FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Piotr (Peter) Bołtuć

NON-REDUCTIVE PHYSICALISM FOR AGI

doi: 10.37240/FiN.2022.10.zs.2

To the memory of Gilbert Harman[1]

ABSTRACT

Creature consciousness provides a physicalist account of the first-person awareness (*contra* Rosenthal). I argue that non-reductive consciousness is not about phenomenal qualia (Nagel's *what it is like to feel like something else*); it is about the stream of awareness that makes any objects of perception epistemically available and ontologically present. This kind of consciousness is central, internally to one's awareness. Externally, the feel about one's significant other's that "there is someone home" is quite important too. This is not substance dualism since creature consciousness and functional consciousness are both at different generality levels of physicalism. Surprisingly, pre-Hegel philosophy of pure subject is more fitting with the current engineering approach than analytical phenomenalism. The complementary view of subject- and object-related perspectives, may come from Fichte's *Wissenschaftslehre*; but here it is placed, securely within the physicalist paradigm. It is essential to the Engineering Thesis in Machine Consciousness, which helps us understand under what general conditions a machine would be first-person conscious, but when it is merely functionally conscious.

Keywords: Machine consciousness, non-reductive physicalism, non-reductive machine consciousness, creature consciousness, non-reductive consciousness; complementary philosophy, *Wissenschaftslehre,* two-tier physicalism.

---

[1] This article is dedicated to the memory of Gilbert Harman, who has shown me at least three things: how to do analytical philosophy without outdoing on analyticity; how to reach beyond the boundaries of philosophy and gently disrespect those boundaries; how to treat a student with true respect. When I have asked Gil about the latter, he just said: Those are my colleagues who were born a bit later than I was. Several of them are to become stronger philosophers than I am. Why would I not treat them as such? *Oh well, Gill has also shown me how not only to love, but also truly respect, one's daughters; me and my five years old are grateful for this.*

1. INTRODUCTION

In Part I we focus on philosophical issues that relate to physicalism. The goal is to explain how complementary treatment of the subjective and objective epistemic perspectives can be consistent with physicalism. This leads to two-tier physicalism. We put forward a controversial thesis that *creature consciousness* is the base of first-person awareness; this is in stark contrast with a common tendency to view it as grounded only in advanced functionalities. Creature consciousness and the functionalist level of analysis at the third-person level of description are those two tiers of functionalism In Part II we apply the main points from Part I to first-person machine consciousness.

2. COMPLEMENTARITY

2.1. Non-reductive two-tear physicalism

Rarely would physicalists accept a non-reductive position on consciousness—and remain *physicalists through and through*. Such view requires accepting two different points:

1) physicalism—seen here as a view that everything has only physical causes, explainable directly or indirectly by the laws of nature, physical, chemical, biological, also mathematical, extending into engineering and computer science, medicine and *the soft sciences*: psychology, sociology, economics and other domains;

2) non-reductive view on first-person consciousness—denying the possibility of always having a reduction from first-person awareness to the third person view on the world or *vice versa*.

I accept both these points. Here is the account of the view:

The conscious experience functions as non-reductive in what psychology calls creature-consciousness. Thus, irreducibility consists in the stream of first-person awareness, not its phenomenal content. This may look like a blunder and incompetence to those philosophers, who dismiss ontologies different from various forms of phenomenalism (for instance, Locke's, Berkeley's, particularly Hume's and Parfit's). They maintain that there is nothing but continuity and connectedness of our experiences? *I say, not even close!*

2.2. Creature-consciousness physicalism

What sense of creature consciousness makes possible its role as the gist of the first-person experiential level? It is the stream of primary awareness, which may or may not be sophisticated in content and seems to have

different strengths among various creatures. We may compare it to the stream of light that comes from an old-fashioned movie projector before one would put in the film; only the latter provides non-trivial *phenomenal* content.

Non-reductive awareness as creature consciousness, is located at the level of bio-chemical specificity, not quite at the level of advanced functionalities, nor at the level of qualia phenomenalism. Creature consciousness, as a biochemical process, is clearly a framework at the physicalist level—few would question this. Consistent functionalism would normally be an application of physicalism as well.

