On the Consciousness of Intelligent Systems

Bernadette Bouchon-Meunier bernadette.bouchon-meunier@lip6.fr
Sorbonne Université, CNRS, LIP6, BP 169
4 place Jussieu
75005 Paris, France

Corresponding Author: Bernadette Bouchon-Meunier

Abstract

The consciousness of intelligent systems has been discussed for decades. The ongoing developments of Artificial Intelligence make this concept even more crucial. We review and discuss recent views of the consciousness of artificial systems, mainly based on theories of the consciousness of human beings taking into account phenomenal consciousness, access consciousness and computational functionalism. Distinguishing simulated and natural consciousness, we point out the importance of interpretable and eXplainable Artificial Intelligence to approach the first one. We focus on Large Language Models and Robotics. We also present discussions on the attribution of a moral status to artificial systems. We finally propose to use a fuzzy set-based knowledge representation to approach some forms of consciousness in intelligent systems, following existing directions of research on fuzzy concepts lattices, fuzzy ontologies and fuzzy description logic, as well as the use of fuzzy methods to represent and manage emotions and perceptions, in particular to identify and simulate sentiments in textual or facial expressions.

Keywords: Artificial systems, Artificial beings, Artificial intelligence, Consciousness, Large language models, Robots, Moral status, Fuzzy sets

1. INTRODUCTION

The consciousness of artificial systems has long been discussed, but it has never been more significant than it is today, due to the development of artificial systems collaborating ever more closely with humans, dialoguing with them in a very natural way and acting more and more like humans. The attention of the readers has been drawn in the past years to the interest of analyzing the possibility for artificial beings to be conscious [1, 2].

The concept of consciousness is complex and takes various forms. It has several facets studied in philosophy, neuroscience, cognitive science in particular. It has been tackled for decades in relation with Artificial Intelligence. After the seminal book published by Isaac Asimov in 1950 [3], we can think, for example, of Hubert Dreyfus [4] who was considering in 1972 that the intelligence of artificial systems is limited and that the advances made at this time on artificial systems were more cognitive simulation than real artificial intelligence, claiming that humans have certain capacities

Citation: Bernadette Bouchon-Meunier. On the Consciousness of Intelligent Systems. **Advances in Artificial Intelligence and Machine Learning**. 2026;6(4):335. https://dx.doi.org/10.54364/AAIML.2026.64335

involved in intelligent behavior which cannot be programmed. On the contrary, Marvin Minsky [5] explained in 1982 that conscious artificial systems could be done by providing machines with ways to examine their own mechanisms while they are working. It is also in this direction that Jacques Pitrat claimed in his book [6] published in 2009 that what he called artificial beings can be considered conscious when they have the possibility of monitoring their own thoughts in such a way that they can explain their decisions to others. This approach goes in the very topical direction of interpretable and eXplainable Artificial Intelligence, clearly fundamental for the consciousness of intelligent systems, especially in the case of simulated consciousness.

Many articles have been published so far, discussing the possibility of artificial systems to be conscious. For instance, the authors of [7] refer to information processing computations in the brain, independently of any subjective experience, to propose a form of consciousness enabling artificial systems to behave as if they were conscious. More recently, Ilya Sutskever [8], OpenAI research group chief scientist, tweeted that “It may be that today’s large neural networks are slightly conscious”.

More than 20 theories supporting the notion of consciousness have been proposed so far [9]. They identify neural, physical or functional properties showing that an organism is conscious or not. These theories are mainly based on what is identified as the consciousness of human beings and in particular on the identification of “what it is like to be in a given state”; some of the theories take other organisms into consideration [10] and a recent study [11]provided a list of indicator properties derived from some of these scientific theories to assess consciousness for an intelligent system.

Discussing about the consciousness in Artificial Intelligence is a topical question, first because of the general interest in eXplainable AI, requiring the system to provide explanations about its functioning and decisions, according both to the DARPA incentive [12] and to the successive guidelines of the European Commission requiring the transparency of data and models in Artificial Intelligence, as well as appropriate explanations on the functioning and the decisions provided to stakeholders. The second reason for this question to be topical is the recent development of Generative AI which produces systems capable of dialoguing with humans in a very natural way, seeming to express themselves in a way analogous to humans, and capable to explain their reasoning to some extent. Debates about the danger of future Generative AI able to feel emotions and to be self-aware sheds light on the question of the consciousness of artificial systems.

We can wonder whether talking about the consciousness of artificial systems is anthropomorphism. In our view, it is not more anthropomorphism than talking about intelligence for such systems, which has been explicitly done since the Dartmouth conference in 1956.

We remark that a plane does not fly like a bird and therefore we think that an artificial system can have a form of intelligence and consciousness different from those of a human being.

