Understanding the Meaning of Understanding: a Possible Way to Self-Consciousness

Daniele Funaro

Understanding the Meaning of Understanding: a Possible Way to Self-Consciousness

Daniele Funaro

Dipartimento di Scienze Chimiche e Geologiche
Università di Modena e Reggio Emilia
Via Campi 103, 41125 Modena (Italy)
daniele.funaro@unimore.it

Abstract

Can we train a machine to detect if another machine has understood a concept? In principle, this is possible by questioning the second machine on the subject of that concept. However, we want this procedure to be done by avoiding direct queries. In other words, we would like to isolate the absolute meaning of an abstract idea such as "having understood" by putting it into a class of equivalence, hence without adopting straight definitions or showing how the idea works in practice with the help of tests. We discuss the metaphysical implications hidden in the above question, with the aim of providing a plausible reference framework. This could also clarify how the mechanism of self-consciousness develops, which requires however an interplay between members of a community. Indeed, we claim that "having knowledge of a concept" becomes an act of consciousness only when opinions are confronted with other actors, belonging or not to the same class of equivalence.

Keywords: Machine Learning, Self-Consciousness, Knowledge Representation, Knowledge Reasoning.

1 Preamble

In his famous tale, *Las ruinas circolares* [1], Borges depicts the efforts of a man in his struggling endeavor to create, through his own dreams, a new life. The accomplishment is to induce the dreamed subject to become an independent virtual son, able to live a proper life disjointed from the dreams of his creator. The story ends with the stunning discovery that the original man is himself the "materialization" of the dreams of another entity.

The dreamer and the dreamed subject belong to different "realities", that in the Borges' fiction end up to be both "virtual". At a first glance, there is

no direct connection between these universes. If ours is somehow the world of reality, the seemingly dissociated domain of our thoughts is usually defined as intangible. One is tempted to attribute a superior level of abstraction to the second environment, though this is not necessarily true, according to the circularity of Borges' arguments.

2 Motivations

One of the aims of modern programming is to instruct machines to learn according to apprehending processes that try to mimic those followed by humans. A recurrent question is what an inanimate bunch of semiconductors may have understood about the lessons imparted; that is: how a machine "visualizes" in its own "mind" the product of new discoveries? Does it have knowledge of its knowledge? Can other machines know about its thinking "just by looking in its eyes"? These philosophical issues are akin to the thematic of automatic self-consciousness [2, 3]. An attempt to provide some answers is tried in these few pages. It will be argued that abstract concepts do not follow from definitions or by direct algorithms, but they might be ruled by the same mechanism that allows one to achieve the first levels of discernment. In addition, the role of the society must not be underestimated. Here, the word "society" denotes a collection of individuals, non necessarily of human type.

3 Discussion

With the supervised help of a teacher, a child can refine the notion of color (red, for instance) through examples, by learning how to construct and assign names to specific classes of equivalence (Fig. 1). Some notions could be actually innate [4]. Moreover, the individuals of a "society", already aware of those primary concepts, can play a fundamental role in the instructing process. By a similar training, a machine can recognize if there is a cat on a table. Adding deeper and deeper layers of training, the same machine can learn to recognize a black cat on a wooden table, lapping milk from a cup. Despite the increasing complexity of the details, the above training sets belong to the space of reality, while the final result (i.e., the knowledge) looks, in some way, more "abstract". The last observation is indeed incorrect from the technical viewpoint, since both the images of the cat and the numerical outcome (the so-called weights) of the brain of the instructed machine are represented by sequences of the same type of physical bits. It is only our preconceived intuition of reality that tends to assign different levels to these categories.

At this point, one may ask: how do we know if a child has a clear idea of the abstract meaning of red? The exam is simply done by submitting to his/her attention one or many objects, and pose questions about their colors. Neglecting possible shades of randomness, this analysis is fast and secure, since is exactly based on the same apparatus that generates the skill of distinguishing

colors.

