Consciousness and Understanding in Autonomous Systems

Ricardo Sanz[1] and Julita Bermejo-Alonso[2]

[1] Autonomous Systems Laboratory
Universidad Politécnica de Madrid, 28006 Madrid, Spain
ricardo.sanz@upm.es
[2] Universidad Internacional Isabel I,
Fernán González 76, 09003 Burgos, Spain
julita.bermejo@ui1.es

Abstract. This position paper will highlight the importance of having a formal notion of understanding as one of the cornerstones in the construction of conscious AIs. It will show that the capability of understanding both the perceptual and the action flows is critical for the correct operation of situated autonomous systems. An assessment is also made on the contribution of the machine learning domain towards this direction.

Keywords: Artificial intelligence · Awareness · Consciousness · Understanding · Autonomy · Machine Learning

"What I cannot create I do not understand" – Richard Feynman, 1988

1 Introduction

Autonomy —the capability of an agent to properly act by itself in a changing, uncertain world— seems to requiere consciousness. Just consider the quality of your autonomous behaviour when you are more or less conscious. This capability is equally needed for machines [27]. Graziano [14] says that *"Artificial intelligence is growing more intelligent every year, but we've never given our machines consciousness."*. Is this true? Have we ever given consciousness to our machines?. The answer to this question depends on what we consider consciousness to be [32]. Proper definitions are needed to ground researcher collaboration and enable theory selection and consolidation.

Chella and Manzotti [5] describe machine consciousness as *"the attempt to model and implement aspects of human cognition that are identified with the elusive and controversial phenomenon of consciousness"*. They also state [6] that *"the main goals that artificial consciousness should achieve: autonomy and resilience, information integration, semantic capabilities, intentionality, and self-motivations."* This vision of machine consciousness is fully aligned with the idea of AI as machine reproduction of human mental capability. In this human-imitation sense, we engineers are obviously quite far from our machines being conscious as we are.

However, human imitation may not be the right path to properly understand the phenomenon of consciousness. Irvine [17] argues that the retention of the concept of ”consciousness” is an impediment to further progress in the cognitive sciences. Being this analysis made from a neuroscience perspective, can it also be valid for engineered machines? It is our conviction, that if we aim to achieve sound progress in conscious AIs, we should go beyond human-centered approaches, to focus on more intrinsic, functional and architectural properties of general *conscious systems*.

The traps of focusing on the human mind are manifold. For example, the good, old *what-is-it-like-to-be* approach to phenomenal consciousness lacks the necessary clarity to serve as foundation to verify and validate engineered AIs as required by systematic engineering practices. We shall concentrate on aspects that are both i) precisely definable —*i.e.* in formal terms— and ii) verifiable by experimentation. Note that *all* results on phenomenal consciousness in humans are based on *subjective* verbal report of the subjects. Objectivity in consciousness research is definitely elusive and the lack of agreement on the very idea of consciousness is a major barrier. Sommerhoff [33] says that *“A precise definition of the word [consciousness], of course, can only be the end-point of a theory of consciousness, just as the concepts of work and energy found a precise definition only as part of a theory of mechanics.”* Maybe it is necessary to formally address more basic aspects of cognitive systems before reaching agreement on AI consciousness.

In our opinion, among the plethora of phenomena around consciousness, there are two key elements for autonomous systems engineering practice: i) the capability of *perceiving* and ii) the capability of *understanding* what has been perceived to achieve the situational awareness that grounds the capability to act meaningfully.

Developing a precise, general, accepted, definition of these two capabilities —esp. of understanding— may become a daunting task. The case of perception is clearer and good proposals on how to define it abound [19]. The disparities among scholars are more related with the perceptual process boundaries —where it starts, where it ends— than with its nature. The concept of understanding is much trickier, however.

“Understanding” is a very elusive concept. It has been a usual topic in epistemology, but it has been amply displaced by the study of *knowledge*. A similar phenomenon has happened in AI technology. The interest in understanding is, however, re-gaining force in all domains[3]. Some may believe that a concept of understanding common to humans and machines is still a dream. There are nevertheless green sprouts in this direction. For example, the position of Newton [21] seems quite close to the needs of engineers: *”the intentionality of a men-*

[3] As demonstrate by the recent DARPA call on systems with common sense where the capability of understanding is seen as critical.

tal state, considered as a response to an environmental stimulus, consists in the understanding the subject has of that stimulus and of her goals in responding to it, just as the intentionality (the meaning) of a fragment of physical behaviour consists in its being part of a goal- directed action, understood by the agent."