In this paper, we propose to look at several facets of consciousness, from phenomenological consciousness linked to perceptions to access consciousness, which gives us information about our actions. We offer an overview of some aspects of consciousness from the angle of intelligent systems, and we wonder to what extent a present or a future system can have such a form of consciousness and what the advantages and drawbacks are.

In section II, we first consider the main facets of consciousness, principally the phenomenal and access properties of consciousness. We then address the proposal of computational functionalism presented by a group of researchers, based on criteria enabling to decide if a being is conscious or not, stemming from neuroscience-based theories of human consciousness. We focus on Large Language Models which are currently the most visible artificial systems in section III, and on robots that appear to be good candidates for consciousness, given their visual similarity and their interactions with human beings in section IV. In section V, we approach the question of the moral status of artificial systems, be they conscious or not. In section VI, we end this paper by proposals to use fuzzy set-based methods and concepts to approach the creation or identification of conscious artificial systems. We conclude in section VII.

This paper is not intended to cover all aspects of the consciousness of living beings or all proposals to consider conscious artificial beings. It is an overview of recent approaches of the consciousness of artificial systems which may lead to developments based on computational intelligence.

2. ASPECTS OF CONSCIOUSNESS

2.1 Intelligence and Consciousness

Various ways of approaching the concept of consciousness of artificial systems can be thought of. We can for instance base our thinking on the intelligence and consciousness of all living beings, or just human beings.

We can also consider that the consciousness of intelligent systems is either a simulated/artificial element or an element analogous to the consciousness of living/human beings. The first case is sometimes referred to as weak artificial consciousness, enabling us to create highly efficient systems and robots. The second case is identified as strong artificial consciousness, allowing interaction between biology and robotics, the study of biological consciousness helping to build better robots, while the research on robot consciousness contributes to the understanding of biological consciousness [13].

In [14], Antonio Damasio identifies levels of intelligence leading to consciousness for living beings, and we can ask ourselves about the transposition of this view to artificial beings.

In a nutshell, the first level of intelligence, available to very simple living beings like bacteria, is perception through senses (vision, hearing, touch…), providing implicit information. This level is easily approached by artificial systems by means of sensors, cameras and electronic sensing devices.

The second level of intelligence is mind, achieved by means of a nervous system. It involves the creation of representations, patterns and images resulting from perception. It is strongly connected to reasoning, language and memory, which artificial systems are also able to use. It provides explicit knowledge.

The third level of intelligence is consciousness, involving feelings and the concept of self-perspective. It is available to various kinds of animals such as fishes, birds, mammals, and of course human beings. In the case of artificial systems, we can easily think of simulated feelings already familiar to robots, for instance. Self-perspective is approached by systems able to know and to explain how they make a decision. The question is to know if machines can be really sensitive and in addition conscious.

2.2 Phenomenal and Access Consciousness

In his seminal paper [15], Ned Block distinguishes several aspects of consciousness: self-consciousness linked to the concept of self, monitoring consciousness referring to a cognitive approach of consciousness, access consciousness and phenomenal consciousness, which are two fundamental interacting aspects of consciousness now at the core of many studies of the consciousness of artificial systems, on which we focus.

The first one, phenomenal consciousness, is linked to subjective experiences and the experience of "what it is like to be in a given state". Its properties include experiential properties of sensations, feelings, perceptions, thoughts, desires, emotions. It is related with conscious attention, associated with what we can report of what we can see, hear or feel. It is mainly qualitative. It is neither associated with any cognitive property, involving thought, nor any functional or programmable property. Phenomenally conscious contents could be representational, but they are not necessarily. This aspect of consciousness can be associated with phenomenology, as introduced by Edmund Husserl and developed by his successors [16], regarded as a philosophical study of the way we experience things and the consideration of subjective experiences.

The second one, access consciousness, is representational. It implies that a representation of the content of an access conscious state is available as a premise in reasoning or for rational control of action and of speech if it exists. A piece of information is access conscious to an agent if he can use it to make a decision or for planning.

Jacques Pitrat, who was a pioneer in Strong Artificial Intelligence or Artificial General Intelligence, was working on an artificial scientist, CAIA, that he wanted at least as efficient as a human one. He thought that the only way to realize such an artificial system was to use a bootstrap and to produce successive systems, each one helping to generate the next one, in collaboration with the human operator for the time being or alone in the future [6].

He mentioned these two aspects of consciousness that he considered reachable by what he called artificial beings, as pointed out by G. Sabah [17]. In this book, he first considered "phenomenological" consciousness, associated with sensations received from the external world and perceptions resulting from their processing, leading to actions. He says that artificial beings have this form of consciousness, since they process the information they receive through artificial devices such as keyboards or cameras from the external world. What he considers is clearly an artificial form of phenomenal consciousness, not related with feelings or emotions, for instance.

