Vol 5 | Issue 1 | Pages 87-94

Trends in Artificial Intelligence
ISSN: 2643-6000

Review Article

DOI: 10.36959/643/307

The Self-Organizing Consciousness: Implications for Deep Learning

Pierre Perruchet[*] *and Annie Vinter*

Université Bourgogne Franche-Comté, Esplanade Erasme, France

Abstract

The success of deep neural networks is impressive, but the behavior of machines is far from mirroring the behavior of humans. Among the most widely acknowledged differences are (1) The time course of learning and the need for big data; (2) The task specificity (artificial networks are unable to build some general competence which could be at least partially transferred to tasks presenting certain analogies for humans) and (3) The type of errors (the errors of deep neural networks are not "human-like"). We suggest that taking the formation of human intelligence as expressed in conscious thinking as a model (instead of the neuronal micro-structure of the brain) could contribute to overcoming these limitations. Conscious states and operations are characterized by a set of constraints, such as selectivity (one can focus on only one set of information at a time) and fast memory decay (most momentary contents of consciousness are doomed to forgetting). These constraints are observed in humans, but they can be directly implemented in computational models. We show that intelligence in humans can emerge, not despite these constraints, but thanks to them, by relying on the amazing power of self-organized systems, as observed in social insects. Implementing self-organization does not seem possible through a purely incremental improvement of current AI architectures. As a consequence, we assert that AI researchers who may wish to introduce algorithms closer to the principles underpinning the shaping of human consciousness have no other alternative than exploring a radically new avenue of investigation.

Keywords

Deep learning, Neural networks, Intelligence, Consciousness, Self-organization

Introduction

The success of neural networks, especially the so-called deep neural networks, that is, connectionist networks comprising several hidden layers (even though the most successful programs of deep learning call also for other algorithms), is impressive. As well-publicized in the media, deep learning networks defeat the best human players in Chess, Go and other complex games, diagnose complex diseases better than the best physicians; self-driving cars are in progress, and language translation services improve at a surprising rate.

However, the behavior of machines is far from mirroring the behavior of humans. Of course, humans also learn to play chess and Go, to diagnose complex disease, to drive cars, and so on, but there are striking discrepancies in the way they do so. Among the most widely acknowledged are:

1) The time course of learning and the need for big data. Humans are typically able to learn a lot from a few, or even a single event, whereas the same event would affect marginally the weight of an artificial network. Artificial networks need very extensive training with an amount of data that far exceed what humans may expect to collect during their whole lifetime.

2) The task specificity. Humans typically acquire competencies that may apply to a large array of different situations due to transfer and generalization. Now, even after extensive training, artificial intelligence exhibits a strong task specificity. Artificial networks are unable to build some general competence which could be at least partially transferred to tasks presenting (for humans) certain analogies. A program for playing Go has no advantage to learn playing chess, not to mention driving or language translation, so that it is now usual to speak about *an* artificial intelligence, a term that would be incongruous for human intelligence.

3) The type of errors. Humans make errors, and certainly more than AIs on average, let alone because they

***Corresponding author:** Pierre Perruchet, LEAD, CNRS 5022, Université Bourgogne Franche-Comté, Esplanade Erasme, Pôle 2AFE 21000 Dijon, France

Accepted: May 19, 2021

Published online: May 21, 2021

Citation: Perruchet P, Vinter A (2021) The Self-Organizing Consciousness: Implications for Deep Learning. Trends Artif Intell 5(1):87-94

SCHOLARS.DIRECT

are prone to stress and fatigue. But the issue bears on the types of errors. The errors of deep neural networks are not "human-like". They can occur in situations that are overtly simple for humans, and they look strange for anyone (including the network designers). The so-called adversarial examples [1] are especially troubling. Misunderstanding the origin of errors casts doubt on the possibility of correcting them, and beyond, on the reliability of the whole system. Reliability is obviously crucial for sensible applications, such as car driving. But, for our concern, errors are important because they provide an invaluable window upon the functioning of the system. Errors are heavily exploited in pedagogical settings, where they may reveal, for instance, that a given learner has completely missed the point the teacher has just explained. This is exactly the feeling - a total lack of understanding - left by an examination of the errors committed by deep learning systems.

This Paper

This paper explores how a drastic change in perspective could contribute to overcoming the limitations of deep neural networks listed above. In a nutshell, our proposal consists in taking human intelligence as expressed in conscious thinking as a model, instead of the neuronal micro-structure of the brain. One could object that our suggestion amounts to a return to symbolic architectures, such as initially proposed by Newell and Simon. This is not the case: In most symbolic models, knowledge coming from human experts is directly implemented into the system as ready-to-use rules or concepts. By contrast, we intend to exploit the way conscious thought and representations are shaped by experience through automatic processes in humans, in the hope of transposing a similar learning mode to the machine. The organization of the paper follows: we first examine how conscious thought emerges in humans, before proposing suggestions to implement similar principles in the artificial intelligence domain.

