AI Consciousness

Piotr Bołtuć[1.2][0000-0002-7210-615X]

[1] University of Illinois (UIS), Philosophy and Computer Science, Springfield, USA
[2] Warsaw School of Economics (SGH), E-Education, Poland
epetebolt@gmail.com

Abstract: We need a clearer ontology, epistemology and subsequently axiology for conscious cognitive architectures, biological and otherwise. This requires us to distinguish between two clusters: 1. functionally conscious cognitive architectures; 2. entities with first-person awareness. The former is essential to AI engineering. The latter relies on the first-person *feel*. Unfortunately, it cannot be built by the early 21st century AI. Neither could it be grasped with mid-20th century verificationist methodology; today, inference to the best explanation allows Non-Reductive Consciousness in psychology and philosophy. There may be a functional link between 1 and 2 (e.g. if NRC provides *phenomenal markers*). Even if NRC turns out not to be directly relevant for the functioning of cognitive systems, it remains indirectly relevant through axiology since we have reasons to care whether other beings have or lack their first-person *feel* of the world (the *Church-Turing Lovers* argument). In this paper I sketch out a deflationary theory of non-reductive first-person consciousness. Reflection on machine consciousness helps define such notion; it is also helped by it since non-reductive consciousness becomes a bit more of *a technical* term employable in visioning future AI.

Keywords: AI Consciousness, non-reductive physicalism, AGI, HLAI, BICA, The Engineering Thesis in Machine Consciousness

1 Non-Reductive Physicalism for AI

Intro: Machines already seem to satisfy important markers of consciousness, "things like self-awareness, knowledge, planning, and a theory of mind"; they even demonstrate *phenomenal* consciousness (contra [Block 1995] but consistently with [Franklin-Baars-Ramamurthy 2008]) and *experience* (*contra* [Chalmers 1995] and [Nagel 1974]. Those are some of the markers set by philosophers and philosophizing psychologists to establish special status of human consciousness; yet, if understood as technical functionalities of a cognitive architecture, they can be satisfied by some of today's AI systems [Schkolne 2018]. Paradoxically, those functional criteria are easier to satisfy than the criteria for conscious systems established by leading AI experts. Modern computer scientists tend to expect higher functional standards for fully conscious robots, such as versatility [Siegelmann 2018], ability for scientific and engineering discoveries [Thaler 2014], understanding 'what is going on' [Sanz 2012], or general intelligence that meets or surpasses human intelligence in thinking and behavior [Goertzel 2014].

The reason why many philosophers stick to non-reductive consciousness, phenomenal qualia or the Hard Problem of Consciousness [Chalmers 1996] is the following: While people and animals have *the feel* of first-person experiences of the world or of their own bodies, such as pain or color, there are no reasons to believe that modern robots have such *feels*. This is because robots lack the first-person epistemic awareness. ***Those philosophers try, and fail, to present their arguments in the third-person, scientific discourse.*** The reason for this failure is that third-person characteristics follow the functions that, by the physical interpretation of Church-Turing thesis, can be Turing-computable and, in principle, can be followed by machines [Deutsch 1985]. Due to the so-called *problem of other minds*, living, conscious organisms can only be described (from the third-person intersubjective perspective) as machines of sorts.

Method: The feel of first-person consciousness cannot be built by the early 21st century AI, or grasped with mid-20th century verificationist methodology [Dennett 1997]. A broader methodology, based on abduction, is better suited to generalize individual testimonies of people's first-person epistemic awareness, through correlations with intersubjectively verifiable facts in their CNR as well as in the external world [Harman 1986]. The latter can establish good inductive reasons for viewing first-person epistemic awareness as the best explanation of those testimonies. Once science learns how the feel of Non-Reductive Consciousness (NCR) is generated in animal brains, this should give us the blueprint for building projectors of consciousness, as well as boundary conditions of such design [Boltuc 2012, 2009, 2007].

There may be a functional link between third-person and first-person approach (e.g. if NRC provides *phenomenal markers* [Baars 1988]). Even if NRC is not directly relevant for the functioning of cognitive systems, it remains indirectly relevant through axiology since we have reasons to care whether other beings have, or lack, their first-person *feel* of the world (the *Church-Turing Lovers* argument [Boltuc 2017]).