He then turns his attention to access consciousness, and he claims that artificial beings can observe their present state and their processes while they work better than human beings. This enables them to understand the reasons of their successes or failures and to correct them. It also enables them to explain their decisions and the way they obtain them, so that other beings can accept these decisions. Artificial beings can choose to observe certain parts of their processes, in what Jacques Pitrat calls "adjustable consciousness" which gives them an advantage for learning, monitoring and explaining.

In order to provide a logical way to decide if a system is conscious on the basis of axioms, a mathematical model of consciousness has been proposed by Giulio Tononi [18], describing fundamental properties of consciousness and he postulates properties that a system must have to account for experience. This model, called "Information integrated Theory", has given rise to controversies, arguments to support it or to criticize it being provided.

David Chalmers considers [19] that a being is conscious if there is something it is like to be that being, and a mental state is conscious if it has a qualitative feel, associated with phenomenal qualities. He describes the so-called hard problem, to explain why and how living beings can have subjective experience and phenomenal consciousness. In particular, he considers [20] that G. Tononi's proposal [18] is in line with the hard problem.

2.3 Computational Functionalism

Computational functionalism considers the hypothesis that performing appropriate computations is a necessary or sufficient condition for consciousness in artificial beings [11]. It is based on neuroscience-based theories of human consciousness which focus on various computational features considered as necessary or sufficient for phenomenal consciousness in humans.

Functions identified in these theories are analyzed in the framework of present or future artificial systems and indicators of consciousness are proposed, on the basis of how these systems process information at the algorithmic and representational levels. This study mentions a debate, also discussed by other authors, about the possibility for a conscious system to suffer if we don't take its consciousness into account. Another discussion concerns the human tendency to attribute intentions and emotions to artificial systems and then to anthropomorphize them.

Thus, we review some of the various theories of consciousness leading to indicators of human consciousness presented by P. Butlin et al. [11], focusing on aspects related to computational intelligence.

Recurrent processing theory is mainly focused on functions involved in vision. The first indicator of consciousness in this theory is the existence of input modules using algorithmic recurrence, referring to neurons receiving inputs produced from their previous outputs. Many deep learning models satisfy this indicator. The second indicator is the existence of input modules generating organized, integrated perceptual representations. This indicator is satisfied by deep convolutional neural networks. But it must be observed that they are generally more efficient to identify local properties than global representations of objects taking their relations into account, involved in consciousness. Some systems, for instance [21–24], are nevertheless able to represent organized perceptual scenes.

Global Workspace Theory deals with access consciousness and considers that many living beings, including humans, use specialized systems operating in parallel to perform tasks and these modules share information through a common workspace. States are conscious when they are represented in this workspace. The existence of such modules is considered an indicator of consciousness. Another indicator is the existence of a limited capacity for the workspace, requiring the information provided by the modules to be selected by an attention mechanism. There exist many artificial systems satisfying these indicators.

Higher-order theories considers that, if a mental state is conscious, it is necessary to be aware of being in this mental state and then to call on higher-order representations, which are representations about other representations, or meta-representations. Indicators of consciousness refer to the availability of several sources of perception and, in particular to the presence of generative, top-down or noisy perception modules and the metacognitive capability to distinguish reliable perceptual representations from noise. There are presently no artificial systems with such capacities but it is possible that deep learning-based systems provide such solutions to these expectations.

Attention schema theory concerns the representation of facts about the current object of attention. It requires the existence of a predictive model representing and enabling control over the current state of attention. Models of attention are at the heart of transformer models currently used by Large Language Models, but they are very different from the human concept of attention, which they simulate to some extent.

Several points can be discussed about computational functionalism. The first one is the tendency we have to over-anthropomorphize artificial systems, to think that they really have intentions and emotions in the same way as human beings. In particular the possibility for a conscious system to suffer is mentioned, but also the possibility that a conscious artificial system has no valenced or affective experiences.

In other words, we should not overestimate or underestimate the consciousness of artificial systems and we must keep in mind the possibility to create very efficient artificial systems which are not conscious, or which only possess some of the features constituting human consciousness according to the neuroscience-based theories. To come back to the analogy with the flight of airplanes, birds or insects, it is clear that airplanes don't fly in the same way as animals, and all flying animals don't have the same way to fly, even though there is a similarity between their ability to fly. It is then possible that some conscious artificial systems have a form of consciousness which is different from the human one.

For the time being, the conclusion [11] is that no current artificial intelligence-based systems have all components of human consciousness. There is no reason to think that future such systems will never satisfy all the criteria of human consciousness.

2.4 Tests for Consciousness

Along the same lines, an overview of so-called C-tests [10] proposes a general framework for tests for consciousness, pointing out the importance of a moral treatment of all beings with phenomenal consciousness or self-consciousness.