We are aware that focusing on consciousness may seem pointless for AI workers, given that a network is neither conscious nor unconscious. In this regard, it is important to mention from the outset that in the view proposed here, being conscious for a mental state is not defined by the presence of a mysterious subjective "color". To be potentially conscious, mental states and operations must satisfy a certain number of constraints such as seriality, limited capacity (one can focus on just one set of information at a time), and quick memory decay (most of the momentary content of consciousness is doomed to forgetting). These constraints are observed in humans, but they may be directly implemented in a computational model. We acknowledge that whether a given mental state or operation fulfills these constraints may be sometimes debatable, but on the other hand, starting from clear-cut definitions is not always possible or desirable. When the term *artificial intelligence* (AI) emerged in the midst of 1950s, its main component, *intelligence*, was far to receive a consensual definition. This indeterminacy could have fueled endless debates about whether intelligence is binary or continuous, unidimensional or multidimensional, and so on. Instead, Melanie Mitchell [2] notes: "*For better or worse, the field of AI has largely ignored these various distinctions*". Interestingly, she adds that in a 2016 report on the current state of AI, a committee of prominent researchers pointed out that "*the lack of a precise, universally accepted definition of AI probably has helped the field to grow, blossom, and advance at an ever-accelerating pace*". (p.20). As for intelligence, anyone has some rough understanding about what it means for a mental state to be potentially conscious, and this understanding may be conceived as sufficiently grounded to go ahead. We surmise that those who now argue that this kind of issue is too elusive to deserve consideration would have, sixty years ago, been left wondering what *intelligence* means exactly, without taking a step forward.

Starting from Perception

The first and principal manifestation of human intelligence is certainly that broadly speaking, everyone consciously perceives the world as it is. "*Perception is, by definition, a meaningful awareness of one's environment and one's perspective on it*." wrote Clore and Proffitt [3]. Regarding visual perception, the first principle Chater states is: "*We 'see' only meaningful organizations (or at least the most meaningful organization the brain can find): Visual chunks, patterns and whole letters, numbers, words, rather than a random scatter of fragments.*" [4]. This kind of claims is not a speculative and optional proposal, it derives from the most fundamental principle of evolutionary biology: As pointed out by Velmans, "*if the experienced world did not correspond reasonably well to the actual one, our survival would be threatened*" [5].

Given that the external world actually comprises numbers and words, one could assume that this principle all simply attests to the fact that mind's structure mirrors world's structure. In other words, percepts would be meaningful because they would be a point-by-point replica of the world. But pushing the observation a little further reveals that this account does not hold, due to two complementary sets of phenomena. First, we may not perceive certain events that are actually in the visual field. Nowhere is this more obvious than in the well-documented phenomenon of inattentional blindness [6], whereby people fail to notice obvious and salient stimuli that they look directly at while their attention is engaged elsewhere (for example towards objects of another category) [7]. Conversely, we may perceive meaningful events even when they are not in the sensory input. We assign immediately a meaning to everything we perceive, even when the sensory data are too sparse to decide. Experimental investigations with bi-stable pictures, in which the sensory input is specially designed to be ambiguous, illustrate this effect. Everyone sees, say, either the young girl or the mother-in-law in the famous image from William Ely Hill, with alternation over time, and never a chimeric and meaningless combination of both. Everything happens as if the brain was unable to create conscious representations that would make no sense for us: At any given time step, nothing in the conscious content is indeterminate or probabilistic in nature.

All these phenomena, and many others such as imagery and hallucinations [8], have led most researchers in the perception area to acknowledge the existence of powerful top-down influences on perceptual phenomena. This impact of cognition on perception finds a compelling support in neuroanatomy. Indeed, sensory cortical areas, including those ensuring the earliest stages of processing, receive much more descending than ascending projections. As noted by Rolfs and Dambacher, [9] "in fact, anatomy tells us that the only substrates of visual processing that are not targeted by top-down feedback are in the retina".

The question is now: How can we conceive of the formation of veridical conscious percepts such as exhibited in daily experiences? A surprisingly simple and powerful solution is proposed in the next section.

Introducing Self-Organization

The solution we propose relies on the existence of dynamical interactions between bottom-up and top-down processes, whereby conscious cognitive contents shape momentary conscious percepts, and reciprocally, conscious percepts progressively built conscious cognitive contents [10]. So doing, we complete the top-down effects evoked above, through the introduction of self-organizing processes to leave the place to an integrative perspective in which the boundaries between perception and cognition nearly disappear.

The concept of self-organization was introduced in the domains of physics and chemistry to account for the emergence of macroscopic structures out of processes and interactions defined at the microscopic level [11]. This concept has been exploited in many domains in the last decades, including biology, cybernetics and of special relevance for our concern, in ethology, to account for the surprising accomplishments of social insects such as ants and termites [12]. It has also been used sporadically in developmental psychology [13,14].