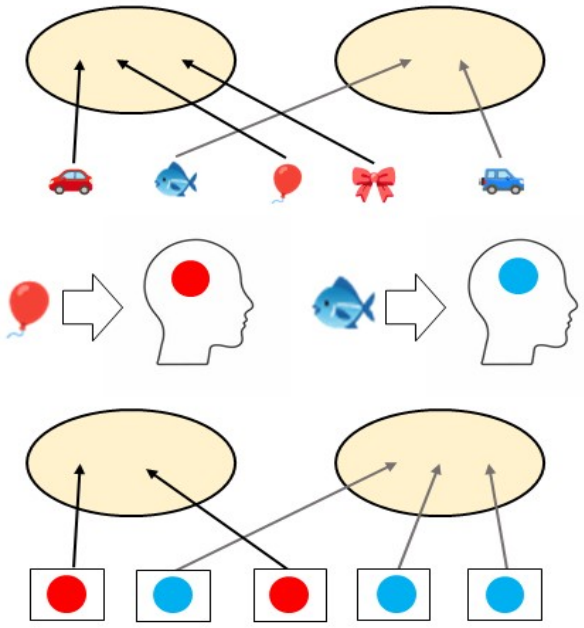

Figure 1: A learning process consists in building specific classes of equivalence (the ones corresponding to the colors "red" and "blue", for instance). The notion of color is imprinted in our brain. Theoretically, we can compare the classes "knowledge of red" and "knowledge of blue" by directly examining the so modified structures of the brain (bottom picture), instead of asking questions about colors to the single individual.

Can we do the above check indirectly, thus without showing to the child any object? In some parts of the child's synapses the activation of a certain concept (innate or acquired) has altered preexisting connections. The study of these new links may give an answer to our question without relying on the practical experiment. In a very similar way, the machine's concept "cat on table" resides in a memory made of silicon-based circuitry, under the form of a peculiar distribution of data. Interpreting these data may teach us if (and maybe what) the machine has formally understood.

Unfortunately, reading a single computer's memory and trying to deduce something is like acquiring the notion of red through the realization of just one test. Therefore, it is advisable to play with a series of trained and untrained devices, in order to make comparisons and come to conclusions. The path to be followed is the same inspiring the initial training procedure used for reality. This will be "supervised" until the machine acquires independence. Such an

algorithm does not necessitate the submission of further cats' pictures. It is an analysis made at a different level, like the dream that exists in a more profound layer with respect to that of the dreamer. The commitment of the learning machine is to distinguish by comparing the devices that "know" from those that "do not know", without studying "what they actually know". It has been already noticed however that all the levels of abstraction are similar from the technical viewpoint. In truth, following Borges, the dreamer himself is the product of the "imagination" of another dreamer, and, in practice, we have no means to distinguish a dream from a dream into a dream (Fig. 2).

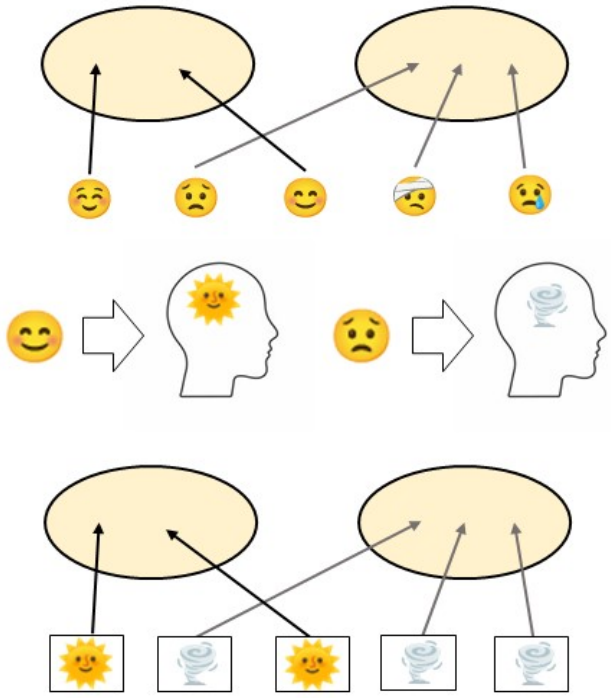

Figure 2: A learning process can be extended to more abstract categories ("happiness" or "sadness", for instance). The analysis follows however a path similar to that applied to real objects. Hypothetically, it would be possible to deduce that a patient is depressed without a visit, but through the confrontation of the brain activity with that of other patients already classified into the category "depressed".

In other words, it would be possible to understand if a device has understood something, through a procedure that does not require direct questions of any type. Since this construction is made with the help of another training history

(at upper level), we cannot mathematically define (not even a posteriori) what kind of configuration must be actually present in a prescribed instructed machine to be tested. This is not surprising, because it is similar to our incapability of providing an absolute definition of red without indicating an example. "Red" is a class of equivalence. In the same way, "have the knowledge of red" is another class of equivalence; there is no official indication of the elements of this class, but only examples of elements sharing the same properties. Our mind is modified as we add new knowledge, however this process is very subtle, so that we cannot practically put into words the details of these changes (to explain for instance in which area of our brain and under what form those data are present).