1.1 Consciousness-Functionalities for Human Level AI

Non-Reductive Consciousness (NRC). Below I touch on three broadly philosophical topics important for today's Human Level AI. The first, versatility, does not require NRC. The second, whether consciousness needs to be humanoid, originates from divergent philosophical assumptions, thus relating to NRC at the framework level. The third, whether, how and to what extent machine consciousness could attain equivalence with animal consciousness, cannot be fully formulated without NRC; yet, it is crucial in the way signaled by [Deutsch 2012], [Goertzel 2017] *et al*, as explained below.

Versatile AI: Today's computing machinery faces a functional gap between the learning phase and implementation phase. During implementation, detail-oriented learning occurs, but structural learning that would change core functional sub-routines does not. This differentiates human intelligence from today's AI, making the latter insufficiently versatile. Thus, we need to build a *life-long learning* AI [Siegelmann 2018, 2003]. This requires us to develop correlations of attentional focus (detail oriented learning) with

self-modeling (structural learning) [Goertzel 2017]. Quite clearly, versatility does not require NRC.

Is AI consciousness necessarily humanoid? Does machine consciousness need to be human-like or could it follow some radically different developmental path/s? The main argument for the former is that human cognitive architecture is the only consciousness we know [Block 2018]. This view comes from prior acceptance of NRC and (oft associated) the privileged access claim, which leads to the argument that we know that other people are conscious only by analogy to our first-person experience. Hence, we could not identify a first-person consciousness radically different from ours; the problem is essentially epistemic. Those who advocate the latter approach [Sanz 2012], are not interested in NRC (as *the feel from the inside*) but about functional consciousness (performing the way a conscious intelligent being would) and they win this debate, for as long as we talk about functionalities of consciousness (f-consciousness). Thus, people working on conscious AI are talking cross-purpose – some about advanced machine functionalities, while others of NRC.

AI and non-reductive first-person consciousness: David Deutsch, claims that 'philosophy will be the key that unlocks artificial intelligence'. Development of AGIs stagnates primarily due to the philosophical misconceptions in formulating problems and methodologies for solving them. Hence, 'the problem of AGIs is a matter of philosophy, not computer science or neurophysiology', and philosophical progress 'is a prerequisite for developing' AGIs [Deutsch 2017]. While Goertzel argues that the "hard problem" of consciousness is sidestepped throughout, via focusing on structures and dynamics posited to serve as neural or cognitive correlates of subjective conscious experience.' [Goertzel 2014], he views NRC as a missing philosophical piece of thorough search on HLAI [Goertzel 2017b].

Those points highlight the need of fully-blown philosophy of machine consciousness, that goes beyond the socio-ethical areas, and to the *core business* of both AI and philosophy. A philosopher's job is to work out a clearer ontology, epistemology and subsequently axiology for conscious cognitive architectures, biological and otherwise.

Projectors of Consciousness. During the 2007 meeting of this forum, Nick Boltuc presented our argument, that AI should be able, in principle, to construct 'projectors of consciousness' [Boltuc 2007]: Once neuroscience learns how first-person consciousness is generated in animal brains, people should be able to reverse-engineer such projectors. We tried to develop such 'Engineering Thesis in Machine Consciousness', while avoiding the philosophical topic of the character of such consciousness.

Yet, it turns out that the attempt to avert from the hard philosophical questions leads to confusions. It seems unclear why we need a theory of consciousness if we can already satisfy the functional demands on conscious systems posed by philosophers. Below, I sketch a highly deflationary theory of non-reductive first-person consciousness to help with this. The theory has no ontology, and needs none.

2 Deflationary Non-Reductive Consciousness for AI

Deflationary view on non-reductive consciousness seems an easier fit with AI than the standard views. The main deflationary step I make is to dispense with the ontological account of first-person consciousness altogether, even such accounts as Chalmers' canonical version of the Hard Problem, defined as the problem of qualities of experience. Looking for some *substance* to pin down the first-person consciousness to is the Cartesian mistake. From the fact that I have a first-person experience of thinking, doubting (or, experiencing a *red* rose) it does not follow that I am a 'thinking thing'; it follows merely that there is experience of first-person *epistemicity*. The idea that it requires any substance is controversial, but the idea that it requires more than one substance clearly violates Ockham's Razor. Asking for ontological base of the epistemic perspective is *one question too many* [Elizabeth of Bohemia 1643].

