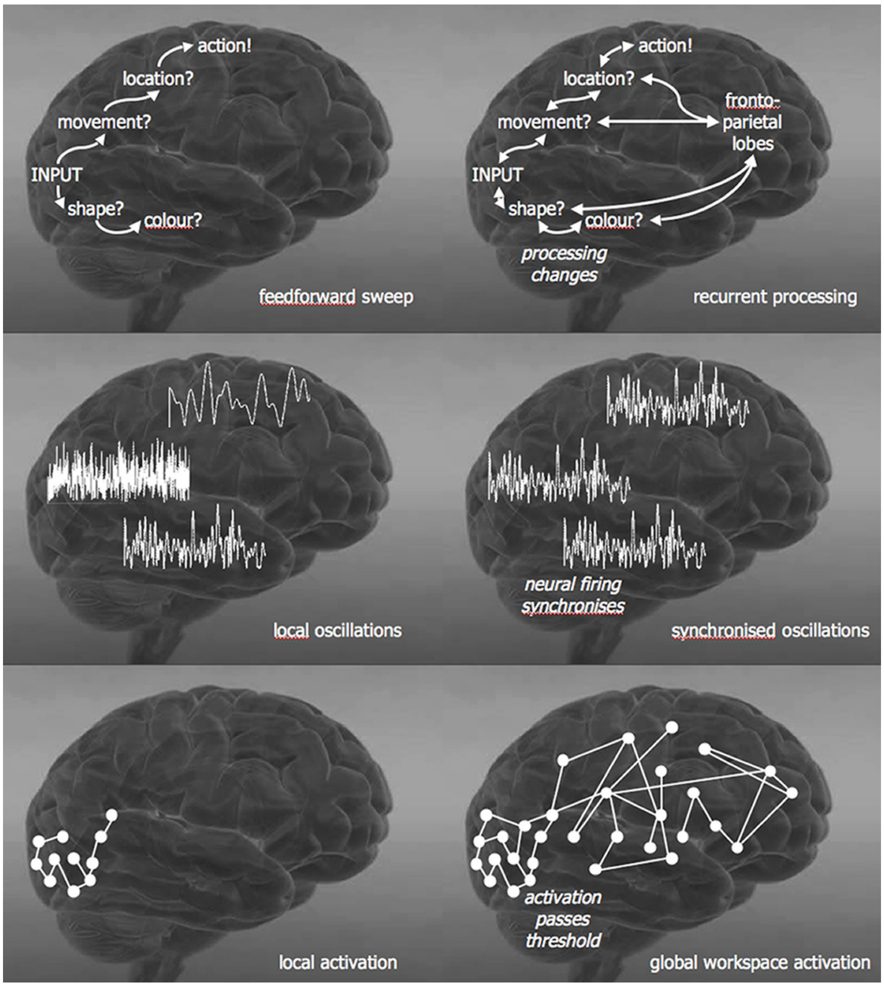

Figure 1. Three examples of "Fame in the Brain" theories on consciousness, depicting the supposed equivalents of visual perception without (left) and with consciousness (right). The top panel represents theories assuming recurrent or re-entrant processing as the crucial feature allowing for consciousness; the central panel represents theories that see synchronised oscillations as a prerequisite for consciousness; while the bottom panel depicts theories that assume global workspace activation is what makes content conscious. Note that, while dissimilar, all theories share the common feature of a clear-cut qualitative difference between unconscious and conscious processing.

From a more theoretical top-down point of view, *Integrated Information Theory* has proposed that conscious states are characterized by the fact that they are both highly *integrated* and highly *differentiated* states. Integration refers to the fact that conscious states are states in which contents are fundamentally linked to each other and hence unified: One cannot perceive shape independently from color, for instance. Differentiation refers to the fact that conscious states are one among many possible states; for each conscious state, there is almost an infinity of alternative possibilities that are ruled out. Thus, only systems capable of both integrating and of representing a wide array of distinct states are capable of consciousness. Based on this hypothesis, one can thus analyze, from a computational point of view, what kinds of systems are capable of such integrated and differentiated representation.