Thus, we seem to have a two-tier physicalism of consciousness:

1. Creature physicalism at the level of creature consciousness
2. Functional physicalism at the level of (always physical) interactions with the environment and information transformation (thinking).[2]

The former provides biological explanations for aware interactions, the latter gives us a phenomenal (potentially meaningful) content. Within psychology of consciousness people are dismissive of the level of creature consciousness as trivial (David M. Rosenthal). I daresay it is not trivial at all. It provides material—in particular bio-chemical—grounding for the first-person stream of awareness.

How would such biochemical explanation give a relevant response to the philosophical problem of explanatory gap? It does not; *does not need to*.

Ned Block illustrates the scientific nature of *ice emerging from water* under certain temperature and pressure conditions. Thus emergence is a physicalist process fully explainable in physical sciences. Creature consciousness is the source of the stream of awareness. Daniel Dennett is right, that asking for a dualistic explanation of this process is begging the question—the process does not require or need such a level. But he is a bit overly entangled with dualities of his student times under supervision of Gilbert Ryle. Some philosophers seem chimed by the question how mechanical functions of matter generate first person "experiences" like the feeling of pain or unique experience of contemplating a certain shade of redness. Such a *Leibniz's mill* may have been justified, sort of, when physics was largely macro-level mechanics and there was a strong presumption towards micro-reductionism—today we know better. Natural sciences include sophisticated chemical and bio-chemical functions that allow for continuous mathematical processing, rather than discrete Turing computing (Sloman, 2020). Not all scientific explanations reduce science to the level of atomic interactions, at least as long as atoms are viewed as mechanical billiard balls.

---

[2] Physicalist theory of thinking was drafted already by Hobbes (not to mention Sextus Empiricus). It has been nicely developed by contemporary neuroscience and artificial intelligence.

Let us think of the two sides of a mirror (explored in some older British stories for the *nice kids*): How is it that the whole magic of reflections, and multicolored refractions at the well-cut edges (let us make it a stylish crystal mirror of the epoch) are possible despite the fact that a mirror is constructed of some wood and glass. Once we know some optics, the story is simple and hardly miraculous. Similarly, burning some wood to get fire seems to have mesmerized deep thinkers since the ice age, at least—but now it does not since we know basic chemistry of the process.

The reason why Dennett may be unable to address the question of first-person consciousness is his deeply rooted hard-core version of verificationism (that comes from Gilbert Ryle). Using Ryle's rough methodology, one is unable to formulate the problem of first-person consciousness; not to mention resolving it. The generalization of people's statements linked to their fMRIs and other neurophysical measures suffices to identify the problem (Thomas Metzinger's old argument)—which does not lead to a dualist impasse in the explanatory gap but may lead to a no-nonsense solution.[3]

Let us expand on the above points. Those who single out the first-person stream of awareness are often labelled "dualists." This brings about peculiar consequences if taken up in the context of creature consciousness, thus demonstrating how non-reductive physicalism is possible. People tend to view first-person consciousness as a sophisticated tool that even most primates lack (Rosenthal, *indirectly* Davidson), whereas it is a rather simple feature, probably present at roaches, ants, definitely at frogs.

To reiterate a novel but unsurprising idea: creature consciousness and functionalism when put together would be resulting in a two-tier *physicalist dualism. We are up to something here.* The two perspectives, first- and third-person, do bring about some kind of a dualism. Yet, this is a *perspectival dualism* (Nagel); *it does not fall into substance dualism,* which—I agree—is a rather bad idea. Substance dualism does not have—and very likely could never have—a good answer to the problem of interaction (Elizabeth of Bohemia).

Eliminativists, who deny first-person consciousness (or, its relevance), are guilty of a dismissive omission of the first-person epistemicity as ontologically relevant, actually necessary, for constructivist foundations of the reality (any non-trivial ontology).[4]

I propose that non-reductive first-person physicalism based on creature-consciousness is the best explanation[5] of first-person's awareness in the

---

[3] Especially if smart, critical minds—like Dennett's—join the game, instead of dragging their feet on its sidelines.

[4] This was discovered by classical German philosophy (especially Fichte), but later dismissed as idealism by most 20th century physicalists, due largely to their controversial take on verification(ism).

[5] And an inference to the best explanation it is.

objective world. *This is the gist of the philosophical framework of this paper.*

2.3. Recent take on complementary philosophy

Here we explore—on a couple of occasions merely mention—some of the instances of complementarity of subject and object within physicalist[6] analytical philosophy.

2.3.1. Double Nagel

Fortunately, physicalist phenomenalism is no longer the only game in town.