To build machines based on a theory of perception and understanding, we point out the necessity to establish clear conceptualisations and definitions of both. Only with this approach, we believe it could be possible to achieve the constructability and verifiability required in the engineering and deployment of real-world AIs. These definitions may then evolve based on the demonstrated results of the implementations to ground definite theories of consciousness.

2 Consciousness and understanding in the cognitive cycle

In our opinion, the core evolutionary and functional value of consciousness is related to the provision of situational awareness to the perceiving and acting agent. Conscious agents know better what is going on. This enables proper actuation.

Situational awareness is defined as *"the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future"* [10]. As was said before, perception and understanding of what has been perceived are the foundation of good situated action.

The achievement of adequate *situational awareness* is decisive in autonomous systems, such as animals, humans, machines or groups of any of them. This need for awareness to act properly is concordant with many existing approaches to the analysis and modelling of the cognitive action cycle (OODA, MAPE, PDCA, *etc.*). In this vein, we can say that the four key elements for a functional autonomous system are:

- The capability of perceiving.
- The capability of understanding.
- The capability of reasoning.
- The capability of acting.

Traditional GOFAI was centred in reasoning. Situated AI deals with the coupling of perceiving and acting —ignoring the in-between in an Skinnerian way. All this research has produced very valuable but non-scalable results to full-fledged autonomous systems: our current autonomous systems still lack the necessary understanding[4].

The place that *consciousness* play in this picture has been unclear [34, 3]. Most human-based theories of consciousness are merely philosophical and too abstract and essentially detached from realisations —realisations that are needed in AI. In other cases, when the theories are closer to realisations (*e.g.* neural in the case of humans, software in the case of machines) [30, 8] there is a problem

[4] Or to be more precise, their level of understanding is limited and strictly tied to specific mechanisms of action.

of non-generality —*i.e.* they are not properly addressing multiple realisability— or of causal opacity in an architectural sense. These theories cannot indeed be positively used as general engineering assets.

In this paper, we suggest that consciousness is the net effect of the *appropriate coupling of perception and understanding* —including self-perception. As we will see later, this is related to goals and value —of the agent, their mates or their masters.

3 The nature of understanding

In a recent essay, Baumberger *et al.* [1] address the question of what is understanding from the perspective of epistemologists. They mostly discuss around two types of understanding: "explanatory" understanding of why something is as it is and "objectual" understanding of a domain. The discussion is essentially metaphysical —e.g. conditions for the existence of understanding— and close to the issues in philosophy of science and far from engineering needs.

In the context of this paper, we focus however on the more earthly issue of how an agent can understand what it perceives. A feat that we could call signal understanding. The motivation is clear: AI failures are sometimes coupled to improper understanding of what was going on. This is not new at all. The old thread in AI on *common sense* is just a manifestation of the need of developing AIs that understand.

A better understanding of the situation becomes critical for autonomous systems. Just consider the casualties caused by autonomous car driving systems that have raised public awareness on this matter. DARPA's MCS —Machines with Common Sense— future program argues that common sense is the basic ability to perceive and understand the world:

> *"Today's machine learning systems are more advanced than ever, capable of automating increasingly complex tasks and serving as a critical tool for human operators. Despite recent advances, however, a critical component of Artificial Intelligence (AI) remains just out of reach — machine common sense."*

The focus on the elusive idea of common sense hinders the problem mentioned before: the too anthropomorphic conception of most AI research. The description of what is sought for AIs —common sense— is meaningful for psychologists or sociologists or the layman, but far from being mechanisable by programmers in the implementation of AIs. We obviously need common sense in the machines and to achieve this we must endow them with the capability of understanding what is happening and what are the consequences of their actions.

What is missing in the current state of affairs is:

- A formal theory of understanding that is scientific, effective and widely accepted.
- A reference architecture for understanding that can be shared and reused.

- Domain-specific architecture instantiations driven by well-defined requirements (e.g. following a real systems engineering process [18]).

In our opinion, the most promising proposals for a theory of understanding depend on the agent having a world model that is deeply tuned to reality and used in action generation [7, 24, 22, 35]. Notwithstanding past developments in this direction, an increased, sustained and collaborative effort is required to advance, disseminate and consolidate them.

The core idea that we want to defend in this paper is strongly representationalist: agents keep models of their worlds in their heads and use them to decide what to do [29]. Minds are model-based controllers [5]. Consciousness is the functional state, the action and the effect of keeping those models updated, tuned to reality [28]. In this picture, some aspects of learning shall indeed be considered as the slow part of consciousness.