Dennett and friends argue to the effect that even the concept of phenomenal qualia as a pseudo-substance commits a version of Cartesian fallacy; yet, they mistakenly presume that this point entails eliminating non-reductive first person consciousness [Dennett 1997]. The mistake is easy to make; it relies on an inadvertent endorsement of Descartes presumption that NRC must have a substance. But a simple move, used by various forms of the double aspect theory, shows otherwise: NRC can best be cashed out as an aspect, or functionality, of whatever monistic substance one is willing to endorse – oft physicalism or neutral monism. Our point here is not to endorse a double aspect theory, which should be viewed merely as the first step in the right direction. Instead, we harken back on a view providing a more explicit picture of how such ontological deflation would work: the *early Russellian monism* [Russell 1921].

Epistemology without ontology. How could we cope with epistemology without ontology? Bertrand Russell sketched out a version of this answer in his *Analysis of Mind* [Russelln1921]; not his *Analysis of Matter* [Russell 1927], which is now viewed as *the canonical version* of Russellian Monism [Alter 2015]. We have two different ways to conceive of the world: *qua* existing, or *qua* perceiving. The former is the ontological framework (or, the *ontological Gestalt*), the latter is the epistemic frame (or the *epistemological Gestalt*). Traditional neutral monism, including Russel's, views some abstract (neutral) ontological level as basic. Yet, is it necessary to postulate, redundantly, neutral matter beyond the material? It is clearer and simpler to maintain the two mutually non-reducible, complementary perspectives. None of the perspectives needs to be viewed as *prima facie* prior, or privileged, although material description is practically dominant as the gist of the empirical knowledge, thus of the sciences.

2.1 Re-Defining the Hard Problem

David Chalmers defines The Hard Problem of Consciousness as the problem of experience [Chalmers 1996]. Roughly, we have access only to our own qualia (experiential/phenomenal qualities, such as taste and color). But isn't it a version of the problem

of other minds? Wouldn't this naturalistic account suffice: I can't feel your pain (numerically speaking), but couldn't I feel pain just like yours, which could be established by scientific knowledge of the exact state of a CNR?

Chalmers seems to respond by asking: How do we know this? But this is a version of the skeptical take on the *other minds*, here applied directly to one's experience (qualia). The empirical work on *reading one's brain*, developed in retrieval of dynamic images from one's visual cortex, by Gallant's lab [Nishimoto 2011], puts into question such skepticism. If phenomenological experience can be translated into the intersubjective language, e.g. images, we attain not only objective phenomenology, but physicalism. Even the Clark-Chalmers extended mind hypothesis undermines important aspects of one's mind's epistemic privacy and primacy [Clark 1998]. Consequently, the Hard Problem should not be defined as a generalization of Nagel 1974] question *What it is like to be a bat?* In his *The View from Nowhere*, [Nagel 1974] workred with a better tool to tackle the problem – with a version of *pure subject*.

Pure subject. Nagel's *The View from Nowhere* is sometimes viewed as a preliminary, and less compelling of Nagel's formulations of the problem of consciousness e.g. [Cavanna 2014]; but this is not Nagel's original attitude. In [Nagel 1986] epistemic first-person perspective is non-reducible to any ontology; it is a view from no ontological location, thus from *nowhere.* This position does not include phenomenal qualia as its substantial part because qualia belong to the objects, and thus are not the essential features of the first-person subjective self. What is left when phenomenal qualia are no longer viewed as the gist of non-reductive consciousness? The first-person epistemic capability to have such experience is left: qualia are merely the most obvious *objects* accessible to the subject. The gist of the epistemic *subject* is its first-person *potentiality* to interact with the world and to co-constitute qualia (or abstract objects, depending on the kind of data-entry).