NagelA

Tom Nagel started with what I call the NagelA set of views—a complementary philosophy, with two basic, ontological (not just epistemic) starting points of his philosophical account of the world.

First, the *objective*, or rather object-related, account that views things as based on measurable, third-person verifiable claims—for instance pictures of a certain mountain that really looks different from various viewing-points were the pictures were taken.

Second, the first-person, or rather subject-related, account based largely on *the feel* of things. In building this double-aspect view (Nagel 1979A; 1986) relied to some degree on Peter Strawson who, in his book *Individuals*, opened up the option of doing non-reductive analytical philosophy of person. However, NagelA, relied largely on classical German philosophy.

As Gilbert Harman (2007) argued "It seems that whatever physical account of a subjective conscious experience we might imagine will leave it completely puzzling why there should be such a connection between the objective physical story and the subjective conscious experience." Harman[7] refers to the essay *Subjective and Objective* (Nagel, 1974, p. 196 f.)," which may be viewed as an early draft of the ontological part of Nagel's masterpiece, *The View from Nowhere* (Nagel, 1986). Nagel highlighted his allegiance to materialism in (Nagel, 1974; 1986). Within materialism, Nagel's question—as presented by Harman—would have a straightforward explanation through the theory of evolution. The fit between one's first-person opinion and the way things are tends to be a good survival mechanism, except for certain exceptions, like those made famous by Daniel Kahneman and Amos Tversky.

---

[6] This focus on physicalism or non-reductive materialism explains merely cursory presence of David Chalmers' theory, which is now a naturalistic version of panpsychism.

[7] In his office at Princeton Harman had Nagel's picture next to Quine's; he told me, these belong to his favorite philosophers.

Nagel's puzzlement with the gap between subjective and objective epistemic perspectives (the latter being "the view from nowhere" from his book's title) is not centered around this kind of explanation. It is metaphysical, the way Johann Gottlieb Fichte in his *Wissenschaftslehre* (and Husserl in his *Ideas*) would have it.[8]

NagelB

In his article *What Is It Like to Be a Bat?* (Nagel, 1974) Thomas Nagel focuses on a rather different, much narrower, picture. Nagel seems to have joined the crowd of his quite distinguished followers (e.g. Jackson; Chalmers) in viewing his early *propaedeutic* paper based on phenomenalism as the gist of non-reductive analytical philosophy. In *What it is Like to be a Bat* we have what I call NagelB (Nagel, 1979B), with Hume-style phenomenalism becoming the main basis of non-reductive philosophy of mind.

Nagel B's theory relies on the exaggeration of our inability to imagine what it is like to be someone else, e.g. a bat. It leads to abandonment of Nagel A's ontology based on complementary subject and object (subjective and objective)—instead, we face phenomenalist, epistemic worries on *the problem of experience.*

*"*What It Is Like to Be a Bat*"* is an easy-to-follow essay, popular among undergraduates. It ignores, comfortably, Kant's victory over Humean skepticism, which was so clearly appreciated, even developed in Nagel's crucial works, mentioned above. The understanding of complementarity of subject and object in ontology, and of the subjective and objective perspectives in meta-epistemology, if we may use this term, is essential to, and clear enough in Fichte's Wissenschaftslehre—and to some degree in the works of Edmund Husserl (e.g. the early parts of his *Ideas* (Husserl 1913), some Neo-Kantians and Kant. But this position is now viewed as wrong and largely abandoned.

The problem of phenomenal qualia as the problem of experience, reaches at least from David Hume, through Derek Parfit (who was the smartest critic of Nagel 1984; whereas McGinn was just destructively clever) to his temporary followers like Frank Jackson, the author of the influential *Black and White Mary* argument, which he has later rejected. Those discussions can easily be reduced to Nagel's *what it is like to be a bat* question*,* which is an unfortunate *lacuna* to take.

The German and British traditions in Nagel

NagelA's more insightful work (Subjective and Objective in Nagel 1979, and *The View from Nowhere*) opened up a complementary framework between object and subject related perspectives—within the view that Nagel

[8] For further discussion see (Boltuc, 2019).

rightfully called *materialism*. This was a discrete, yet steady-handed, move of classical German philosophy from the realm of idealism to that of physicalist non-reductive materialism. It can be maintained that Nagel put forth a materialist redescription of Fichte-style pure transcendental subject.