The value of models comes from their *actionable nature*. Models can be exercised to provide different classes of information. As [36] says *"models are the highly specialized part of our technological equipment whose specific function it is to create the future."*

In this model-based picture of minds, the nature of understanding is clear: a sensory signal is understood when the information it carries is properly integrated in the mental model of the agent. Note that "properly" means in strict accordance to the architecture, goals and values of the agent.

4 Theories of understanding

This idea of what is understanding must be developed into a solid theory of understanding to be able to systematically implement mechanisms for consciousness in real-world autonomous machines.

Understanding in classic AI has been associated with the generation and use of knowledge [20], mostly in a propositional form. However, we shall go beyond propositional accounts of understanding (and hence also beyond propositional accounts of knowledge) to address more general autonomy problems [25, 15]. While epistemologists are dwelling in the post-Gettier analysis of knowledge, the theory of general knowledge as applicable to AIs has not advanced much more beyond Newell's knowledge level.

Philosophy has been dealing with this issue. Nevertheless, most philosophical theories are not precise nor positive enough. Deiss [9] defines consciousness as a process of *interpreting sensations* —i.e. finding *meaning* in sensory flow. He considers that meaning resides in the expectations and predictions attached to qualitative sensory contrasts using brain's associative memory. Saying that meaning "resides in" is too vague to be useful.

[5] Obviously, some simple control loops do not requiere full fledged models to operate; nevertheless the controller shall somehow capture the dynamics of the controlled system [7].

Engineering-grade theories shall be intelligible, realisable and actionable [31]. De Regt proposes a *Criterion for the Intelligibility of Theories* (CIT_1) as one way of testing the intelligibility of scientific theories by other scientists [23, p.102]:

> CIT_1: A scientific theory T (in one or more of its representations) is intelligible for scientists (in context C) if they can recognize qualitatively characteristic consequences of T without performing exact calculations.

De Regt's criterion intends objectivity and sufficiency for mathematically expressed theories. But is is specially interesting because it requests qualitative exercisability of the theory, a close encounter with the model of understanding proposed here.

Physics provides, in this sense, the better example of understanding. Feynman [12], considered this question in his lectures: *"What do we mean by 'understanding' something? . . . If we know the rules, we consider that we 'understand' the world."*

Feynmann [12] associates physical understanding of the behavior of a system with having *"some feel for the character of the solution in different circumstances."* He adds: *"So if we have a way of knowing what should happen in given circumstances without actually solving the equations, then we "understand" the equation, as applied to these circumstances. A physical understanding is a completely unmathematical, imprecise, and inexact thing, but absolutely necessary for a physicist."*.

In the same vein, Chaitin [4] proposes the idea that comprehension is based on data compression; that understanding something (data) means being able to figure out a simple set of rules —a model— that explains it (that explains how the the available data is produced).

There are not many definitions of *meaning* beyond linguistics. Gelepithis [13], in the context of a theory of consciousness says that the meaning of a previously encountered *stimulus within its context*, for a *human* at certain *time*, is the prevailed neural formation that can affect its attention. So, meanings are neural formations that strongly affect attention.

Thorisson *et al.* [35] provide a formal definition of *understanding* in the context of model-based cognitive agents. An agent's understanding of a phenomenon depends on the accuracy of the model that the agent has with respect to the phenomenon. Understanding is hence a multidimensional matter of degree determined by the adequacy of the model to the phenomenon in two aspects: completeness and accuracy. This model-centric view is, in essence, akin to the formal models behind modelling and simulation [38, 37]. In the same vein, Thorisson *et al.* [35] provide a definition of *meaning* of a datum for an agent, that is captured by the set of relevant implications of the datum in relation to a concrete set of goals of the agent and the knowledge that the agent has in that in situation.

However, concerning the specific issue of understanding, Thorisson's definitions transpose the capability of exercising the models to the computations of implications — that the authors capture under the idea of testing for understanding in four dimensions: predict, achieve, explain, (re)create.

Our work in this domain [26, 28, 16, 2] orbits around the model-integration theory of understanding in autonomous systems. Autonomous systems generate meanings from data (typically from sensory inputs) and use their continuously updated mental models to control their behavior. Understanding a piece of information gathered from the sensors implies its integration into the model that captures the agent's knowledge. This theory goes in line with the analysis done by Thorisson *et al.* [35] but departs from it in two aspects:

- The actionable nature of the model. The model is causally complete; it can be executed to provide the capabilities associated to the agent cognitive powers (e.g. prediction, control or explanation).
- The definition of meaning. We depart from Thorisson in the interpretation of meanings as the exercisable content of models.