The concept of self-organization does not easily fit into a formal definition. We retain below two main, related features. First, the process does not need support, control, correction, or supervision by any external agent, a list including a putative mental unconscious. Second, order arises from direct dynamical interactions between multiple systems, and between these systems and their environment. "Dynamical" stands here for "across time", and in a psychological context, the time scale may be the few minutes or hours of a laboratory experiment, the months or years of practice needed to acquire a given expertise in adults, or still the whole lifetime from infancy to old ages. But what is meant is not, or not only, the fact that the same interactions occur continuously between two (or more) static structures over the considered period. What happens in general are reciprocal changes of configuration. To simplify, considering two systems A and B, the action of A transforms B1 into B2, and B2 in turn transforms A1 into A2. The next processing step will concern A2 and B2, which may differ from A1 and B1 either quantitatively or qualitatively, and so on recursively.

This integrative perspective is fully consistent with current neurosciences. Miskovic, et al. [15] wrote: "*we believe that analogies to dynamically reverberating loops and principles of reciprocal causation provide a much closer approximation to the ways that brains function, and that these ideas necessitate a more thoroughgoing reevaluation of many cognitivist axioms. It is quite possible - indeed it seems likely - that static distinctions between perception, cognition, and emotion reflect much more about historical intellectual biases in the field of cognitive science than about the true operations of the brain/mind*". Along the same lines, in a commentary co-signed by 16 authors, Hackel, et al. concluded: "*The distinctive experiences of seeing and thinking do not reveal a natural boundary in brain structure or function. The idea that the brain contains separate "mental organs" stems from an ancient view of neuroanatomy* [16]. *Modern neuroanatomy reveals that the brain is better understood as one large, inter-connected network of neurons, bathed in a chemical system, that can be parsed as a set of broadly distributed, dynamically changing, interacting systems* ".

Let us consider how a system exploiting dynamical interactions between mental contents can explain the formation of veridical percepts, using as a support an experimental paradigm largely exploited after the seminal paper of Saffran, et al. [17]. Participants hear a continuous flow of syllabic speech (e.g.bupadapatubi), which is composed, unbeknown to them, by the repetition of a few (artificial) words (in our example: bupada and patubi). Even though there is no pause or any other prosodic markers of word boundaries, subjects, whether infants, children or adults, discover the words, often after only a few minutes of exposure to this continuous artificial language. This situation was originally investigated in the context of studies on language acquisition, but the paradigm is considered here as an instantiation of a more general issue, namely discovering the genuine world units whenever they are not salient in the sensory input.

Our interpretation, which has been implemented in a computer program PARSER [18], is that subjects first perceive sequences of a few consecutive syllables (using syllables rather than, say, phonemes, as the initial shaping units are inconsequential). Each perceived sequence, construed as the content of a single attentional focus, is the stuff of mandatory mechanisms of associative learning and memory, so that its components are chunked together to compose a new multisyllabic unit. In most cases, these new units do not match the actual words of the artificial language. In the excerpt above, the speech flow may have been segmented, for instance, as "bupa" and "dapatubi". The key point of the proposed account is its top-down component: These chunks, whether correct or not, become the new units of processing, in place of the initial units (the syllables in this case). As such, they shape subjects' further perception, so that subjects now perceive a random sequence of a few newly formed units whenever they are present in the input. These new units may be in turn chunked together to compose new units, and so on recursively.

One could conjecture that by proceeding as described above, the system quickly becomes encumbered with a plethora of incorrect or increasingly longer candidate units. This does not happen, however, because in PARSER, processing units are doomed to forgetting, an ubiquitous phenomenon which again fits well with conscious experiences. The words emerge naturally (i.e., without any supervision) from the far more numerous irrelevant units that are created on the fly, all simply because, in the speech flow, words (such as "bupada") are more frequent and more cohesive than chunks straddling word boundaries (such as "padapa "). Likewise, although no arbitrary limit is posited for the length of units, excessively long units, when created, do not survive (as most sentences heard in everyday life are forgotten in their verbatim form), because long sequences are not frequent and cohesive enough. The cohesiveness of an unit may be measured as the contingency between its components, which depends on the bidirectional conditional probabilities between them. PARSER turns out to be sensitive to contingency without calling for any mental computation, because, as known for half of a century, associative processes, which are at the core of the model, are sensitive to contingency [19]. For instance, "dapa", which straddles word boundaries in the example above, will be eliminated because "da" and "pa" reoccur often in other contexts, hence interfering with "dapa". To sum up, the relevant word units emerge from the recurrent interplay between bottom-up and top-down influences, with elementary associative learning and memory processes being applied to continuously evolving mental units (and, crucially, not only to some "objective" primitives of the speech). The emerging representations match the veridical units of the language, because (1) Ubiquitous associative learning processes lead to the selection of the most cohesive units, and (2) The language units are defined by the cohesiveness of their components[1]. Note that in simulating the perception of a continuous speech flow as a succession of short sequences that evolved across time, PARSER just mimics the conscious experience of anyone faced with an unknown sequential material, and moreover, the model's evolving units are consistent with the changing phenomenal experiences of the learners throughout the learning session.

