A Physicalist Causally Oriented Foundation For a Conscious Machine based on the Spread Mind

Riccardo Manzotti[1]

[1] IULM University, Milan, 20143, Italy
riccardo.manzotti@gmail.com

Abstract. Consciousness is here taken as the capability to experience something else than one's internal physical underpinnings as it happens in human beings. It is an empirical fact that, when we are consciousness we experience, say, the external world rather that what takes place inside our brain. This is puzzling. To solve this puzzle the literature has considered various options that are not obviously implementable in a machine. In fact, to address this issue, authors have always appealed to *ad hoc* ontological phenomena – e.g., Tononi's IIT, emergent properties, qualia, computational models – that are not part of the standard picture of the physical world. The problem of these solutions is that they add an unexplained principle to explain consciousness, but they do not explain why the physical world should have such an additional feature. However, recently a different solution has been put forward, named *Spread Mind*, that does not require any metaphysical addition to the physical world and that is perfectly compatible with the physical world as it is. In this paper, I will summarize the Spread Mind and I will propose how can such a solution can be used to design and implement an artificial conscious machine.

Keywords: First Keyword, Second Keyword, Third Keyword.

1 Consciousness and machines

It is fair to say that consciousness is a physical phenomenon and, as such, there is no a priori reasons as to why it might not be implemented in a machine. However, to be able to do so, a physicalist model of consciousness is required. If consciousness were an immaterial phenomenon, machine consciousness would be impossible. However, so far, consciousness has allegedly been an unsolvable challenge to physicalism because consciousness seems capable of doing something that does not fit with the standard view of the physical world. Such a capacity, which has led to several widespread formulations – Levine's epistemic gap, Chalmers's hard problem, Nagel's what it is like [1–3] – can be summarized as the capacity of experiencing something that does not obviously seem to be instantiated in one's body.

Consider an example that works both in human and tentative conscious AI systems. First, consider the human case. I see yellow and yet, inside my brain, nothing is yellow. How is that possible? How can my neural activity be associated with a property that is not physically there? In AI, the question is when an AI system will experience the meaning of the information it processes rather than being a pure syntactical operator.

In successful standard perception, a neural process occurs in the brain and, as a result, I experience yellow. However, how do the bits that take place in my head get the meaning of yellow has never been explained properly and led to the famous gap between semantics and syntax in information processing [4]. Bits, per se, do not have any meaning. And the problem is even more unsolvable in AI since in AI we cannot appeal to mysterious emergent properties hidden in the brain. When an AI system will experience something that, at the best of our knowledge, is not physically inside the AI system. For instance, when will an AI system experience yellow even if inside its physical implementation nothing is yellow?

By and large, the tradition has responded to this puzzle deploying three classes of models that have revealed to be largely unsatisfactory. Regarding consciousness and AI, it is informative to compare them briefly, albeit with some simplification, to make it clear why the proposed solution may be the only option for machine consciousness.

1.1 Denial of the problem

In the first group there are those authors that approached the issue of consciousness by denial – consciousness is either a collective delusion or a conceptual mistake. Accordingly, there is nothing that has to be explained. Consciousness is only a complex behavior that will be explained in the same way in which AI has dealt with intelligence, by means of incremental progresses on individual problems. In this class, functional models such as Global Workspace are the prevailing options [5, 6]. These models are oblivious to anything but the functional aspects of consciousness. In fact, such models do not address the problems of consciousness as such, they address only the cognitive functions that, in human beings, appear to be closely related with consciousness (memory, attention, cognitive workspace). These approaches are welcomed by the AI community because they seem to avoid metaphysical questions about the nature of consciousness. They are also connected with the separation between Strong Machine Consciousness and Weak Machine Consciousness, where they part with the latter [7]. While the proponents of these approaches defend the thesis that these incremental approaches will sooner or later get to the issue of consciousness [8], there is neither empirical evidence nor conceptual proof that this will be the case. They often ground their optimism on some loose analogy with other phenomena. For instance Todd E. Feinberg claimed that “From these neural features arise consciousness in a way comparable to how the complex property of life naturally arises from the interactions of its chemical and cellular components.” [9] Yet, there the analogy between life and consciousness is a very poor one. They have in common only the fact that, until a few years ago we were ignorant about both. However, so far these approaches have failed to address the key problem: how can an AI system see yellow. Is cognitive/computational power enough or something else is needed?