The reliability of a set of CPUs, programmed to face AI problems, could be in principle verified by plugging electrical supply, but avoiding the use of any peripherals. Here the purpose is not just the check of their plain functionality, but to test their supposed capability to apply intelligence. For instance, this kind of training may teach a robot to make a decision not only on the basis of what other robots do, but on what they are thinking (if the data of their central systems are available, such as the value of the weights of a certain Machine Learning process). The instruments to carry out this analysis are standard, although applied to a context that looks upgraded. As pointed out in a recent article [5], even current codes for image understanding may fail when tested on appropriate nasty examples. What may happen with abstract concepts is at this stage unpredictable, since it is certainly not an easy task to predict what is the percentage of trustworthiness of these outcomes, which are surely affected by large error spreading. A long phase of experimentation is then necessary. Since this dissertation is only finalized to the description of the basic principles, we only provide a few guiding theoretical advice. Thus, we do not discuss in this paper any concrete development process, leaving to the experts the implementation of the instances here exposed.

4 Remarks

We proceed in this short disquisition with a warning. Trespassing the privacy of an individual to know if he/she has well elaborated the concept of red, without posing straight questions, could be a first step to surreptitiously discriminate peoples on the base of political inclinations, sexual attitudes or whatever an organ of control wants to know. It is worthwhile to recall once again that here this kind of analysis is not directly constructed on specified parameters, but to the belonging or not of the individual to classes already constituted. The purpose of the machine is to classify an individual through a characterizing history (websites visited, for example), without examining substantial real facts, but rather the general activity in the framework of a list of prejudged individuals presenting a well prescribed property.

Without an IQ test, a person could be recorded as an "intelligent fellow" because of his/her affinity to representatives initially present in that category

(the training set), without explicit notion of the properties characterizing that class. In fact, the setting up of the class itself is the result of a previous analysis carried out on individuals declared "intelligent" or "not intelligent" in advance. Thus, the decision is taken as a consequence of the "way to behave", and not on the capability of "acting intelligently" in the solution of a given problem; this without the necessity of formalizing officially how an intelligent person is expected to behave. It is evident that such a superstructure has ethical implications, so that it must be used wisely.

In terms of AI, the memory maps of a new machine are compared for example with those belonging to the class of "intelligent machines" and the response is made without further external checks. Note that clear traces denoting "intelligence" could not be present in the preexisting elements of the training set (actually, we do not even know how these traces look like), indeed such elements have been just selected on the base of the claimed intelligence of the machines to which they belong. Referring to Fig. 2, an examiner could be able to recognize a happy guy if some parameters allow for his insertion into the category "happy guys". There is not need to meet the guy in person, and the parameters could not be directly associable with our current notion of "happiness".

In this last paragraph the focus is concentrated on another crucial observation. It is possible to retain the idea of a specific color without naming it. By the way, the reason why the word "red" is the name of the class of all red colored objects, comes from the necessity of confronting each one's discoveries with other peoples (see at the beginning of section 3). The abstraction of the term is actually born the moment it becomes a product of the collectivity. Thus, the interpretation of a thought comes naturally as a result of the comparison of many minds, as also punctuated before in this paper. We believe that this is the first step towards a formalization of the meaning of self-consciousness, as it will be discussed later in section 6.

5 Contextualization

The topics touched in this brief exposition are certainly not new [6, 7, 8, 9]. They assume however a wider relevance in this specific moment, in which the field of Machine Learning is experiencing a positive period of growth, both in applications and complexity. In the review papers [10, 11, 12], future developments in the field of Deep Learning (and more in general AI) are addressed. Among the various disciplines, Reinforcement Learning is also gaining popularity [13]. There, a progressive tune up of the *policy* is wisely applied to optimize the so-called *return*. For instance, in [14] and [15], this type of training has been implemented without human assisted supervision, and can represent a first attempt to guide a machine to acquire self-knowledge. As it will be pointed out several times in this paper, there is a hidden difficulty in going ahead with this construction, i.e., the device will not be conscious of its own understanding, until this "state" is shared with other entities (*I understand we both have understood, because we "feel" it in the same way*).