PARSER was able to discover all the words and only the words of Saffran, et al.’s language [17], as well as more complex languages in which the simple frequency of co-occurrence was controlled [20], after exposure to still shorter samples of speech than human subjects. It has proven to compare favorably with other models relying on more or less complex computations. Words are discovered without requiring the very sophisticated inferential processes involved in a Bayesian framework [21]. And PARSER is sensitive to more relevant statistical measures (namely: Both backward and forward transitional probabilities between syllables) than the connectionist model often involved in this research context (the Simple Recurrent Networks, SRN, which exploits only the forward transitional probabilities). Also, PARSER learns the words whatever their length, whereas TRACX [22], which relies on a connectionist architecture that basically works as an autoassociator network, is unable to extract monosyllabic words.

Towards a Generalization

Whatever the achievement of PARSER and the support it brings out for applying the concept of self-organization to mental events, there is a striking gap between the few words of an artificial language and the units of real world. Moreover, even if a real-world scaling is arguably plausible, building veridical units seemingly fall short of addressing the full issue of human intelligence.

In principle, generalization of self-organizing processes from words to objects should be straightforward, given that between-component contingency is a relevant measure of internal consistency for both kinds of units. However, the same problem arises for words and objects. Shifting from artificial words to words of natural languages operates a drastic change in complexity. A mental unit corresponding to a word from natural languages is much more than an auditory or graphical pattern, it has a meaning, which is not limited to a dictionary definition, it may be linked to its usual surrounding words, and so on. Likewise, a unit corresponding to an object may include its form and color, but also its label, the context in which it can appear, its affordance (the possibility of action it affords), its affective valence, and several other elements. Thus one may wonder whether simple associative processes are powerful enough to handle increasingly large and sophisticated mental units. Such an objection may seem lethal to those of the workers on connectionism who believe that their feed forward networks, which start from low-level features from the beginning to the end of learning, implement an updated version of earlier associative theories. In fact, in our view, the opposite is true: Connectionist networks represent a deep impoverishment of earlier theories. A standard principle of associative learning is that associations bear on representations, and that, to quote one of the leading theoreticians of animal learning, "*the representation of external events that can enter into such associations may be quite complex*"[23].

Creating veridical representations, i.e. representations that are isomorphic to the structure of the world, could account for a number of phenomena that are often linked to sophisticated computations and rule-based reasoning. Clark and Thornton put forward the idea that neural systems “*trade representation against computation”* [24]. Let us consider the notion of transitivity. In the linear ordering tasks, two premises are presented, such as: A is longer than B and B is longer than C. Participants have to judge whether the conclusion A is longer than C, is correct. It may be assumed that people solve this task because they have some formal notion about the transitivity of the expression “longer than,” and that they apply the transitivity rule to the problem at hand. However, it is far simpler to assume that people have built a conscious representation of the premises in the form of a linear array, and then "read" the response to the question directly on this representation. There is now a consensus about the idea that

[1] The parallel with the self-organizing processes accounting for the complex behavior of social insects is striking. Thermites, for instance, follow a succession of nest building phases: In the first phase, the insects fly in a random pattern, followed by a pillar construction phase, then an arch construction phase, ending with a dome construction phase. All these phases can be explained by the interactions between two simple behavioral principles (a) Thermites move in the direction of the strongest pheromone gradient; and (b) They deposit building materials at the strongest point of concentration, and the physical properties of the pheromones (their diffusion through a gradient) [40,41].

people proceed in this way, and more generally, that much of reasoning depends on mental models [25]. The Shastri and Ajjanagadde model relies on the same general view [26]. Representations become able to provide a model of the world in which some structural relations that have not been encoded as such can be directly "read," instead of being computed through analytical inference processes. Similar ideas have been developed in other research domains, including the instance-based model of categorization [27], the lexicalist approaches to syntactic processing [28], and the memory-based theories of automatism [29] and procedural learning [30].

The Self-Organizing Consciousness as Inspiration for Artificial Intelligence

The Self-Organizing Consciousness model and the neural networks share some of their underlying principles. Of primary importance is their emphasis on learning and, more specifically, on associative forms of learning. But there are also several striking differences. In most neural networks, learning is supervised, while the most fundamental forms of learning are the natural by-product of attentional processing. The input of the networks consists in low-level features which do not change throughout training, whereas the main learning principle we put forward is that the cognitive units evolve at each processing step. Artificial networks manipulate gradual weight between meaningless "neurons" without possible matching with world components, whereas we posit the primacy of discrete and meaningful cognitive units. And finally, most artificial networks follow a bottom-up, or feed-forward, mode of processing, whereas the model we suggest heavily relies on dynamical interactions between bottom-up and top-down processes.