Now we face Tom Nagel's two theories of first-person consciousness: NagelA based on the gist of Fichte, Husserl and Kant, and Nagel B, based largely on Hume's phenomenalism. Nagel A and Nagel B read like two different philosophers—the former a much better one, but maybe a bit ahead of his times.

This is part of a big philosophical picture. Even when eminent German scholars try to highlight Kant's ideas helpful in modern cognitive science (T. Schlich, A. Newen), they start with a disclaimer that they do not try to endorse "Kant's idealism" and then merely search for scraps that remain when the structure of Kant's theory gets demolished. This attitude comes from philosophical predominance of British empiricism, narrowly understood, and abandonment of philosophical methods not approved by narrow such empiricism.

I view defending the heritage of Nagel A from the brute-force dominance of Nagel B as one of the main challenges for non-reductive physicalism.

2.3.2. Russell's short-lived analysis of mind

Bertrand Russell, in his *Analysis of Mind*, proposed a somewhat similar view. Each object, e.g. a chair, can be grasped as a set of objective parameters, as well as through phenomenal content (Hume-style), The full picture available to us, is composed of those two ontologies combined.

Yet, both Russell and Nagel fell back onto the *third-person accounts of physical object taken as a given*. In his *Analysis of Matter* Russell has taken matter, investigated from the third-person object-related perspective, as the primary and only substance. It may have been too early, in Russell's times, to understand scientific views broadly enough to keep his *complementary* physicalism, or to turn it into a two-tier physicalist theory.

We are now going to explore and expand first-person subject's complementary relation with third-person functionalism, referring to Harman's under-appreciated functionalism of concepts.

2.3.3. Harman's functionalism of concepts

Gilbert Harman attempted to open a non-reductive window within functionalism—the window based largely on semantics. This was done well in his underpublicized "functionalism of concepts" (Harman, 1990). The claim is, briefly, that we cannot fully understand first-person statements from the third person perspective. Therefore, we need to analyze the functions of concepts, both within their first- and third-person contexts of use. (Harman,

1990; 1993; 2007). This is based on Harman's argument "Knowledge that P requires being able to represent its being the case that P. Limits on what can be represented are limits on what can be known." This point originates from Harman's interest in Wilhelm Dilthey's (1989 (1883)) *Das Ferstehen* as a special first-person epistemic perspective (Harman, 1993; 2007).

> "With respect to pain and other sensory experiences there is a contrast between an objective understanding and a subjective understanding of what it is like to have that experience, where such a subjective understanding involves seeing how the objective experience as described from the outside translates into an experience one understands from the inside." (Harman, 2007)

"(I)n philosophical semantics" Harman distinguishes "between accounts of meaning in terms of objective features of use and translational accounts of meaning" (Harman, 2007, pp. 2–3). This approach provides him with an elegant account of the explanatory gap. Harman tries to use those issues in translation for "understanding what it is like for another creature to have a certain experience [...] To understand what it is like for the other creature to have that experience is to understand which possible experience of one's own is its translation" (ibid.) The context of translation theory gives Harman an objective reference frame, where phenomenal qualia and sort of hermeneutics come through without their usual drift towards a dualism of sorts.

Harman refers to the Lewis–Nemirow interpretation of *what it is like arguments*, developed in reference to Jackson's *Black and White Mary's case* – which claims that one lacks an ability (e.g. to identify the red objects), not lacking any knowledge. To this Harman responds: "For them, understanding what it is like to have a given experience is not an instance of knowing that something is the case, a conclusion that I find bizarre" (Harman, 2007, p. 3).

At Harman's seminar in epistemology (Spring 1991), after his presentation "Can Science Understand the Mind?" where he presented a crisp version of this argument, I developed a related point: Ability can be translated into knowledge when we take into account informational content of the programming that it takes to have a robot perform the task at hand (Boltuc, 1998a, 1998b).

To sum up, Harman argues for the following point: "purely objective account of conscious experience cannot always by itself give an understanding of what it is like to have that experience. There will at least sometimes be an explanatory gap. This explanatory gap has no obvious metaphysical implications" (Harman, 2007, p. 3). *This is exactly the outline of a non-reductive view on subjective and objective perspectives, kept neatly within physicalism.*

Writing this paper, I may be slightly more radical now, by trying to identify a physicalist emergence base, not only of the first-person content, but of

the first-person stream of pre-awareness that allows the subject-related perspective itself.