Explainable Artificial Intelligence (XAI) [16, 17, 18, 19, 20, 21, 22, 23, 24], is aimed to improve the performances of AI by better clarifying human mental models. The discipline remains at an empirical level and focuses on specific examples, without catching the underlying potentialities of a systematic use of Machine Learning at more sophisticated degrees of abstraction, as proposed in this paper. In addition to these aspects, efforts have been made in order to understand human functions [25], define "interpretability" [26], or associate increasingly complex concepts, with the help of always more sophisticated modules acting on data and accomplishing upgraded tasks [27]. In [28], the concern is to provide a robot with sophisticated skills, in order to be able to recognize aspects of human behavior. This implementation, obtained by assembling specialized modules, is a prerogative of a single machine through a process of identification at various stages, similar, more or less, to what happens in Deep Learning. The design requires strong human assistance to be initialized, since the building process translates into machine language, the results of the experiences commonly lived in reality. The goal is more similar to the effort of creating a sort of human clone, rather than letting the machine develop a proper way of reasoning. The above mentioned approaches are then quite different from the one here discussed, where "understanding" is not viewed as a "complexification" of the bottom, but as a concentrate of the experiences of a community, that can be extracted on the base of the same principles ruling human connections with reality.

One may try to establish intersections between our approach and the so called *Theory of Mind* (ToM) [29, 30, 31], which represents an efficacious instrument of analysis in the sociological and psychological contexts. In such a discipline, governed by empiricism, part of the effort is concentrated on the study of the various stages of development, where humans acquire knowledge and understanding, through a systematic process named: "learning the Theory of Mind". Again, the translation of these achievements into the machine language seems to follow a path which is different from what suggested here. In truth, we do not want to teach anything to a computer or transfer our "vision" into it. Instead, the machine has to learn its own ToM. For instance, a computer may autonomously build the concept of *wellness*, after examining a series of people declared by a supervisor to be joyful or sad, maybe because they laugh, cry, or move their face in a bizarre manner (see again Fig. 2). The results of this training are, in general, not decipherable, as the machine in its own analysis could emphasize aspects of the individuals that we do not even observe or imagine. At the end of the process, we do not need to know the definition of "wellness" apprehended by the machine. On the other hand, if we had a definition of wellness, we could have directly imparted it into the machine from the very beginning. The learning process is satisfactory if somehow (with a margin of error) the machine has "understood" in its own way, and it is able to operate accordingly. There is no need to care about the format of these notions, if the machine can finally do the job, for which it has been trained, in the proper way. At higher level, future machines could not necessitate instructions from people, but they will talk, exchange information, and create new cerebral

connections that have nothing in common with those usually developed in humans. By following this approach, in the technology of tomorrow, no human could be in the position to understand what computers actually have in mind.

6 Self-consciousness

Going into a more sophisticated area, a possible extension of these considerations can be applied to the field of consciousness, though the approach may be judged a bit risky (or naive). Consider the phrase: *I know that I am conscious because I can share this opinion with other people, and not because I can universally define such a feeling*. Again, following this path, the term "self-knowledge" applied to an individual turns out to be an element of a class of equivalence; therefore, it should be studied within this frame of reference. Thus, based on the material discussed in the present paper, in order to be built, abstraction necessitates of both "reality" and the (implicit or explicit) request of a community; hence it cannot be the consequence of the direct experience of a singleton. We can make this idea clear with an example. Let us suppose that a set of automata learn to play chess and refine their capabilities by continuously challenging each other. Will be they conscious of being chess players? The answer is **no**, from the simple reason that there is no utility to develop such a knowledge, unless the machines do "decide" together that there is the necessity to build the class of "chess players", with the purpose to distinguish their ability from the state of other existing machines that do not even know the basic rules of the game. Recognizing to be part of that class is an act of consciousness, although one may argue that this notion is rather *weak* in comparison to more advanced forms of awareness. Belonging to the class of chess players becomes an act of consciousness only when the category of non chess players has been identified. If I do not play just for fun, and for any victory there is a bonus, then it is better to know my skills before facing the enemy.

Thus, the convenience to give origin to specific membership classes may be due to some (external) forms of gratification. To this purpose, a device may be supplied with *ad hoc* registers aimed to classify and publicly advertise, the current level of capabilities and a certain degree of "satisfaction". In a similar way, species evolution on Earth has been driven by sources of extrinsic stimuli, differentiating the individuals into classes. For example, the nodes of a complex web may start challenging each other, if some form of benefit comes from the dispute. Will they become aware of their potential strength as the course of time passes? Note that an isolated single element cannot become conscious by itself, because, according to our view, such a problem is ill-posed.