This point distinguishes British phenomenalism of Chalmers, Parfit, Price, Hume, Berkeley and Locke from the one inspired by German classical philosophy from *Leibnitz's mill*, and Fichte's *transcendental subject* (the most reduced notion of passive subjective self), to Cohen, Husserl's *pure transcendental self* and also [Nagel 1986]. The latter seems able to distill subjectivity beyond all the accidental features, and grasp pure *epistemicity*, or stream of awareness, distinct from the objects of attention (including phenomenal content), which allows for more clarity on what the gist of first-person consciousness is – its potential for subjective experience, not a specific kind of content.

Identity of NRC does not consist in phenomenal content. Peter Unger **[**Unger 1990] illustrates this point: Imagine a pair of philosophical twin (complete identity, including mental content till point T-now). But at T you and your philosophical twin are in phenomenally identical, yet numerically different rooms. Both of you know that one of the twins is going to undergo excruciating pain. Would you be indifferent which of the twins it shall be, you or the other one? If identical phenomenal content were the sole thing that matters [Parfit 1986], one should be indifferent. But no indifference. Based on this case, an underlying numerical identity of the stream of consciousness (either

based directly on epistemic *locus* of that stream of consciousness, or – naturalistically – on identity of its bodily emergence base), is what matters; not just its content.

At the level of most reduced analysis, the subject and object are best viewed as complementary basic non-reducible entities, while at the higher ontological level the epistemic locus may, as an empirical fact, originate from physical emergence base. Hence, we reach *deflationary* non-reductive physicalism. For the stream of awareness to manifest itself, it needs to be projected upon some content (phenomenal or informational) *of* such awareness. Yet, identity seems to be retained through the stream of awareness, not through continuity and connectedness (or any other features) of its content. Unger's seems more persuasive than Parfit on this.

We may re-define the hard problem of consciousness as not the problem of qualities of phenomenal content, but instead as the problem of unique access to one's own first-person grasp of the stream of awareness. There is a practical sense to this philosophical approach. If a nurse discovers that a patient regained consciousness after surgery, she does not care about details of the patient's content of consciousness; she cares just about the stream. The patient may have lost all the memories, and thereby continuity and connectedness of the content\; she may also lose most of her perceptive capabilities, some body parts, character threats etc., yet still regain the first-person consciousness. Identity stays with the *locus of awareness* – not with phenomenal content. The content-related questions come naturally, but they tend to come next.

Practicalities for future AI: Reflection on machine consciousness helps us define the notion of non-reductive deflationary consciousness within the framework of physicalism: clarity that AI brings into functional consciousness makes it quite visible that the only little thing that today's (or, near future) AI lacks, compared to human beings and many animals, is NRC. On the other hand broad reflection on AI is also helped by the seemingly just philosophical notion of NRC. We should see it as a natural phenomenon. produced in animal brain, details of which near-future neuroscience is likely to discover in great detail. Hence, there is little if any space for the aura of mysterianism on NRC.

We should be quite clear that our robots are still far from fully human level consciousness. AI satisfies some of the functional criteria for machine consciousness, including self-awareness, knowledge, planning, reaction to phenomenal markers (such as color, sound or temperature), or even creativity. In the near-future AI is undoubtedly going to be versatile and, later, attain human level general intelligence (HLAI). But the functionality of first-person epistemicity (NRC) – while naturalistically defined (in this paper and in [Boltuc 2019] – is the feature that robots still do not have. This is important for humanoid robots, especially artificial companions, since, due to the lack of NRC 'there is nobody home' behind their more and more elaborate caring, sexual, social and moral behavioral subroutines. This justifies the call not to treat today's robots as moral patients *tout court*, even though they may function as moral agents. There is nothing naturalistically impossible, or *prima facie* wrong, in attaining NRC for robots, we just do not have it now, and probably not for quite a while.