2.3.4. Subject-object complementarity

The basic subject-object relationship is what creates both the epistemic and ontological backgrounds for the existence of any phenomena. The complementary, view on the essential and mutually irreducible position of the subject and object, first as conceptual *atoms*, and next as originators of the complementary epistemic and ontic perspectives, turns out essential for formulating the ontological structure of autonomy for humans and other advanced animals alike. It is also relevant for advanced AIs.

The lack of the epistemic subject's primary interaction with the epistemic object results in the lack of consciousness *tout court*, though it does not affect complex conscious functions (if viewed from the outside). Having opened the philosophical background—so alien to, even detested by, philosophers of the post-Lockean, Humean schools—we need to demonstrate the same at the more practical level of everyday experience. While the first part of this paper may be viewed as complex and overly *philosophical*, the latter part may be viewed as simplistic. However, simple is good when thinking about first-person awareness.

Attempts to place non-reductive consciousness on very advanced functional structures relies on a misguided assumption. The assumption is that non-reductive consciousness is unique to humans (and maybe the more intelligent of us) and also that it bestows, by itself, a strong moral status; this we deny.

All the philosophical footwork stressed above is here to provide a background to the physicalist interpretation/s of the first-person stream of awareness, relied upon in the final argument.

3. STRAWBERRIES AND CREME

From now on, this paper focuses on robots enjoying (or not) some *strawberries and crème*; this is of course reference to the large topic of phenomenal qualia, as well as the topic of first-person experience by robots.

We explore, briefly, Mori's *uncanny valley* and the case of *Church-Turing lovers* (Boltuc, 2017B; 2011). This leads to the argument how future robots may perhaps have non-reductive consciousness, based on physicalist grounds. This requires a jump from "creature consciousness" as a bio-psychological term, to creature consciousness as an advanced chemical feature (Sloman, 2020), not limited to "mentations" or similar specificities of biological life (*contra* Searle). Carbon based chemistry, in accordance with

bioengineering, trends as a fruitful mix with inorganic chemistry and physical sciences broadly understood.

*

It is "idiotic to make a computer enjoy strawberries and crème," as Alan Turing pointed out (Turing, 1950). This statement sounds right in Turing's famous article but only because computers known to Turing would have no way to get *the feel* of anything.

3.1. Spooky enjoyments

Some of the robots today can exhibit behaviors consistent with enjoyment, which is often viewed as spooky (Mori, Boltuc, 1998; 2017; 2021). This attitude is consistent with *the uncanny valley effect*, first highlighted by Mori. It is a situation in which artifacts (e.g. dolls, robots, even pictures of human beings or animals), under certain conditions, may be perceived as spooky. Those that are not overly humanoid may be viewed as pleasant, often cute (some shape, color, size conditions apply). Those that are very good representations of humans, e.g. realistic—though often not *hyper-realistic*—portraits or sculptures representing human beings, are often quite accepted (unless they have identifiable content- or form-based triggers of negative feelings). However, pictures, sculptures or dolls that are humanoid, yet not quite good enough as human representations, are perceived as spooky or *uncanny*. This is especially striking if such artefacts encroach on privacy or sexuality—if so, they are likely to be viewed as *disgusting*; if they are large and overpowering—they are viewed as dangerous (Boltuc, 2021). Humanoid dolls that visibly enjoy the strawberries and crème, would probably be spooky—if they did so well enough to mimic human expressions of culinary enjoyment, but not quite well enough to be truly realistic human or animal representations.

There is a rather philosophical, or psychological, explanation why such enjoyment by robots would be spooky, unless it was very tastefully represented. Spookyness in the instance of a robot ostensibly enjoying strawberries and crème—not to mention erotic pleasures—seems to come from the obviousness of the fact, that they do not experience such enjoyment, thus lacking the proper causal chains from enjoyment to its behavioral expression.

Evolutionarily, if someone fakes their feelings, the attempt often masks insincere, hostile, or exploitative intentions. There are good reason why faking one's feelings may look strikingly spooky, since it may in fact be dangerous.

While working on this topic, I came upon the second uncanny valley—the valley of perfection. Human being comprise an inherent doze of imperfection, even when they perform the tasks that are well defined by the stand-

ards of efficiency or of an art (e.g. a required "program" within competitive ice-skating). The non-spooky humanoid artefacts must also be human-like also in terms of *our perfect imperfections*. Thus, a robot being too efficient or too "perfect" in imitating human behavior would also be "uncanny"—that is the second uncanny valley, *the uncanny valley of perfection* (Boltuc, 2011, 2017, 2021).