These differences lead to doubt that implementing a self-organizing model is possible through a purely incremental improvement of current AI architectures. Adding still more layers or still enlarging the training sets, even if one assumes an exponential growth of power, could at best lower the rate of errors, but overall looks as a dead-end. We suggest that AI researchers who may wish to introduce algorithms closer to the principles underpinning human consciousness have no other alternative than exploring a radically new avenue of investigation, in which conscious thinking is taken as a model instead of the neuronal structure of the brain.

Discarding neuronal-like organization as a model may appear as outrageously provocative. Indeed, mimicking, even loosely, the interconnectivity of biological neurons seems the best way to obtain the same results as neurons, namely, intelligent behavior. Moreover, the convolutional neural networks have been directly inspired by the discoveries of the Nobel prizes Hubel and Wiesel regarding the neuronal architecture of the visual system, half a century ago, and hence seems to be a promising approach. These arguments are worthwhile, but they are not compelling. No one knows how the activation of biological neurons can lead to mental states. Assuming that the activation of roughly similar artificial nodes will be sufficient to achieve the simulation of mental events seems a bit illusive, and as far as we can say today, not in the process of being validated. Moreover, the suggestion of "discarding neuronal-like organization" should not be seen as a denial of the biological roots of consciousness. Conscious thought is obviously the end-product of biological processes, and therefore, taking conscious functioning as a model does not amount to ignore the neurobiological underpinnings. Artificial neural networks simulate the neural level without any attempts to reproduce the biochemical mechanisms involved in living organisms. Our suggestion of simulating mental events without any attempts to reproduce the neuronal mechanisms involved in humans amounts only to a shift in the level of analysis, which can hardly be described as heretical. Note in addition that the strong reliance on top-down processing intrinsic to our approach (see below) could be thought of as more faithful to the overarching structure of the brain than mimicking some of its components.

What are the concrete implications? Needless to say, the implications do not consist in tagging as "conscious" some or all of the processes involved in a computational model, whatever they may be. To be considered as potentially conscious, a representation or operation needs to comply with precise and generally restrictive criteria, which are the only ones of importance. At the extreme, the explicit reference to consciousness could even be passed over in silence, provided that the model meets the relevant criteria. For instance, McCauley and Christiansen state that PARSER is "perhaps the most influential chunking model devoted to implicit learning and word segmentation" [31], without any mention, whether positive or negative, that the model simply reproduces the learners' conscious thoughts throughout learning.

A first consequence of taking mental/conscious functioning as a model is the exclusive reliance on meaningful and discrete units. This constraint goes much further than a simple localist coding of information, which only refers to the input layer. All time steps are concerned. To perceive and understand the world, shaping the structure of representations over the structure of the world seems to be the best approach. Now, it is indisputable that the world is made up of discrete units, from the grain of sand to the stars, going through objects, plants, animals or humans (not to speak of atomic and subatomic levels). Of course, these units and their interactions generate statistical regularities between elementary features as a by-product, and hence, capturing and exploiting these regularities may allow to simulate to some extent the understanding of structural relationships. This is the principle of any connectionist approach. However, this approach is limited because the world is not causally governed by statistics and probabilities, and starting from these by-products is likely to never be able to trace back to the actual causes and consequences.

A second crucial consequence is a massive top-down organization, in striking contrast with the usual feedforward networks. One could argue that recurrent networks implement some top-down structure, but their general objective, mainly processing sequential material, is very different. The Long Short-Term Memory (LSTM) networks are a bit nearer of what we intend to mean, but remains nevertheless far behind [32]. The top-down organization typical of conscious processing is the most radical that we can imagine: What is perceived is only composed of a few previously learned units, although possibly in new combinations. The end-result is that most of the sensory information objectively available at a given time step is overlooked. Of course, the units that turn out to be retained are not selected randomly. They depend on the properties of attention. In particular, elements in close temporal or spatial contiguity are more likely to be captured in a same attentional focus than spread out elements, and new combinations are privileged.

A question immediately arises: Why is it necessary to implement in overpowered computers what is usually thought of as a damaging limitation of human minds, such as the attentional bottleneck? The main response is that in order to create representational units isomorphic to the world components, considering a restricted set of features is certainly the best, if not the unique solution. Human associative learning is at the root of our most fundamental acquisitions, and an ubiquitous, although often ignored, observation is that what becomes associated is always limited to the very few features that capture attention at a given moment, or in other words, to our fleeting conscious contents. This is true for conditioning [33], implicit learning [34], and statistical learning [35]. There is an overwhelming number of associations that could be detected in the input by an efficient, advanced system such as a deep learning network, but, arguably, most of these correlations are useless or misguiding. If extracting exhaustively the complex network of associations present in the world was adaptive for humans, it would be quite surprising that evolution has led to restrict associative or statistical learning to the very few elements falling into the conscious focus.