1.2 Intentionality and AI

In the second group there are those authors that believe that to experience yellow, there is no need to instantiate anything yellow, and yet it is possible to address yellow in the external world by means of some relations – aboutness, representation, intentionality. The advantage of this approach is that, once again, the real problem of consciousness – how can a conscious system have an experience of yellow – has been dodged. A vehicle – be it a neural activity or electronic patterns in a silicon chip – does not need to instantiate anything fancy, it is enough to be in the right relationship with its meaning. Thus, a bit does not need to be yellow, it is enough that it *represents* yellow. Unfortunately, so far, no one has been able to suggest a way to implement the kind of representation that leads to consciousness in a physical system, neither biological nor artificial. Notwithstanding the copious literature on the topic of the naturalization of representation [10–12], the only working notion of representation is functional – i.e.. *x* represents *y* if *x* has the function to stand for *y* in some circumstances. A functional representation is, of course, coherent with AI – from neural networks to symbolic systems – but it does not cope with the issue of consciousness.

1.3 Special properties

The third broad group of solutions is represented by those solutions that put forward the hypothesis that consciousness requires some additional ingredients – e.g., panpsychism, quantum phenomena, emergent properties, qualia, hidden aspects of information, etc. A popular example of this explanatory strategy is offered by Tononi's theory of integrated information (IIT) [13, 14] . Let's consider it given its recent success in the field of consciousness studies. Tononi has proposed that the processing of information creates something, which he has called integrated information and that such a phenomenon is what we call consciousness. This is surely an interesting hypothesis, but a hypothesis that substitutes the problem of consciousness with the problem of integrated consciousness. First, it does not explain why there must be anything like consciousness. Second, it does not explain why consciousness should have its properties (where the yellow comes from). Finally, integrated information is suspicious because its existence is epiphenomenal. It does not do anything. In fact, it cannot be directly observed, it can only be postulated. To recap, IIT is akin to a Russellian turtle – namely something that is placed epistemically under something else without having any better foundation for itself.

2 The Spread Mind: A Mind-Object Identity Theory

The shortcomings of the previous approaches seem to hamper any attempt to design a conscious AI system [15, 16]. How can an AI system see yellow? How can a brain see yellow? By and large, how can a physical system see yellow? After considering all possible alternatives I have elsewhere put forward an alternative model that addresses this issue and that might provide a conceptual foundation for consciousness in physical systems – be them biological or artificial. This model is called the Spread Mind (SM) and is based on the possibility of a Mind-Object Identity [17, 18]. I will outline briefly this model and then I will explain why it might allow to conceive conscious AI systems. SM is based on two straightforward and rather simple hypotheses about the nature of consciousness and the nature of the physical properties we experience.

First, SM consider an identity between conscious experience of a property instantiated by an external object and the very property instantiated by the object itself. In brief, rather than following any of the three approaches abovementioned, SM suggests considering a complete mind-object identity. In standard successful perception, whenever conscious experience of an object occurs, the object is there together with the ensuing neural activity. In brief, SM considers the possibility that there is an identity between the external properties instantiated by the external objects and one conscious states. What is the experience of yellow, according to SM? It is the yellow that takes place in the external object, for instance a banana. The advantage of SM is that there is no need to suppose anything more than the physical world. There are no special conscious states inside the brain and there are no special properties.

Second, SM stresses the intrinsically relative nature of physical properties – a notion possibly extended to the external objects themselves. In brief, SM suggests that all consciously experienced properties are physical properties instantiated by external objects relatively to one's body as in *relative* velocity. Thus, yellow is neither a disposition nor an absolute physical property. Yellow occurs relative to the causal circumstances offered by one's body and the external world.

Third, SM solves the aforementioned problem of intentionality by considering the identity between the physical underpinning of one's conscious state and what that conscious states is supposed to represent. There is no longer any need to introduce a mysterious arrow pointing from the internal states of the internal states to the external world. The agent's body has only the role to bring together relative properties in the external world.

SM suggests shifting the ontological basis of consciousness from the physical stuff inside the agent to the physical stuff in the external world. In this way, the problem of intentionality and semantics gets explained away. There is no longer any need to explain how bits inside a computer get linked to their external meaning, because the physical thing that is supposed to be one and the same with a mental state is no longer the bit inside the brain, but the external object. So, for example, there is no longer any need to explain how is that a bit is about an apple, because the physical state that is suggested to be one and the same with the experience of the apple is not the bit inside the body of the agent but the apple itself, which is just red as one's experience of the apple.

The external object relates to the body by means of a chain of causal processes not differently from those that takes place internally to the body. The idea that the processes taking place inside the body are somewhat closer to one's self and thus more adapt to be the basis of one's consciousness is parochial at best. From a physical perspective there are just physical processes. Some of them take place inside the body and some of them take place in the causal surrounding of the body. Is there any reason why we should prefer the former to the latter? As a matter of fact, I do not see any unless one assumed that the center of one's consciousness is the body, which would be question begging.

3 A Physicalist Model for Conscious AI systems

SM can be used to design and implement conscious AI systems [20, 21]. These are the key points:

1. A conscious experience of a property *p* is simply the property *p*.
2. The property *p* exists relative to the causal circumstances offered by the body of the agent.

A series of consequences follows.