Recently, Ben Goertzel pointed out that human actions, practices and even ethical values are paraconsistent. Thus, a humanoid companion—or a robot meant to cooperate with humans smoothly (instead of patronizingly) would need to follow largely a paraconsistent logic (Goertzel, 2021a; 2021b; 2021c). This paraconsistency may be construed as a way to avoid the uncanny valley of perfection

3.2. Church-Turing Lovers revisited

The Church-Turing Lovers[9] is a rather clear case where we have good reasons to care whether an agent has or lacks the first-person feel. It involves advanced artificial companions, which perform many (or all) functions that a human companion, or a significant other, would perform. They even look the same (at the right level of granularity), function in a society, may even make money and partake in reproductive process (at least as a surrogate parent)—thus they are functionally equivalent to socially astute humans. However, while they have functional, human level consciousness, they lack the first-person feel what the world looks and feels- like for them. In general, we have overwhelming reasons to believe that they lack first-person awareness. Would one have reasons to care whether his or her significant other is a Church-Turing Lover or a human being?

If one is supposed to care about the feelings, in particular, the inner feel of one's significant other—for non-instrumental reasons—this would be a *futile task* if she or he had no first-person stream of awareness. The meaning of interrelationship with another human consciousness would be lost.

Such relation with a machine may be *functionally* satisfying. However, there would be no real *I-Though* relationship (Bubber). This is why, we should care about one's partner having their awareness, not just a chain of useful functions, including memory, recall and so on.

One other philosophical point needs to be made. If first-person awareness is the sole relevant difference (different constructions of a human and a robot will be viewed as functionally irrelevant at the desired level of granularity), and if first-person awareness so understood has no influence on actions or other functions of a robot, this would lead to epiphenomenalism (roughly, irrelevance of awareness).

---

[9] The name comes from David Deutsch's physical interpretation of the Church-Turing Thesis.

Yet, the very knowledge, or justified suspicion of one's significant other that there is nothing it is like for one's companion to feel their love and other personal experiences, this would create a non-epiphenomenal (at least in the sense of *not-futile*) reason to care about their first-person awareness or first-person consciousness. Thus, if one wants to avoid epiphenomenalism within functionalist determinism (thus, irrelevance of first-person consciousness) one has a reason for doing so based on the Church Turing Lovers and their inability for first-person consciousness. This is based, at the very least, on the fact that people cannot justifiably care about your feelings. which is substantial to many meaningful relationships.

In Section 4 we argue that future robots should be able to have first-person consciousness.[10] In the current subsection, I have demonstrated one reason to believe that such attempt is not quite frivolous, futile or unreasonable.

3.3. Let the Robots Enjoy their Strawberries!

A merely functionalist (soft-AI) approach to first-person consciousness is based on imitating, or faking, the true, aware feelings, such as lack of first-person input, which is evolutionarily unnatural. Back to the uncanny valley effects, we may repeat that if somebody clearly fakes their deep emotions, e.g. love or joy, it makes sense to view such behavior as spooky, since they may cover insincere, even hostile intentions.

Here comes the question I entertain in my early article on the Turing's strawberries [Boltuc 1998B]. Isn't it possible for a robot to *actually enjoy* strawberries and crème? This would require many things, such as sensory equipment fit for tasting a dish, coded as a positive (and specific to a given taste) motivators, with relevant, sort of semantic, phenomenal content.

Kevin O'Regan's sensorimotor approach to ontology and its practical consequence—sensory substitution[11]—may perhaps pave the way to at least theoretical possibility of such enjoyment. Since we can play with various functionalities both of brains and sensorimotor apparatus, there may be some elbowroom for combining artificial or bioengineered elements in an artificial agent.

Contra O'Regan, phenomenal content of first-person consciousness also requires much more—a stream of first-person epistemic awareness, or, to

---

[10] Strong objections of Tom Metzinger (IJMC) to risk constructing machine first-person consciousness when we do not quite know what we are doing are duly noted. Yet, they belong to a different discussion. Briefly, at some point we are likely to know what we are doing in constructing those; then and only then my point on first-person consciousness for advanced AI would be justifiably practical.