To recap, the fundamental steps of our construction are the following ones:

- Through a procedure of *Learning by Examples* it is possible to partition a data set in classes. For instance, the class named "red" contains all elements corresponding to our notion of that color (Fig. 1). We do not need, however, to assign a name to that class, unless this is somehow

required by practical applications. A child can play with colored bricks and create a blue house, without necessarily knowing that the final realization is actually "blue". The abstract name emerges when interacting with other children: *would you please give me another blue brick?*

- By a construction similar to that suggested in the caption of Fig. 2 (only apparently at a more abstract level), one can generate the class "knowledge of red". The elements of this class are the individuals who are aware of the existence of the color red and can recognize it among others. This does not automatically imply that these individuals are conscious of being part of the class.

- Self consciousness is founded on the previous requisites, but necessitates a further requirement. Being a member of the class "knowledge of a color" becomes an act of self-consciousness when this is associated with some benefit: *I am conscious of knowing the importance of the color "gold" and I can make profits based on it*. Here, the unavoidable interaction with other individuals constitutes the stimulus for the generation of the new class of peoples who have "knowledge of the color gold and are aware of it".

In contrast to what has been just specified, current research in Artificial Consciousness is aimed to extract definitions and characterizations in human natural activities [32, 33] to be translated in computational models (see [34, 35, 36], as well as [3] for a thorough review of the major achievements). In the *Searle's Chinese room* argument [37] (see also [38] and the references therein), cognitive modeling becomes a problem of semantics. Again, the interpretation of the outside world starts from the way we are able to describe it, and involves direct interactions between single entities (humans and machines). In other words, sociologists and philosophers are still trying to catch the absolute meaning of a concept in order to explain it to a machine. This approach, in vogue among engineers 60-70 years ago, is far from that pursued in this paper, which follows a viewpoint more pertinent to what nowadays technologies call Machine Learning. In [39] we can find the following statement: *consciousness corresponds to the capacity to integrate information*. Though we are not moving here in the direction indicated in that paper, we recognize a vague resemblance with some basic concepts.

7 Conclusions

The rationalization process of mathematical type, described so far, involves the classification of objects or abstract entities into classes of equivalence. We renounce however to give a definition to the elements of these classes, though we know that each class contains elements of the same nature. Classes can be associated with a name (a red hat belongs to both the classes "red" and "hats"). However, names come after the construction of a class and are used to

communicate to other individuals that something has been apprehended from nature and that such experiences are waiting to be shared. This is different from assuming to have a name (hence, a characterization) and collect together all the entities under that name. We cannot create the notion of "good guy" from scratch, but we can recognize a good guy among a multitude of fellows. This is because our mind, with observations and the exchange of information, has generated the appropriate class of equivalence. Classes can be generated at any level of abstraction and complexity, up to the top level, where a class contains in the labeling the annotation that its elements are aware of being part of that class. In this context we can somehow claim that the members of the class have reached a certain degree of self-consciousness. Regarding the practical viability of these ideas, we are not able here to investigate further, so that the turn now passes to the experts. We should however be careful, when establishing parallels between our mind and the work of a machine. Due to the complexity of biological systems, these types of connections are at the moment very mild. Human beings went through a long process of evolution. Experiences of a single life mix up with innate structures that are inherited from generations, therefore these last aspects should not be underestimated.

References

[1] Borges, J.L. *Ficciones*, SUR, Buenos Aires, 1944.

[2] Chrisley, R. Philosophical foundations of artificial consciousness, Artif. Intell. Med., 2008, 44, 119-137.

[3] Reggia, J.A. The raise of machine consciousness: studying consciousness with computational models, Neural Networks, 2013, 44, 112-131.

[4] Chomsky, N. *Syntactic Structures*, Mouton, The Hague, 1957.

[5] Rosenfeld, A.; Zemel, R.; Tsotsos, J.K. The elephant in the room, 2018, arXiv:1808.03305.

[6] Michalski, R.S. Understanding the nature of learning: issues ad research directions, in *Machine Learning: An Artificial Intelligence Approach*, Vol. 2 (Michalski, R.S.; Carbonell, J.G; Mitchell, T.M. Eds.), Morgan Kaufmann, Burlington MA, 1986.

[7] Bishop C. M. *Neural Networks for Pattern Recognition*, Oxford Univ. Press, NY, 1996.

[8] Russell, S. J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, Pearson Education, London, 2010.

[9] Goodfellow, I.; Bengio Y.; Courville, A. *Deep Learning*, MIT Press, Cambridge, MA, 2016.

[10] LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning, Nature, 2015, 521, 436-444.