Some of the reasons making that extracting the whole pattern of correlations would be detrimental to the discovery of genuine world units are worth examining. A rather trivial reason is that contextual information is often irrelevant for the discovery of genuine units. A pencil is usually on a table or desk. But a pencil on a high shelf or under a refrigerator is still a pencil. In an unusual location, the object doesn't reduce its probability of being a pencil from, say, 95% to 55%. But there is another, more subtle reason. World's units are hierarchically organized, and correlations make sense only within one level of organization. To illustrate with language, children create words or a few words units from syllables, and for this task, statistical relationships between syllables are relevant. Statistical learning is also useful to understand or generate syntactically correct sentences [36]. However, for these tasks, statistical relationships are relevant only if words are taken as new coding units. Relationships between, say, the last syllable of a word and the first syllable of the next word, which were relevant to discover the words, act now as noise. This illustrates the principle coined as "complexity from noise" by Atlan [37], whereby the noise at a given hierarchical level constitutes the information at another hierarchical level. As a consequence, extracting all present associations at once is doomed to failure or approximation, because mixing several hierarchical levels amounts to mix relevant and irrelevant associations at a given time step. It is common to speculate that in deep learning networks, hierarchy is encoded into successive layers, the abstraction level being increased from every layer to the next [38]. However, this hypothetical process is assumed to occur on each processing step, whereas in the operations we suggest, the units from a given hierarchical level must be consolidated before shifting to the next.

Is an Approach Modeled on Conscious Mode of Functioning Worth Exploring?

Implementing all these constraints implies deep and costly changes. The resulting model would need to manipulate symbols, or at least discrete units. Neural networks use discrete units only in the input layer under the form of low-level and unchanging features, and in the final step of processing when a discrete response (e.g., "dog") must be selected from a set of possibilities on the basis of probabilistic information (e.g., dog = 60%, cat = 40%). They play no active role. If the system is devised to mimic conscious thought, then the functional elements of processing have to be discrete and possibly complex units. They have to shape the coding of incoming information, and they must be the basic stuff of memory and learning processes. Current neural networks are notoriously ill-suited for this task. Note that old-fashioned symbolic systems are not better. Conscious units are not the ready-to-use and immutable rules and concepts coming from human experts: They are built by experience, and they evolved at each time step as a function of the exposure to the data.

Where does this lead us? Since its inception by Newel and Simon on the one hand, and Rosenblatt's perceptron on the other hand, the field of AI has constantly balanced between symbolic and connectionist architectures, so that they are now seen as the unique viable options. Implementing our framework could require breaking this dualistic landscape, in order to open a third generation of models. Note that current hybrid architectures [39], in which a connectionist network send its output to a symbolic system, do not do the job better, because they miss the dynamic induced by the fact that discrete units are the sole matter of processing.

PARSER suggests a promising direction. At any given time step, the units built at earlier steps serve to handle the current input, and they are the matter for the very same processes that were involved in their formation. As a consequence, the hierarchical structure of the world is naturally taken into account, without any "manual" change or intervention. The system works because the same general property is assumed to define the units at all hierarchical levels, namely the mutual dependency or cohesiveness between units components, and the same processes (mainly forgetting and associative processes) are assumed to be able to exploit this property whatever the complexity of the units. However, PARSER is more like a toy model providing a proof of concept in a restricted domain. Building an AI that scales up PARSER's basic principles to address worldwide issues remains a challenging task for AI experts.

Whether this task is worth undertaking depends on whether such new AIs would be likely to overcome the problems of deep neural networks. We listed three main problems in the introductory section. Regarding the time course of learning and the need for very huge databases, a positive response appears obvious, as suggested by the fact that PARSER learns to segment an artificial language with a much more limited amount of exposures to the data than all the concurrent models. In PARSER, a provisional unit is created on the fly, which is reminiscent of a common experience of each of us: The nearly immediate (although sometimes short-lived) learning of, say, the name of a new acquaintance, from a single episode.

Task specificity certainly raises the strongest reluctance to talk about intelligence in deep networks. There are a number of reasons leading to expect the mitigation of this concern in a model based on self-organizing consciousness. The formation of meaningful units is certainly a major one. Indeed, a new situation can often be construed as a recombination of learned units. To illustrate with language, processing any sentence as a whole would lead to start from zero when faced with a new sentence. However, a new sentence is nothing else that a recombination of known words, so that words that have been previously learned can be directly exploited. It appears again that the exhaustive analysis of whole scenes typical of deep learning systems is a dead end.