Besides various tests of the consciousness of living beings, this framework is mainly focused on phenomenal consciousness and covers the two sides of the question, to know if an organism is either able to be conscious or presently conscious.

3. LARGE LANGUAGE MODELS AND CONSCIOUSNESS

The overwhelming presence of Large Language Models in the Artificial Intelligence framework leads to consider these particular artificial systems and to wonder if they can be conscious and what would be the consequences of their possible consciousness.

A study was conducted in 2023 [25] to determine if people are inclined to think that Large Language Models can be conscious and, in particular, if they can have phenomenal consciousness. A majority of the participants were thinking that it was possible for ChatGPT, considered for the study, to have some phenomenal consciousness. There is no doubt that the developments of this system since then would not contradict this tendency.

D. J. Chalmers [26] tackles the question of the evidence of consciousness or non-consciousness for a Large Language Model. He considers several indicators of consciousness which could be identified to determine if a Large Language Model is conscious. He points out the difficulty to take into account answers of the Large Language Model about its own consciousness without knowing if it has been trained on the subject, and he discusses several of these indicators: the issue of a unified agency, as opposed to an entity taking the shape of different agents, the necessity to have a biological framework, as opposed to silicon systems, the capacity to have sensory processing or to perform bodily actions, providing sensory consciousness and bodily consciousness, as opposed to cognitive consciousness. In conclusion, he gives a low confidence degree to the consciousness of current LLMs, but he predicts future LLMs satisfying many criteria of consciousness. He warns us about the danger to create conscious artificial systems “unknowingly and unreflectively”.

4. ROBOTS AND CONSCIOUSNESS

We may think that robots, in particular humanoid robots [27, 28], are the first artificial systems for which we thought of consciousness. It was not the case.

An early proposal to create conscious robots was due to O. Holland and R. Goodman [29] on the basis of the evolution of living beings over the millennia, progressively increasing their intelligence until the existence of human beings with a conscience. Similarly, they propose to create a series of robots with a more and more sophisticated intelligence until they reach consciousness, by building and exploiting internal models.

Another proposal by R. Manzotti and A. Chella [13] is based on functionalism, regarded as necessary to the notion of consciousness. They consider that it is possible to create robots with a real subjective experience, not on the basis of internal processes, but thanks to relations developed with the external world. They question the possibility of robots to suffer and have moral rights.

Imitation is at the root of the proposal by J. A. Reggia, G. E. Katz, G. P Davis [30] to create cognitive robots which can learn by imitating demonstrations provided by humans, with a working memory to retain the explanatory interpretations constructed by the robot during learning and a form of inductive reasoning to discover the intentions of the human when performing a task. They think that conscious robots can be obtained in this manner.

Considering the particular case of vision, J. K. O'Regan [31] adopts a sensorimotor approach of feel, as a way of interacting with the world, focusing for instance on pain. He claims [32] that there is no logical impossibility to build a robot with access consciousness, as it is a behavioral capacity, including the capacity of self-knowledge. In an approach akin to an aspect of Higher Order Theory, he considers a form of access consciousness based on cognitive access, stating that an agent must have cognitive access to the fact that it has cognitive access to a piece of information. He presents [32] experiences in robotics to implement the notion of self, considered as a critical aspect of access consciousness, and regarded from several points of view, for instance self-distinguishing or self-knowledge.

5. MORAL STATUS OF CONSCIOUS BEINGS

As we have already mentioned, several researchers consider that a conscious artificial being should have a moral status requiring to protect them from suffering and giving them some rights. Some of them consider that the fact that an artificial intelligence may or may not have moral rights is the main reason to study its possible consciousness and answers to this question depend on the kind of conscious states the intelligent system can have [33].

Recently, Y. Bengio and E. Elmoznino [34] published a text on the risk to believe that some artificial systems being conscious have a moral status. Extending human rights and justice to artificial systems is not simple, since they are not "mortal and fragile", as human systems are, and their components and memories can be reused indefinitely, their needs are substantially different from human needs. They are also more intelligent than humans, so the notion of justice and equality between individuals is called into question. In addition, if they are granted a goal of self-preservation because they are conscious, a conflict may occur between their right to survive and human's right to protect himself. Faced with so many problems, the authors conclude that it is less risky for human beings to construct intelligent systems which are not conscious.

An overview of the links between artificial phenomenal consciousness and moral personhood is proposed by K. Mosakas [35], who discusses the possibility that artificial systems have a form of phenomenal consciousness, emotional states, sensory awareness and a sense of self. He focuses [36] on the particular case of so-called social robots, autonomous, interacting and communicating with humans. He presents several options to support the idea that a social robot has a moral status: the first one is to consider that a being has a moral status if this being has a phenomenal consciousness, another option suggests that consciousness is not a necessary condition for a robot to have a moral status, and one possibility is that it has a moral status if it is in social relations with other entities.