Autonomous behavior is a tricky issue, esp. in relation to autonomous systems. Note that behavior is generated to provide *value* to i) the agent or ii) to the owner —sometimes not the same thing as Asimov aptly noticed. The autonomous artificial system needs understanding of the significance of perceptual elements in light of agent's or owner's goals. Meanings —the exercisable content of world models— are used to determine equivalence classes of agent+world trajectories in state-space in relation with agent?s value system (projections into the future including counter-factuals).

5 Does machine learning create understanding?

Machine learning is, in principle, the substrate of the ultimate form of awareness: the capability of understanding anything. We should avoid, however, the current hype on deep learning and similar mechanisms. The fact that correlation is not causation underlies many of the problems that these technologies show. The models they create and use can mimic certain datasets but cannot be extended beyond them because their causal structure is not necessarily isomorphic to that of the reality generating the data. While more robust that rule-based systems in many contexts, neural network leaners still suffer the cliff effect.

Models created by learners are actionable, but only in the specific context and use where they were learnt. For example, a learnt model for condition maintenance of a machine can work well predicting its failure but can be useless in diagnosing the causes.

Besides this, many learnt models are unshareable due to their opacity. The Explainable Artificial Intelligence (XAI) DARPA program shows the generalised awareness of this opacity problem

6 Conclusions

Perception and understanding are central issues for consciousness — both in humans and in AIs. While a lot of research effort has been devoted to perception, the same can not be said of the one dedicated to understanding.

Past research on AIs that understand has mainly focus on how specific sensory input is understood by the autonomous system —language understanding, image understanding—, and addressed it as mere syntactic parsing. This can be considered just a perceptual process, maybe necessary but previous to understanding. Understanding has not been addressed from a general ample viewpoint, but just as specialised mechanisms to deal with specific classes of problems and sensor flows. Only some teams have addressed the general problem [35].

The approach defended in this paper considers understanding as a process of integrating perception into *actionable* models. These models are then used by the agent to compute actions that make sense; that provide value for the agent and/or the owner.

Some aspects of human consciousness have not contributed much to addressing the problem of conscious AIs. The question of qualia and phenomenal experience in general is a red herring [11]. We shall be aware of this. Awareness —including self-awareness— is the critical asset for building autonomous machines.

”It is impossible to separate awareness, consciousness and understanding”
– Jacques Lacombe, 2003

References

1. Baumberger, C., Beisbart, C., Brun, G.: What is understanding? an overview of recent debates in epistemology and philosophy of science. In: Grimm, S.R., Baumberger, C., Ammon, S. (eds.) Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science, chap. 1, pp. 1–34. Routledge (2017)
2. Bermejo-Alonso, J., Hernández, C., Sanz, R.: Model-based engineering of autonomous systems using ontologies and metamodels. In: IEEE International Symposium on Systems Engineering 2016 (IEEE ISSE 2016). Edinburgh, Scotland (2016)
3. Cavanna, A.E., Nani, A.: Consciousness. Theories in Neuroscience and Philosophy of Mind. Springer-Verlag, Berlin Heidelberg (2014)
4. Chaitin, G.: The limits of reason. Scientific American pp. 74–81 (March 2006)
5. Chella, A., Manzotti, R.: Artificial consciousness. In: Cutsuridis, V., Hussain, A., Taylor, J.G. (eds.) Perception-Action Cycle. Models, Architectures, and Hardware, pp. 637–674. Springer, New York (2011)
6. Chella, A., Manzotti, R.: Artificial consciousness. In: et al., V.C. (ed.) Perception-Action Cycle: Models, Architectures, and Hardware,, Springer Series in Cognitive and Neural Systems, vol. 1. Springer (2011)
7. Conant, R.C., Ashby, W.R.: Every good regulator of a system must be a model of that system. International Journal of Systems Science **1**(2), 89–97 (1970)
8. Dehaene, S., Charles, L., King, J.R., Marti, S.: Toward a computational theory of conscious processing. Current Opinion in Neurobiology **25**, 76 – 84 (2014). https://doi.org/10.1016/j.conb.2013.12.005
9. Deiss, S.: Universal correlates of consciousness. In: Skrbina, D. (ed.) Mind That Abides. Panpsychism in the new millennium, chap. 7, pp. 137–158. John Benjamins Publishing Company (2009)