References

1. Alter, T. Nagasawa, Y. (eds.) Consciousness in the Physical World: Perspectives on Russellian Monism. Oxford University Press (2015)
2. Baars,B.J. A Cognitive Theory of Consciousness Cambridge University Press, (1988)
3. Block, N. Engineering Machine Consciousness (Symposium Slides), IACAP, Polish Academy of Science, Warsaw (2018)
4. Block, N. On a Confusion about a Function of Consciousness. Behav. Brain Sci., 18, (1995)
5. Boltuc P. Subject is no Object; Complementary Basis of Information. In: Burgin, M., Dodig-Crnkovic, F (eds.) Philosophy and Methodology of Information, vI, World Scientific (2019)
6. Boltuc, P. Church-Turing Lovers. In Lin, P., Abney, K., Jenkins, R., Eds Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence; Oxford University Press 214–228 (2017)
7. Boltuc, P. The Engineering Thesis in Machine Consciousness. Techne Res. Philos. Technol. 16, 187–20 (2012)
8. Boltuc, P. The Philosophical Issue in Machine Consciousness. Int. J. Mach. Conscious, 1, 155–176 (2009)
9. Boltuc, N., Boltuc P. "Replication of the Hard Problem of Consciousness in AI and Bio-AI: An Early Conceptual Framework" in AI and Consciousness: Theoretical Foundations and Current Approaches, eds. Chella, A.; Manzotti, R (Merlo Park, CA: AAAI Press), 24-9 (2007)
10. Cavanna, A. E.; Nani, A. Consciousness. Theories in Neuroscience and Philosophy of Mind. Springer (2014)
11. Chalmers. D. The Conscious Mind. Oxford University Press (1996)
12. Clark, A., Chalmers D. The Extended Mind Analysis 58 (1):7-19 (1998)
13. Dennett, D. Kinds of Minds: Towards an Understanding of Consciousness, Basic Books (1997)
14. Deutsch Philosophy will be the key that unlocks artificial intelligence, *The Guardian* 3 Oct (2012)
15. Deutsch, D. Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer. Proc. R. Soc. Ser. A 400, 97–117 (1985)
16. Elizabeth of Bohemia. Elizabeth to Descartes 6 May (1643). Ed. Shapiro, L. The Correspondence between Princess Elizabeth of Bohemia and Rene Descartes, CUP, (2007)
17. Franklin, S.Baars-< Ramamurthy, U. Phenomenally Conscious Robots? APA Newsletter on Philosophy and Computers, Fall, 08/1, 2-4 (2008)
18. Goertzel B. From Abstract Agents Models to Real-World AGI Architectures: Bridging the Gap: in: Everitt, T. Goertzel, B. Potapov, A. (eds.) Artificial General Intelligence. Springer (2017)
19. Goertzel B. Conscious AI, Lecture Notes, The Consciousness Workshop, Shanghai (2017b)
20. Goertzel, B. Characterizing Human-Like Consciousness: An Integrative Approach; OpenCog Foundation, Hong Kong, China September 7, (2014)
21. Harman G. Change in View MIT Press (1986)
22. Nagel, T. The View from Nowhere, (1986)
23. Nagel, T. The Philosophical Review, Vol. 83, No. 4, 435-450 (1974)
24. Nishimoto, S. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies Current Biology 21, 1641–1646, October 11, (2011)
25. O'Regan, K. Why Red Doesn't Sound Like a Bell: Explaining the Feel of Consciousness, Oxford University Press, (2011)
26. Parfit, D. Reasons and Persons. Oxford Univ. Oxford University Press, (1986)

27. Piccinini, G. Mind Gauging: Introspection as a Public Epistemic Resource, Grad Expo, University of Pittsburgh, Pittsburgh, PA, (2001)
28. Russell, B. *Analysis of Matter*, Spokesman, Nottingham (1927)
29. Russell, B. Analysis of Mind, London: George Allen and Unwin; New York: The Macmillan Company, (1921)
30. Sanz, R. Hernandez, C. and Sanchez-Escribano, M.. Consciousness, action selection, meaning and phenomenic anticipation. International Journal of Machine Consciousness, 4(2):383–399, (2012)
31. Schkolne, S. Machines Demonstrate Block's Consciousness in: Becoming Human Exploring Artificial Intelligence & What it Means to be Human 05/16 (2018)
32. Siegelmann, H. Lifelong Learning Machines (L2M) https://www.darpa.mil/program/lifelong-learning-machines, Last accessed 2018/12/17
33. Siwgelmann, H. Neural and Super-Turing Computing Minds and Machines 13 p. 103-114 (2003)
34. Thaler S. Synaptic Perturbation and Consciousness, International Journal of Machine Consciousness, Vol. 06, No. 02, 75-107 (2014)
35. Unger, P. Identity, Consciousness and Value Oxford Univ. Press (1990)