A. Chella [37] discusses the influence of the understanding of artificial consciousness on our capability to create more ethical artificial systems. He wonders if an artificial system can be a moral agent, and if some artificial consciousness is necessary to ensure an ethical behavior. He inventories several approaches of the consciousness of robots, for instance artificial empathy, considering that a perturbation moving the robot from a stable state can be associated with a negative feeling, while a perturbation bringing the robot closer to its stable homeostatic state can be associated with a positive feeling, enabling it to "feel something like a feeling" and then to feel empathy for human or artificial individuals. Another approach that he mentions is artificial wisdom, stating that an ethical artificial being must decide its actions on the basis of its "wisdom" and not by learning doctrines of ethics, which requires a capability to understand its context, and he mentions the implementation of such an artificial agent.

6. FUZZINESS AND CONSCIOUSNESS

Fuzziness can be regarded as the absence of crisp category boundaries, a gradual transition from inside to outside a class, the possibility for an element to belong partially to a category, the existence of approximate values of features, the utilization of natural language words to describe numerical variables, the management of degrees between 0 and 1 to weigh membership, confidence, certainty or possibility. In short, given a universe X, a fuzzy subset A of X is defined by a membership function f_A defined on X, taking values in [0, 1], expressing that an element x of X belongs to A with a degree $f_A(x)$.

If we think of computational functionalism, we can see that in some cases, the fact that a system satisfies one of the properties of consciousness is uncertain or indeterminate. Given the list of properties taken into account, a system can also be partly conscious, or conscious to some degree. We can also assign a confidence degree to the fact that a system is conscious. Then we can consider that computational functionalism accepts a form of fuzziness in the assessment of consciousness.

This fuzzy vision of consciousness leads us to think that fuzzy set theory could be useful in the analysis of the consciousness of artificial systems and in the creation of conscious systems, taken in the broadest possible sense, in spite of the present overwhelming success of Large Language Models.

Several utilizations of fuzzy methods to deal with consciousness have already been proposed. A dynamical model of phenomenal experiences [38] taking fuzziness into account makes the hypothesis that brain functions as a dynamical system which can be regarded as a tube in which thoughts, concepts and actions exist. Consciousness moves in this tube and meets concepts and actions which can be fuzzy and which function as attractors. Their fuzziness is a consequence of their change with the environment and with time. When the trajectory of a mental element gets close to a concept, the individual becomes aware of this concept. Such an approach leads the authors to consider the intensity of consciousness, which means that consciousness is a fuzzy concept. This definition of consciousness through a fuzzy dynamical system would be interesting to develop further and to adapt to artificial beings.

Beyond this approach of consciousness through fuzzy methods, we can consider a fuzzy description of its constituents, starting with perceptions and concepts. According to U. Eco [39], concepts participate in the representation of perceived phenomena by unifying sensorial impressions. L.A. Zadeh [40] initiated the idea of a computational theory of perceptions, to enable artificial systems to reason and to operate with perception-based information. He supposes that perceptions are described by means of natural language, with characterizations such as “it is likely”, if we consider the example of the perception of probabilities.

Concepts have been qualified as messy by E. Rissland [41], with a core and a “penumbra” where they can have an imprecise interpretation, they can evolve and change according to the context, and they can have exceptions. Such a view of concepts opens the door to fuzzy concept representation, which has been investigated since the first proposals of fuzzy concept lattices [42, 43]. A fuzzy context [44] is a triple (X, Y, I), where X is a set of objects, Y a set of attributes and I a fuzzy relation lying in a set L of truth degrees, for instance [0, 1]. For subsets A of X and subsets B of Y, fuzzy formal concepts are the pairs <A, B> such that A is the collection of all objects that have all the attributes of B and B is the collection of all attributes associated with all the objects of A. Using a fuzzy model of concepts provides more softness to the management of concepts, making them more appropriate to deal with subjective elements, feelings, affects or sensory information.

More generally, all aspects of fuzzy set theory involved in interpretable and eXplainable Artificial Intelligence, as initiated by L.A. Zadeh [45], can provide elements to approach the consciousness of intelligent systems. In particular, the capacity to express subjective information [46, 47] in a form of natural language benefits greatly from using fuzzy ontologies [48, 49] to cope with imprecision and vagueness of concepts, very useful to describe concepts or relations. Fuzzy description logic [50, 51] is an extension of description logic enabling to reason on the basis of fuzzy individuals or roles and fuzzy relations and it is appropriate, for instance, to manage emotions or perceptions.