1. A conscious AI system is necessarily an embodied and situated system.
2. Consciousness is not instantiated by the body of the agent, nor by the interactions between the agent and its environment.
3. Consciousness is the causal world, the body of the agent brings into existence by providing the right causal circumstances.
4. A conscious AI system is a physical system that has the same causal structure of a human body.
5. Consciousness is not a computational property, but a set of external causes.

SM places consciousness in the external world without any need of appealing to additional ontological principle. Moreover, SM is different from behaviorism or enactivism, because it focuses on external objects rather than on processes or on interactions.
As most models of consciousness, SM allows predictions about how design a conscious AI system (Table 1). The biggest advantage of SM is its ontological parsimony. It gives to the body the role of being the causal fulcrum of a collection of external causes (the relative properties *p*). The body of the agent does not have the role of instantiate consciousness as it happens in other theories – e.g, Tononi's IIT. The body of the agent has only the role of bringing together objects that are suggested being identical with consciousness. These has several advantages with respect to functionalism and enactivism. On the one hand, functionalism suggests an identity between functional states and conscious states leaving open the question of the ontology of such states and why they should have the properties we acknowledge in our experience. Enactivism suffers of similar problems dealing with the ontology and properties of sensory motor contingencies/affordances/enactions (which is why recent enactivists have withdrawn on cognition alone [19]). On the other hand, SM does not need any new ontological layers. The relative external objects and their properties are what the world is made of, regardless of cognitive agents. No emergence or additional properties are needed. Moreover, there is no need to match the alleged conscious states with properties of the external world, since consciousness and external world are supposed to be one and the same.

Table 1. Comparisons between consciousness models.

Model	Nature of consciousness	Conscious AI systems
Substance dualism	A substance outside the physical world	Impossible

Integrated Information (Tononi)	An additional epiphenomenal phenomenon whose existence is to yet uncertain	Achievable yet epiphenomenal
Cognitive approaches (Baars)	An epiphenomenal offshoot of certain functional processes	Achievable yet unrelated with consciousness
Mind-Brain Identity (Smart)	A property of biological processes	Impossible
Spread Mind	External relative yet physical objects	Achievable and causally effective

References

1. Chalmers DJ (1996) The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press, New York
2. Levine J (1983) Materialism and qualia: The explanatory gap. Philos Quarterly 64:354–361
3. Nagel T (1974) What is it like to be a Bat? Philos Rev 4:435–450
4. Searle JR (1984) Minds, brains, and science. Harvard University Press, Cambridge (Mass)
5. Baars BJ (1997) In the Theather of Consciousness. The workspace of the mind. Oxford University Press, Oxford
6. Shanahan MP (2010) Embodiment and the Inner Life. Cognition and Consciousness in the Space of Possible Minds. Oxford University Press, Oxford
7. Seth AK (2009) The Strength of Weak Artificial Consciousness. Int J Mach Conscious 1:71–82
8. Seth AK (2016) The real problem. Aeon 11:1–11
9. Feinberg TE (2018) Unlocking the " Mystery " of Consciousness. Sci Am 1–9
10. Millikan RG (1984) Language, Thought, and other Biological Categories: New Foundations for Realism. MIT Press, Cambridge (Mass)
11. Dretske FI (1995) Naturalizing the Mind. MIT Press, Cambridge (Mass)
12. Petitot J, Varela FJ, Pachoud B, Roy JM (1999) Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science. MIT Press, Cambridge (Mass)
13. Tononi G (2004) An information integration theory of consciousness. BMC Neurosci 5:1–22
14. Tononi G, Boly M, Massimini M, Koch C (2016) Integrated information theory: From consciousness to its physical substrate. Nat Rev Neurosci 17:450–461.
15. Manzotti R, Chella A (2018) Good Old-Fashioned Artificial Consciousness and the Intermediate Level Fallacy. Front Robot AI 5:1–10.
16. Chella A, Manzotti R (2012) Jazz and machine consciousness: Towards a new turing test. In: AISB/IACAP World Congress 2012: Revisiting Turing and His Test: Comprehensiveness, Qualia, and the Real World, Part of Alan Turing Year 2012
17. Manzotti R (2017) The Spread Mind. Why Consciousness and the World Are One. OR Books, New York
18. Manzotti R (2017) Consciousness and Object A mind-object identity physicalist theory,

Advances i. John Benjamins Pub., Amsterdam

19. Hutto DD, Miyn E (2017) Evolving Enactivism. Basic Minds Meet Content. The MIT Press, Cambridge, USA

20. Manzotti R, Jeschke S (2016) A causal foundation for consciousness in biological and artificial agents. Cogn Syst Res 40:172–185.

21. Manzotti R, Chella A (2015) The Causal Roots of Integration and the Unity of Consciousness. In: Poznanski RR, Tuszynsky JA, Feinberg TE (eds) Biophysics of Consciousness. World Scientific, Singapore, pp 189–229