A generalized top-down architecture also prevents the formation of over-specialized processes. For humans, a number of questions that arise whenever one postulates that the incoming information is processed exhaustively on a bottom-up mode by unconscious processes, are not even an issue in our view. For instance, in the standard framework, the formation of categories needs ad-hoc mechanisms to explain why distinct objects are processed in the same way. By contrast, the formation of categories naturally follows top-down processing, because letting aside idiosyncratic details happens as a necessary consequence of this mode of processing. Imposing the same meaning to superficially different objects leads naturally to deal with these objects as if they belonged to a same category, and by the same token, enlarges the scope of this category for further processing episodes. Behaviors such as transfer, generalization, and analogy making may receive the same kind of explanations. Instead of postulating that the incoming information is processed in all its detail, and subsequently analyzed to examine whether the present input is similar on some dimensions to known situations, positing the primacy of top-down processing eliminates the problem: A new situation is directly perceived through the filter of familiar units, and these units are modified in turn to encompass future events. Ironically, once again, rejecting the existence of a powerful symbolic unconscious paves the way for much more parsimonious accounts of mental events. The advantage of shallow top-down processing could be directly transposed to AI. Arguably, the prevalent reliance of neural networks on feed-forward algorithms leads to hyper-specific responses, which have only limited relevance to related situations.

Finally, what about errors? Taking the conscious mind as a model certainly does not prevent the occurrence of errors: Humans themselves are notoriously error- prone, especially during a learning episode. However, strange, non "human-like" errors typical of deep neural networks should be avoided, notably because the guided selection of a few features, mimicking attentional selection in humans, should prevent the exploitation of spurious associations. Moreover, given that only meaningful units would be implied all along the learning process, potential errors would be easier to understand and therefore easy to repair. Among other consequences, this would lead to much more reliable systems.

Concluding Remarks

In humans, adaptive conscious contents may emerge naturally from the dynamical interplay between the properties of conscious thought on the one hand, and between these properties and the external world on the other hand.

Endorsing this view led us to argue that deep neural networks, despite their indisputable successes, are engaged in the wrong way as long as they intend to mimic human thought. The AI researcher Melanie Mitchell sees it as likely that "*the supposed limitations of humans are part and parcel of our general intelligence [2]. The cognitive limitations forced upon us by having bodies that work in the world, along with the emotions and "irrational" biases that evolved to allow us to function as a social group, and all the other qualities sometimes considered cognitive "shortcomings," are in fact precisely what enable us to be generally intelligent rather than narrow savants. I can't prove it, but I think it's likely that general intelligence can't be separated from all these apparent shortcomings, in humans or in machine*". We believe that the AI mainstream, grounded on the belief that intelligence rests on boundless processing capabilities, housed in ever-larger memories and ever-faster processors fell into a conceptual trap.

We suggest that taking conscious thought as a model rather than the neuronal micro-structure of the brain should be one of the new avenues to be explored. Our view is that no genuine understanding can emerge from an analysis of the statistical properties of world elementary features, no matter how complete and extensive this analysis may be. In the alternative framework we suggest, discrete units are processed in the context of a top-down architecture, which ensures that the incoming information is dealt with, as far as possible, as a function of the units created in previous experiences. PARSER, our computational model of word segmentation, may serve as an illustration of this radically different approach. Note that in PARSER, in keeping with Mitchell's intuition, relevant units are extracted, not "despite of", but thanks to, what is usually construed as processing shortcomings. Without selective attention and forgetting, PARSER would not work. However, as fully acknowledged, PARSER is hardly more than a proof of concept for a set of principles. We hope that these principles, and more generally the idea of complex representations emerging through self-organization, may serve as an incentive to elaborate powerful new algorithms for an AI that would more closely approximate human intelligence.

References

1. Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. ICLR: 1412.6572.
2. Mitchell M (2019) Artificial Intelligence: A Guide for Thinking Humans. Farrar, Straus & Giroux, New-York.