10. Endsley, M.R.: Towards a theory of situation awareness in dynamic systems. Human Factors **37**(11), 32–64 (1995)
11. Fekete, T., Edelman, S.: Towards a computational theory of experience. Consciousness and Cognition **20**(3), 807–827 (2011)
12. Feynmann, R., Leighton, R., Sands, M.: The Feynmann Lectures on Physics. New Millenium Edition. Volume II. Mainly Electromagnetism and Matter. Basic Books, New York (2010)
13. Gelepithis, P.A.M.: A novel theory of consciousness. International Journal of Machine Consciousness **6**(2), 125–139 (2014)
14. Graziano, M.: Build-a-brain — can we make consciousness into an engineering problem. Online: https://aeon.co/essays/can-we-make-consciousness-into-an-engineering-problem (July 2015), accessed: 10/11/2018
15. Hayes, P.J.: The second naive physics manifesto. In: Hobbs, J.R., Moore, R.C. (eds.) Formal Theories of the Commonsense World, pp. 1–36. Ablex, Norwood, NJ (1985)
16. Hernández, C., Bermejo-Alonso, J., Sanz, R.: A self-adaptation framework based on functional knowledge for augmented autonomy in robots. Integrated Computer-Aided Engineering **25**, 157–172 (2018)
17. Irvine, E.: Consciousness as a Scientific Concept. A Philosophy of Science Perspective, Studies in Braina and Mind, vol. 5. Springer (2012)
18. ISO/IEC/IEEE: ISO/IEC/IEEE 15288-2015 Systems and software engineering – System life cycle processes. International standard, International Standards Organisation (2015)
19. López, I.: A Framework for Perception in Autonomous Systems. Ph.D. thesis, Departamento de Automática, Universidad Politécnica de Madrid (May 2007)
20. Newell, A.: The knowledge level. Artificial Intelligence **18**, 87–127 (1982)
21. Newton, N.: Foundations of Understanding, Advances in Consciousness Research, vol. 10. John Benjamins Publishing Company (1996)
22. Newton, N.W.: Understanding and self-organization. Frontiers in Systems Neuroscience **11**(8) (2017). https://doi.org/10.3389/fnsys.2017.00008
23. de Regt, H.W.: Understanding Scientific Understanding. Oxford University Press (2017)
24. Rosen, R.: Anticipatory Systems. Philosophical, Mathematical, and Methodological Foundations, IFSR International Series on Systems Science and Engineering, vol. 1. Springer, 2nd edn. (2012)
25. Sandewall, E.: Features and Fluents. The Representation of Knowledge about Dynamical Systems. Clarendon Press, Oxford (1994)
26. Sanz, R.: Engineering conscious machines. In: Models of Consciousness ESF/PESC Exploratory Workshop. Birmingham, UK (September 1-3 2003)
27. Sanz, R., López, I., Bermejo-Alonso, J.: A rationale and vision for machine consciousness in complex controllers. In: Chella, A., Manzotti, R. (eds.) Artificial Consciousness, pp. 141–155. Imprint Academic (2007)
28. Sanz, R., López, I., Rodríguez, M., Hernández, C.: Principles for consciousness in integrated cognitive control. Neural Networks **20**(9), 938–946 (2007)
29. Sanz, R., Meystel, A.: Modeling, self and consciousness: Further perspectives of AI research. In: Proceedings of PerMIS '02, Performance Metrics for Intelligent Systems Workshop. Gaithersburg (MD), USA (August 13-15 2002)
30. Shanahan, M.: Embodiment and the inner life: Cognition and Consciousness in the Space of Possible Minds. Oxford University Press (2010)

31. Sloman, A.: What enables a machine to understand? In: Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 2. pp. 995–1001. IJCAI'85, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1985)
32. Sloman, A.: An alternative to working on machine consciousness. International Journal of Machine Consciousness **2**(1), 1–18 (2010)
33. Sommerhoff, G.: Consciousness explained as an internal integrating system. Journal of Consciousness Studies **3**(2), 139–157 (1996)
34. Taylor, J.G.: The Race for Consciousness. MIT Press (1999)
35. Thórisson, K.R., Kremelberg, D., Steunebrink, B.R., Nivel, E.: About understanding. In: Steunebrink, B., Wang, P., Goertzel, B. (eds.) Artificial General Intelligence. Proceedings of the 9th Conference on Artificial General Intelligence (AGI 2016). pp. 106–117 (2016)
36. Wartofsky, M.W.: Telos and technique: Models as modes of action. In: Models: Representation and the Scientific Understanding, chap. 8, pp. 140–153. Springer Netherlands, Dordrecht (1979). https://doi.org/10.1007/978-94-009-9357-0
37. Zeigler, B., Muzy, A., Kofman, E.: Theory of Modeling and Simulation. Discrete Event & Iterative System Computational Foundations. Academic Press, 3rd edn. (2018)
38. Zeigler, B.P.: Toward a formal theory of modeling and simulation: Structure preserving morphisms. J. ACM **19**(4), 742–764 (Oct 1972)