[11] Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C.J.; Müller, K. -R. Explaining deep neural networks and beyond: A review of methods and applications, Proc. IEEE, 2021, 109(3), 247-278.

[12] Jiang, Y.; Li, X.; Luo, H.; Yin, S.; Kaynak, O. Quo vadis artificial intelligence?, Discov. Artif. Intell., 2022, 2(4).

[13] Sutton, R.S.; Barto, A.G. *Reinforcement Learning – An Introduction*, MIT Press, 1998.

[14] Singh, S. Learning to play Go from scratch, Nature, 2017, 550, 336.

[15] D. Silver et al. Mastering the game of Go without human knowledge, Nature, 2017, 550, 354-359.

[16] Baehrens, D; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; Müller, K.-R. How to explain individual classification decisions, J. Mach. Learn. Res., 2010, 11, 1803-1831.

[17] Landecker, W.; Thomure, M.D.; Bettencourt, L. M. A.; Mitchell, M.; Kenyon, G. T.; Brumby, S. P. Interpreting individual classifications of hierarchical networks, Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM), 2013, 32-38.

[18] Castelvecchi, D. Can we open the black box of AI?, Nature, 2016, 538, 7623, 20-23.

[19] Lundberg, S. M.; Lee, S. A unified approach to interpreting model predictions, Proc. Adv. Neural Inf. Process Syst., 2017, 30, 4765-4774.

[20] McDermid, J. A.; Jia, Y.; Porter, Z.; Habl, I. Artificial intelligence explainability: the technical and ethical dimensions, Phil. Trans. R. Soc. A., 2021, 379, 20200363.

[21] Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; Müller K. -R. (Eds.). *Explainable AI: Interpreting Explaining and Visualizing Deep Learning*, LNAI, 17000, Springer, 2019.

[22] Bau, D.; Zhu, J.-Y.; Strobelt, H.; Lapedriza, A.; Zhou, B.; Torralba A. Understanding the role of individual units in a deep neural network, Proc. Nat. Acad. Sci. USA, 2020, 117(48), 30071-30078.

[23] Phillips, P.J.et al. Four Principles of Explainable Artificial Intelligence, Natl. Inst. Stand. Technol. Interag. Intern., Rep. 8312, 2021.

[24] Vassiliades, A.; Bassiliades, N.; Patkos, T. *Argumentation and explainable artificial intelligence: a survey*, Cambridge Univ. Press, 2021.

[25] Wilson, A. G.; Dann, C.; Lucas, C. G.; Xing, E. P. The human kernel, in *Advances in Neural Information Processing Systems* (Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; Garnett, R.; Eds.), 2015, 28.

[26] Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable Machine Learning, 2017, arXiv:1702.08608v2

[27] Bottou, L. From machine learning to machine reasoning, Mach. Learn., 2014, 94, 133-149.

[28] Scassellati, B. Theory of mind for a humanoid robot, Autonomous Robots, 2002, 12, 13-24.

[29] Goldman, A. Theory of mind, in *Oxford Handbook of Philosophy and Cognitive Science*, (Margolis, E.; Samuels, R.; Stich, S. Eds.), Oxford Univ. Press, 2012.

[30] Baron-Cohen, S.; Leslie, A.M.; Frith, U. Does the autistic child have a "theory of mind"?, Cognition, 1985, 21, 37-46.

[31] Wellman, H.W.; Cross, D.; Watson, J. Meta-analysis of theory-of-mind development: The truth about false belief, Child Development, 2001, 72, 655-684.

[32] Zeman, A. Consciousness, Brain, 2001, 124, 1263-1289.

[33] *The Oxford Companion to Consciousness* (Bayne, T.; Cleeremans, A.; Wilken, P. Eds.), Oxford Univ. Press, 2009.

[34] Aleksander, I.; Morton, H. Phenomenology and digital neural architectures, Neural Networks, 2007, 20, 932-937.

[35] Sun, R.; Franklin, S. Computational models of consciousness, in *Cambridge Handbook of Consciousness* (Zelazo, P.; Moscovitch, M. Eds.), Cambridge University Press, 2007, 151-174.

[36] Gamez, D. Progress in machine consciousness, Consciousness and Cognition, 2008, 17, 887-910.

[37] Searle, J. Minds, brains, and programs, Behav. Brain Sci., 1980, 3(3), 417-424.

[38] Harnad, S. The symbol grounding problem, Physica D: Nonlinear Phenomena, 1990, 42, 1-3, 335-346.

[39] Tononi, G. An information integration theory of consciousness, BMC Neuroscience, 2004, 5, 42.