Meta-representations

Finally, a rather different idea has been proposed by the philosopher David Rosenthal in the form of his "Higher-Order Thought" theory of consciousness, which assumes that at any given time, a representation is conscious to the extent that one has a thought that one is conscious of that representation. In other words, a representation of some content becomes conscious when it becomes possible to think *about* that content. Mere "fame in the brain" is therefore neither sufficient nor necessary to make a representation conscious in this perspective; what is needed instead is the occurrence of meta-representations that redescribe in specific ways lower-level representations. Thus, it is in virtue of this re-representation that first-order content becomes conscious content.

Modelling Consciousness in Neural nets

Computational models of consciousness are few and in between, and for good reason: Since phenomenal consciousness ("what it is like") cannot be characterized in functional terms, it is difficult to propose plausible mechanisms for it. Thus, most existing models have focused on accounting for the functional differences between computations carried out with or without consciousness. For instance, at the Université Libre de Bruxelles (ULB), we have developed connectionist models that are able to capture human performance in implicit learning tasks. Implicit learning occurs when people exhibit behavioural adaptation to a stimulus environment without awareness of what they have learned. Natural language learning is often considered to involve such implicit learning, for we all learn to express and correct utterances without intending to do so, and without necessarily acquiring verbalisable knowledge of grammar. In the laboratory, implicit learning is explored by means of experimental situations in which people are exposed to an ensemble of complex stimuli and asked to process them in some way. For instance, people may be asked to memorize numerous meaningless strings of letters (e.g., "TXVPPS", "PTVXS", &c.). Unknown to them, all these stimuli have been generated based on a finite-state grammar. After memorization, people are then told that the strings had all been generated based on a set of rules, and are now asked to decide whether novel strings are "grammatical" or not. Participants complain that they know nothing of a grammar, but they are told to try their best based on intuition. The main result in these experiments is that people consistently perform better than chance when classifying these novel strings are grammatical or not, yet they find themselves unable to verbalize how they make their decisions, thus exhibiting a dissociation between their performance and their awareness of the knowledge they have acquired—implicit learning.

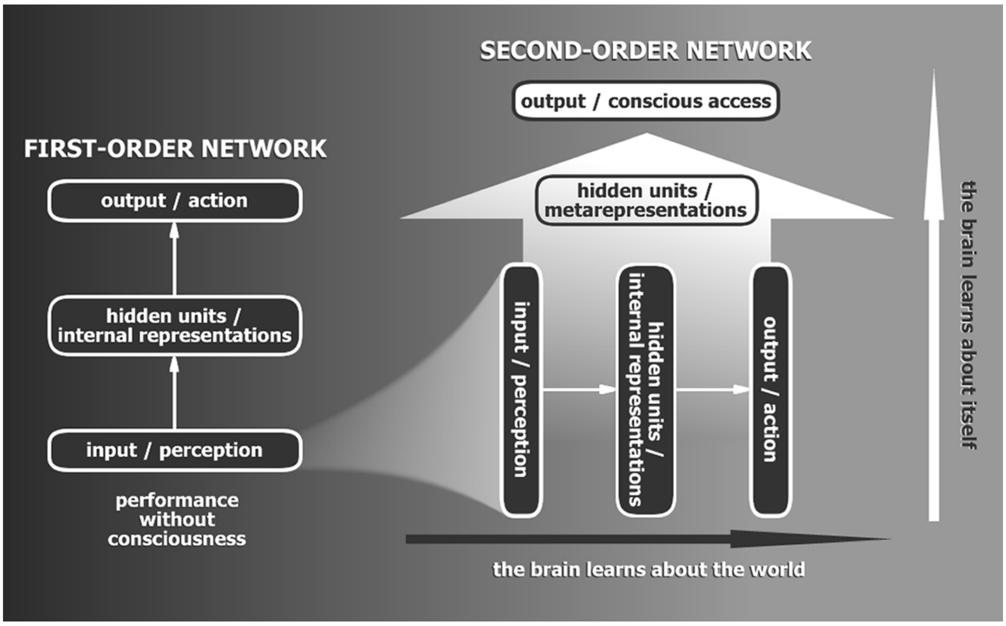
SECOND-ORDER NETWORK
output / conscious access
FIRST-ORDER NETWORK
hidden units / metarepresentations
output / action
the brain learns about itself
hidden units / internal representations
input / perception
hidden units / internal representations
output / action
input / perception
performance without consciousness
the brain learns about the world