[11] Sensory substitution is to change stimuli of one sensory modality into stimuli of another sensory modality. For instance, people with damaged eyes have visual experiences re-transcribed as music (that identifies various colors and shapes) that goes to visual cortex. After re-learning, they "see" through the sounds.

put it in more "Continental" terms, it requires a *locus of consciousness and permanence* [Shalom].

Summing up: The above discussion is to show the difference between functional consciousness defined in the third-person perspective and first-person awareness (always first-personal). The former is similar to Dennett's *agential stance* with a machine acting as if it was conscious like a human being. The latter may or may not be functionally relevant (if not, it is epiphenomenal, which would not be very attractive). Under normal circumstances, it is functionally relevant, although it may function like a catalyst, not quite an agent *tout cours*.

Below, I briefly explore some of the consequences of non-reductive consciousness for robots and other advanced AI agents.

4. THE ENGINEERING THESIS REVISITED

The above philosophical refection brings us closer to addressing the point: What kind of machine would have non-reductive consciousness, like humans do? The answer is: the ones with *creature consciousness*. *Puzzling, is not it? Is not creature consciousness a biological feature of some animals?* Yet, are not biological animals physicalist creatures?

4.1. Robots with creature consciousness

The engineering thesis in machine consciousness is a claim that, within a physicalist framework, we should expect all things to be prone to physical explanation (if not currently, then in the future, as science develops). First-person consciousness is a real phenomenon, even though perceived by every individual separately. Such testimonials are confirmed by neuroscientific observation. If first-person consciousness is real in in the physicalist universe, then—in principle—one should be able to re-engineer it (Boltuc, 2009; 2012). This seems to follow from the David Deutsch's physical interpretation of the Church-Turing thesis (Deutsch 1985).

By proposing, in the current article, the hypothesis that the emergence base (or *locus*) of first-person stream of awareness is *creature consciousness*—we make a conceptual step towards defining the kind of consciousness that would be expected in any first-person aware beings. Some recent papers argue that such consciousness may belong to the parts of neurons, plants or even fungi.

This is not the endorsement of panpsychism (though I explored this option in (Boltuc, 2010)). Panpsychism requires a hypothesis that conscious-

ness is sort of like a physical substance (say, moisture) that is in the air on Earth, but in some places, there is more of it, leading to clouds, rain, puddles, lakes and so on. For consciousness, we would require nervous systems and brains—maybe even spiritual beings—that are the locations of large amounts of well-structured consciousness.

For now, until there are strong reasons to the contrary, we view consciousness as an emergent property of neural substrates; however, we see no reason to assume its ubiquitous nature. The difference is philosophically important since panpsychism assumes that consciousness is one of the substances in the universe and the other is matter, which results in dualism. Chalmers' panpsychism and earlier forms of spiritual monism (Baruch Spinoza, Gottfired Wilhelm Leibniz, George Berkeley) may claim that everything is consciousness, or at least has a conscious aspect; this is a monistic non-materialist system—sometimes idealism like (Berkeley's) and sometimes neutral monism (Spinoza's). Those big-picture systems seem to neglect the options open within non-reductive physicalism, which is much better confirmed by our experiences.

Organic chemistry is a natural science just like non-organic chemistry—bioengineering works for both kinds. If we can bioengineer creature consciousness that is not a clone or permutation of an animal brain but uses other techniques, and if it is an "organic" part of an entity capable of carrying advanced AI and central to that task, we should be open to the idea that the robot would be *prima facie* conscious. It would be a potential carrier of first-person conscious capabilities, such as phenomenal experiences, and their precondition: first-person awareness. That would be a first-person conscious bio-electronic system.

We may also think of a non-carbon-based system doing the same. But understanding creature consciousness based on non-organic chemistry would involve a much more complex process of discovery, belonging to more remote future. Except, if AGI capable of multiplying human intellectual capacity becomes a reality and gets to work, among the many other things, also on the project to create *creature consciousness* in non-standard chemical substances.

The theoretical problems concerning such an organic (and later maybe inorganic) creature, would be both: to estimate what substances, and in what structures, would likely be first-person conscious, and then the need to define what is and what is not a new being (as opposed to biological cloning and other relatively standard kinds of reproduction). It is important for organic chemistry research to determine whether the organic brain-like structure is a part of machine, or results in a cyborg with an implanted (and no doubt re-engineered) animal or hybrid animal-human brain.