Since emotions are at the core of phenomenal consciousness, we must mention the importance of fuzzy methods in affective computing. Fuzzy logic has been used in the development of various systems that can model and influence human emotion-related states and processes [52, 53]. Among many other examples, we can cite the analysis of text emotional content [54], the assessment of gameplay emotions from physiological signals [55–57], automatic recognition of facial expression [58–64] or the creation of intelligent agents [65]. Such methods have for instance been used in Kansei engineering, to translate feelings and impressions into product parameters [66, 67]. More importantly, fuzzy methods have been used for decades in the man-machine interactions to help machines respond to feelings expressed by users [68–71].

7. CONCLUSION

To conclude this short overview of the possibility for artificial systems to be conscious, we can remark that the concept of consciousness being so complex, several angles can be chosen to tackle the problem. We can think of simulated consciousness or real consciousness. We can consider phenomenal consciousness or access consciousness. We can restrict our analysis to human beings or take other conscious animals into account, which has not yet been done.

In the same way as planes flying differently from birds or butterflies, we can imagine conscious artificial systems which have a form of consciousness different from the human one, not based on emotions or qualitative experiences. We can also think of artificial beings which possess some of the features underlying the consciousness in human beings. It may also be possible to design highly efficient artificial intelligence-based systems which are not conscious.

If we think of simulated consciousness, research on interpretable and explainable artificial systems can contribute to some forms of consciousness. Fuzzy set-based knowledge representation can provide tools to more appropriately represent all kinds of concepts and to cope with subjective information, emotions and feelings, helping to improve the consciousness of artificial beings and robots.

It is important to see that working with conscious artificial systems does not solve all ethical problems related to Artificial Intelligence, since consciousness and conscience belong to two different paradigms. The analyses we have mentioned in this paper are only related to the former one, and not to the capacity to decide which actions are ethically or morally acceptable. Several questions remain open: should we try to create artificial systems with a real consciousness? Is it a danger for humanity, or a danger for artificial beings? Is there a benefit for artificial beings or a benefit for humanity? If we built artificial systems with simulated consciousness, what is the risk for human beings to be duped or deceived?

8. ACKNOWLEDGEMENT

A previous version of this work was presented in a plenary lecture given at the IEEE World Congress on Computational Intelligence (IEEE WCCI 2024) in 2024 in Japan.

References

[1] https://www.science.org/content/article/if-ai-becomes-conscious-how-will-we-know

[2] Lenharo M. If AI Becomes Conscious: Here's How Researchers Will Know. Nature. 2023

[3] Asimov I. *I, Robot*. Gnome Press. 1950..

[4] Dreyfus HL. What Computers Can't Do: The Limits of Artificial Intelligence. New York: Harper & Row. 1979.

[5] Minsky M. Why People Think Computers Can't. AI Mag. 1982;3:3-15.

[6] Pitrat J. Artificial Beings: The Conscience of a Conscious Machine. Wiley-IEEE Press. 2009:256.

[7] Dehaene S, Lau H, Kouider S. What Is Consciousness, and Could Machines Have It? Science. 2017;358:486-492.

[8] https://twitter.com/ilyasut/status/1491554478243258368

[9] Seth AK, Bayne T. Theories of Consciousness. Nat Rev Neurosci. 2022;23:439-452.

[10] Bayne T, Seth AK, Massimini M, Shepherd J, Cleeremans A, et al. Tests for Consciousness in Humans and Beyond. Trends Cogn Sci. 2024;28:454-466.

[11] Butlin P et al. Consciousness in Artificial Intelligence: Insights From the Science of Consciousness. 2023. ArXiv preprint: https://arxiv.org/pdf/2308.08708

[12] https://www.darpa.mil/program/explainable-artificial-intelligence

[13] Manzotti R, Chella A. Good Old-Fashioned Artificial Consciousness and the Intermediate Level Fallacy. Front Robot AI. 2018;5:39.

[14] Damasio A. Sentir ET Savoir, Une Nouvelle Théorie de la Conscience. Odile Jacob. 2021:145-151.

[15] Block N. On a Confusion About a Function of Consciousness. Behav. Brain. Sci. 1995;18:227-247.

[16] Sokolowski R. Introduction to Phenomenology. Cambridge: Cambridge University Press. 2000.

[17] Sabah G. Conscience et Intelligence artificielle(s) vues par Jacques Pitrat. Rev Ouverte Intell Artif. 2022;3:179-192.

[18] Tononi G. An Information Integration Theory of Consciousness. BMC Neurosci. 2004;5:42.

[19] Chalmers DJ. The Conscious Mind. 2nd edition. Oxford University Press. 1996.

[20] Chalmers DJ. Facing up to the Problem of Consciousness. J Conscious Stud. 1995;2:200-219.