Figure 2. Illustration of metarepresentations in a connectionist network. On the left is a first-order network that can use its internal representations to solve a task, but has no access to these representations. On the right, this first order network becomes the input of a higher-order network, which can access the knowledge of the first network, and which can therefore in theory become conscious as the brain learns about itself, for instance what this first-order knowledge means for the system.

In our lab, we have mostly explored how well Elman's *Simple Recurrent Network* (SRN) is able to capture human performance in such and similar situations. This type of back- propagation network consists of a layer of *input units*, representing perception, connected to a layer of *hidden units*, which in turn feed into a layer of *output units*, representing action. The SRN has an additional set of *context units*, which are a copy of the hidden units at time that form part of the input at time t+1 (hence *recurrent*). Using this network, we have been able to reproduce various empirical results in the implicit learning literature, showing for instance that when trained on artificial grammar learning material just as human participants were, the network develops a rich and abstract set of internal representations that actually represent the structure of the grammar used to generate the strings. However, and this is a crucial point, the network can use this knowledge, but it does not know that is possesses it. The sophisticated representations that the network has developed over training constitute knowledge that is *in* the network, but not knowledge *for* the network (Figure 2, on the left). Just like human participants, the network is not aware that it has learned anything, nor does it know the structure of the knowledge that it has acquired. Hence the question that motivates our current work: What mechanisms are required to make it possible for the network to "know that it knows"?

Plug a network into another

One simple possibility that we have recently explored is to train a second network to observe the internal states of the first (Figure 2, on the right). In this way, knowledge in the first-order network becomes knowledge *for* the second-order network. This second network can then be trained to perform different tasks, such as predicting the first network's state, or evaluating the extent to which the first network has learned its task well. In a recent study, we have shown that such second-order networks can be trained to wager (place bets) on the first-order network's performance in a cognitive task. Interestingly, wagering can be taken as a measure of awareness, as follows: If the decisions you take are conscious, informed decisions, you will also know when you are correct and when you are not. For instance, to follow up on the artificial grammar learning example described above, if you have conscious knowledge about the rules of the grammar, then you will know when your classification decisions are correct and when they are not. Thus, when asked to bet on your own decisions, you will place correct, advantageous bets. If, on the other hand, you have no idea when you are correct and when you are wrong (that is, when your knowledge is unconscious), your wagering will be at chance even while your classification performance may be better than chance. Our modelling work captures these patterns of associations and dissociations very nicely, and interestingly suggests that consciousness may be something we learn rather than something we have. This leads to the hypothesis that the brain is continuously learning about its own information processing, thus developing models of its own workings, and that this self re-description is a crucial computational principle that differentiates conscious from unconscious processing. Proteus is not around the corner, but a step at a time may get us there. ø

Further reading:

- *Several questions regarding thought in computers, including on consciousness and thought thought, are addressed in detail on Mapping Great Debates: Can Computers Think? http://www.macrovu.com/CCTGeneralInfo.html*
- *For important papers on any issues regarding consciousness, visit David Chalmers' monumental Online Papers on Consciousness http://consc.net/online*
- *Two recent papers regarding modelling of consciousness: Maia, T. V. & Cleeremans, A. Consciousness: converging insights from connectionist modelling and neuroscience. Trends in Cognitive Sciences, 2005, vol. 9, pp. 397-404; and: Cleeremans, A., Timmermans, B., & Pasquali, A. Consciousness and metarepresentation: a computational sketch. Neural Networks, 2007, vol. 20, pp. 1032-9.*
- *Watch out for Wilken, P., Bayne, T., & Cleeremans, A. (Eds.). The Oxford Companion to Consciousness. Oxford University Press. Currently in press, expected early 2009.*