3. Clore GL, Proffitt DR (2016) The myth of pure perception. Behav Brain Sci 39: e235.
4. Chater N (2019) The Mind is Flat: The Illusion of Mental Depth and The Improvised Mind. Penguin books.
5. Velmans M (1998) Goodbye to reductionism. In SR Hameroff, AW Kaszniak & AC Scott (Eds.), Toward a science of consciousness II, p. 51. Cambridge, MA: MIT Press.
6. Mack A, Rock I (1998) Inattentional blindness. MIT Press, Cambridge, MA.
7. Most SB (2013) Setting sights higher: Category-level attentional set modulates sustained inattentional blindness. Psychol Res 77: 139-146.
8. Howe PD, Carter OL (2016) Hallucinations and mental imagery demonstrate top-down effects on visual perception. Behav Brain Sci 39: e248.
9. Rolfs M, Dambacher M (2016) What draws the line between perception and cognition?. Behav Brain Sci 39: e257.
10. Perruchet P, Vinter A (2002) The self-organizing consciousness. Behav Brain Sci 25: 297-388.
11. Nicolis G, Prigogine P (1977) Self-organization in non-equilibrium systems. J WileySons, London.
12. Camazine S, Deneubourg JL, Franks NR, et al. (2003) Self-Organization in Biological systems. Princeton University Press.
13. Thelen E, Smith LB (1994) A dynamic systems approach to the development of cognition and action. MIT Press, Cambridge MA.
14. Smith LB (2005) Cognition as a dynamic system: Principles from embodiment. Develop Rev 25: 278-298.
15. Miskovic V, Kuntzelman K, Chikazoe J, et al. (2016) Representation of affect in sensory cortex. The Behav Brain Sci 39: e252.
16. Hackel LM, Larson GM, Bowen JD, et al. (2016) On the neural implausibility of the modular mind: Evidence for distributed construction dissolves boundaries between perception cognition and emotion. Behav Brain Sci 39: e246.
17. Saffran JR, Aslin RN. Newport EL (1996) Statistical learning by 8-month-old infants. Science 274: 1926-1928.
18. Perruchet P, Vinter A (1998) PARSER: A model for word segmentation. J Memory Lang 39: 246-263.
19. Rescorla RA (1968) Probability of shock in the presence and absence of cs in fear conditioning. J Comp Physiol Psychol 66: 1-5.
20. Perruchet P, Poulin-Charronnat B (2012) Word segmentation: Trading the (new but poor) concept of statistical computation for the (old but richer) associative approach. In: P Rebuschat J, N Williams (Eds) Statistical learning and language acquisition, De Gruyter Mouton, Berlin, Germany 119-143.
21. Frank MC, Goldwater S, Griffiths TL, et al. (2010) Modeling human performance in statistical word segmentation. Cognition 117: 107-125.
22. French RM, Addyman C, Mareschal D (2011) TRACX: A Recognition-Based Connectionist Framework for Sequence Segmentation and Chunk Extraction. Psychol Rev 118: 614-636.
23. Mackintosh NJ (1997) Has the wheel turned full circle? Fifty years of learning theory 1946-1996. The Quarterly J Experiment Psychol 50: 879-898.
24. Clark A,Thornton C (1997) Trading spaces: Computation representation and the limits of uninformed learning. Behav Brain Sci 20: 57-90.
25. Khemlani S, Johnson-Laird PN (2012) Theories of the syllogism: A meta-analysis. Psychol Bull 138: 427-457.
26. Shastri L, Ajjanagadde V (1993) From simple associations to systematic reasoning. Behav Brain Sci 16: 417-494.
27. Brooks LR (1978) Nonanalytic concept formation and memory for instances. In: E Rosch, B Lloyd, Cognition and categorization. Lawrence Erlbaum, Mahwah NJ.
28. Bates E, Goodman JC (1999) On the emergence of grammar from the lexicon. In: B MacWhinney, The emergence of language, Lawrence Erlbaum, Mahwah NJ.
29. Logan GD (1988) Towards an instance theory of automatization. Psychol Rev 76: 165-178.
30. Chen Y, Orr A, Campbell JID (2020) What is learned in procedural learning? The case of alphabet arithmetic. J Exp Psychol Learn Mem Cogn 46: 1165-1177.
31. McCauley SM, Christiansen MH (2019) Language learning as language use: A cross-linguistic model of child language development. Psychol Rev 126: 1-51.
32. Hochreiter S, Schmidhuber J (1997) Long Short-term Memory. Neural Comput 9: 1735-1780.
33. Mackintosh NJ (1975) A theory of attention: Variations in the associability of stimuli with reinforcement. Psychological Review 82: 276-298.
34. Hsiao AT, Reber AS (1998) The role of attention in implicit sequence learning: Exploring the limits of the cognitive unconscious.
35. Wang FH, Zevin JD, Trueswell JC, et al. (2020) Top-down grouping affects adjacent dependency learning. Psychon Bull Rev 27: 1052-1058.
36. Williams JN, Rebuschat P (2011) Statistical learning and syntax: What can be learned and what difference does meaning make? In: Statistical Learning and Language Acquisition, P Rebuschat and JN Williams, De Gruyter Mouton, 237-264.
37. Atlan H (1972) L'Organisation biologique et la Théorie de l'information Hermann Paris (rééd 1992).
38. Wlodarczak P (2019) Deep Learning in eHealth In: VE Balas SS Roy D SharmaP Samui (Eds) Handbook of Deep Learning Applications Springer.
39. Sun R, Merrill E, Peterson T (2001) From implicit skills to explicit knowledge: A bottom-up model of skill learning. Cognitive Science 25: 203-244.
40. Hill N, Bullock S (2015) Modelling the Role of Trail Pheromone in the Collective Construction of Termite Royal Chambers. Proceedings of the European Conference on Artificial Life 43-50.
41. Kugler PN, Turvey MT (1987) Information natural law and the self-assembly of rhythmic movement. Erlbaum Associates, Hillsdale NJ.

DOI: 10.36959/643/307