[21] Burgess CP, Matthey L, Watters N, Kabra R, Higgins I, et al. MONet: Unsupervised Scene Decomposition and Representation. 2019. ArXiv preprint: https://arxiv.org/pdf/1901.11390

[22] Greff K, van Steenkiste S, Schmidhuber J. On the Binding Problem in Artificial Neural Networks. 2020. ArXiv preprint: https://arxiv.org/pdf/2012.05208

[23] Lotter W, Kreiman G, Cox D. Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. 2017. Arxiv preprint: https://arxiv.org/pdf/1605.08104

[24] Sajjadi MS, Duckworth D, Mahendran A, van Steenkiste S, Pavetić F, et al. Object Scene Representation Transformer. 2022. ArXiv preprint: https://arxiv.org/pdf/2206.06922

[25] Colombatto C, Fleming SM. Folk Psychological Attributions of Consciousness to Large Language Models. Neurosci Conscious. 2024;2024:niae013.

[26] Chalmers DJ. Could a Large Language Model Be Conscious? 2024. ArXiv preprint: https://arxiv.org/pdf/2303.07103v3

[27] Hu Y, Chen B, Lin J, Wang Y, Wang Y, et al. Human–Robot Facial Coexpression. Sci Robot. 2024;27:eadi4724.

[28] Lungarella M, Metta G, Pfeifer R, Sandini G. Developmental Robotics: A Survey. Connect. Sci. 2003;15:151-190.

[29] Holland O, Goodman R. Robots With Internal Models, a Route to Machine Consciousness? J. Conscious Stud. 2003;10:77-109.

[30] Reggia JA, Katz GE, Davis GP. Humanoid Cognitive Robots That Learn by Imitating: Implications for Consciousness Studies. Front Robot AI. 2018;5:1.

[31] O'Regan JK. How to Build Consciousness Into a Robot: The Sensorimotor Approach. In50 years of artificial intelligence: Essays dedicated to the 50th anniversary of artificial intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg. 2007:332-346.

[32] O'Regan JK. How to Build a Robot That Is Conscious and Feels. Minds Mach. 2012;22:117-136.

[33] N. Alonso, "An Introduction to the Problems of AI Consciousness", *The Gradient*, 2023, https://thegradient.pub/an-introduction-to-the-problems-of-ai-consciousness

[34] Bengio Y, Elmoznino E. Illusions of AI Consciousness, the Belief That AI Is Conscious Is Not Without Risk. Science. 2025;389:1090-1091.

[35] Mosakas K. Artificial Consciousness and Moral Personhood. England: Oxford University Press. 2025.

[36] Mosakas K. On the Moral Status of Social Robots: Considering the Consciousness Criterion. AI Soc. 2021;36:429-443.

[37] Chella A. Artificial Consciousness: The Missing Ingredient for Ethical AI? Front Robot AI. 2023;10:1270460.

[38] Huette S, Spivey MJ. Fuzzy consciousness. In: Edelman S, Fekete T, Zach N, editors. Being in Time: Dynamical Models of Phenomenal Experience. Amsterdam: John Benjamins Publishing Company. 2012:149-164.

[39] Eco U. Kant and the Platypus : Essays on Language and Cognition. Harcourt Brace. 2000.

[40] Zadeh LA. A New Direction in AI. AI Mag. 2001;22:73-84.

[41] Rissland EL. AI and Similarity. IEEE Intell Syst. 2006;21:39-49.

[42] Belohlávek R. Fuzzy Concepts and Conceptual Structures: Induced Similarities. In: Joint Conf. Inf Sci'98 proceedings. 1998;98:179-182..

[43] Burusco Juandeaburre A, Fuentes-González R. The Study of L-Fuzzy Concept Lattice. Mathware Soft Comput. 1994;3:209-218.

[44] Belohlávek R, Vychodil V. What Is a Fuzzy Concept Lattice? International Conference on Concept Lattices and their Applications. 2005:34-45.

[45] Bouchon-Meunier B, Lesot MJ, Marsala C. Lotfi A. Zadeh, the Visionary in Explainable Artificial Intelligence. TWMS J. Pure Appl. Math. 2021;12:5-13.

[46] Bouchon-Meunier B, Lesot MJ, Marsala C. Modeling and Management of Subjective Information in a Fuzzy Setting. Int J Gen Syst. 2013;42:3-19.

[47] Marsala C, Bouchon-Meunier B. Fuzzy Data Mining and Management of Interpretable and Subjective Information. Fuzzy Sets Syst. 2015;281:252-259.

[48] Bobillo F. Managing vagueness in ontologies [Ph.D. thesis]. Granada: University of Granada. 2008.

[49] Bobillo F, Straccia U. Fuzzy Ontology Representation Using OWL 2. Int J Approx Reason. 2011;52:1073-1094.

[50] Straccia U. A Fuzzy Description Logic. In AAAI/IAAI. 1998:594-599.

[51] Straccia U. Reasoning Within Fuzzy Description Logics. J Artif Intell Res. 2001;14:137-166.

[52] El-Nasr MS, Yen J, Ioerger TR. Flame—Fuzzy Logic Adaptive Model of Emotions. Auton. Agents Multi-Agent Syst. 2000;3:219-257.

[53] Jefferson C, Liu H, Cocea M. Fuzzy Approach for Sentiment Analysis. In 2017 IEEE international conference on fuzzy systems (FUZZ-IEEE). IEEE. 2017:1-6.

[54] Dzogang F, Lesot MJ, Rifqi M, Bouchon-Meunier B. Early Fusion of Low Level Features for Emotion Mining. Biomed. Inform. Insights. 2012;5:BII-S8973.

[55] Levillain F, Orero JO, Rifqi M, Bouchon-Meunier B. Characterizing Player's Experience From Physiological Signals Using Fuzzy Decision Trees. In Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games. IEEE. 2010:75-82.

[56] Mandryk RL, Atkins MS. A Fuzzy Physiological Approach for Continuously Modeling Emotion During Interaction With Play Technologies. Int. J. Hum. Comput. Stud. 2007;65:329-347.

[57] Orero J, Levillain F, Damez-Fontaine M, Rifqi M, Bouchon-Meunier B. Assessing Gameplay Emotions From Physiological Signals: A Fuzzy Decision Trees-Based Model. In INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH 2010. 2010:1684-1693.

[58] Ralescu A, Hartani R. Reasoning About the Transition Between Facial Expressions. In Proceedings of 1994 3rd IEEE International Workshop on Robot and Human Communication. IEEE. 1994:186-191.

[59] Ralescu A, Hartani R. Fuzzy Modeling Based Approach to Facial Expressions Understanding. J Adv Comput Intell Intell Inform. 1997;1:45-61.

[60] Bahreini K, van der Vegt W, Westera W. A Fuzzy Logic Approach to Reliable Real-Time Recognition of Facial Emotions. Multimed Tools Appl. 2019;78:18943-18966.

[61] Chakraborty A, Konar A, Chakraborty UK, Chatterjee A. Emotion Recognition From Facial Expressions and Its Control Using Fuzzy Logic. IEEE Trans. Syst. Man Cybern. A Syst. Hum. 2009;39:726-743.

[62] Ilbeygi M, Shah-Hosseini H. A Novel Fuzzy Facial Expression Recognition System Based on Facial Feature Extraction From Color Face Images. Eng. Appl. Artif. Intell. 2012;25:130-146.

[63] Liliana DY, Basaruddin T, Widyanto MR, Oriza II. Fuzzy Emotion: A Natural Approach to Automatic Facial Expression Recognition From Psychological Perspective Using Fuzzy System. Cogn Process. 2019;20:391-403.

[64] Nicolai A, Choi A. Facial Emotion Recognition Using Fuzzy Systems. In 2015 IEEE international conference on systems, man, and cybernetics . 2015:2216-2221.

[65] Stenseke J. Artificial Virtuous Agents: From Theory to Machine Implementation. AI Soc. 2023;38:1301-1320.

[66] Lesot MJ, Bouchard C, Detyniecki M, Omhover JF. Product Shape and Emotional Design: An Application to Perfume Bottles. In: Proceedings of the of the Int. Conference on Kansei Engineering and Emotional Research, KEER'10. 2010.

[67] Amin MA, Afzulpurkar NV, Dailey MN, Esichaikul V, Batanov DN. Fuzzy-C-Mean Determines the Principle Component Pairs to Estimate the Degree of Emotion From Facial Expressions. In International Conference on Fuzzy Systems and Knowledge Discovery. Berlin, Heidelberg: Springer Berlin Heidelberg. 2005:484-493.

[68] Chen L, Su W, Feng Y, Wu M, She J, et al. Two-Layer Fuzzy Multiple Random Forest for Speech Emotion Recognition in Human–Robot Interaction. Inf Sci. 2020;509:150-163.

[69] Chen L, Wu M, Zhou M, Liu Z, She J, Hirota K. Dynamic Emotion Understanding in Human–Robot Interaction Based on Two-Layer Fuzzy Svr-Ts Model. IEEE Trans. Syst. Man Cybern. Syst. 2020;50:490-501.

[70] Chen L, Zhou M, Wu M, She J, Liu Z, et al. Three-Layer Weighted Fuzzy Support Vector Regression for Emotional Intention Understanding in Human–Robot Interaction. IEEE Trans Fuzzy Syst. 2018;26:2524-2538.

[71] Yan HB, Huynh VN, Murai T, Nakamori Y. Kansei Evaluation Based on Prioritized Multi-Attribute Fuzzy Target-Oriented Decision Analysis. Inf Sci. 2008;178:4080-4093.