The Engineering Thesis in Machine Consciousness tells us that, within naturalism, at some point we would be able to build non-reductive con-

sciousness in a machine. Moral and practical aspects of such project belong to different papers (including Boltuc 2018B)—while the option may be scary it is also deeply enticing.

REFERENCES

N. Block, *On a Confusion about a Function of Consciousness*, Brain and Behavioral Sciences, 1995, 18 (2), 2FFH27–247.

P. Boltuc, T. P. Conelly, *Uncanny Robots of Perfection*, in: Brain-Inspired Cognitive Architectures for Artificial Intelligence, Gudwin R. et al. (eds.), BICA*AI 2020, Springer Science, 2021, pp. 56–68.

P. Boltuc, *Subject Is No Object*, in: Epistemic Basis of Information, M. Burgin; Gordana Dodig-Crn fhkovic (eds.), Philosophy of Information, World Scientific, 2019, pp. 3–39.

_____ , *Strong Semantic Computing*, Procedia Computer Science, 123, Feb. 2018A, pp. 98–103; https://www.sciencedirect.com/science/article/pii/S1877050918300176

_____ , *Cognitive Agents: Is There a Moral Gap Between Human and Artificial Intelligence?*, in: Information, Communication and Automation Technology Ethics in the Knowledge Society, Tzfestas S. (ed., NOVA Science Publishers (2018B).

_____ , *Church-Turing Lovers*, In: Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence; P. Lin, K. Abney, R. Jenkins (eds.) Oxford University Press: Oxford, UK, 2017; pp. 214–228.

_____ , *Non-reductive Consciousness as Hardware*, APA Newsletter on Philosophy and Computers, 2015, p. 14 (2, Spring).

_____ , *The Engineering Thesis in Machine Consciousness*, Techné: Research in Philosophy and Technology, 16 (2, Spring), 2012, pp. 187–207.

_____ , *What is the Difference between Your Friend and a Church Turing Lover*, The Computational Turn: Past, Presents and Futures? C. Ess; R. Hagengruber Aarchus University, 2011 pp. 37–40.

_____ , *A Philosopher's Take on Machine Consciousness*, in: Philosophy of Engineering and the Artifact in the Digital Age, V. E. Guliciuc (ed.), Cambridge Scholar's Press, 2010, pp. 49–66.

_____ , *Qualia, Robots and Complementarity of Subject and Object*, World Congress of Philosophy, Boston 1998B; http://www.bu.edu/wcp/Papers/Mind/MindBolt.htm

_____ , *Reductionism and Qualia*, Epistemologia, 4, 1998A, pp. 111–130.

M. Bubber, *I and Thou*, W. Kaufmann (trans.), Charles Scribner's Son, New York 1970.

D. Chalmers, *Panpsychism and Panprotopsychism*, Amherst Lecture in Philosophy, 8, 2013.

D. Davidson, *Rational Animals*, in: Actions and Events: Perspectives on the Philosophy of Donald Davidson, E. Lepore, B. McLaughlin (eds.), Basil Blackwell, New York 1985.

D. C. Dennett *The Intentional Stance*, MIT Press, 1981.

D. Deutsch, *Quantum Theory, the Church–Turing Principle and the Universal Quantum Computer*, Proceedings of the Royal Society, 400 (1818), 1985, pp. 97–117.

W. Dilthey, *Introduction to the Human Sciences*, R. Makkreel, F. Rodi (eds.), Princeton University Press, Princeton, NJ 1989 (1883).

Princess Elisabeth of Bohemia, René Descartes, *The Correspondence between Princess Elisabeth of Bohemia and René Descartes*, L. Shapiro (ed., trans.), University of Chicago Press, 2007.

B. Goertzel, *Exploring Open-Ended Intelligence Using Patternist Pilosophy*, At: IS4SI, Philosophy and Computing, Sept 14, 2021; https://www.youtube.com/channel/UCQ3w2Jpi6DQf9aK51A0LLFA

_____ , *Paraconsistent Foundations for Probabilistic Reasoning, Programming and Concept Formation*, ArXiv abs/2012.14474, 2020.

_____ , *The Hidden Pattern. A Patternist Philosophy of Mind*, Brown Walker Press (FL), United States, 2006.

G. Harman, *Explaining an Explanatory Gap*, APA Newsletter, 6 (2), Spring, 2017.

_____ , *Reasoning, Meaning, and Mind*, Oxford University Press, 1999.

_____ , *Immanent and Transcendent Approaches to Meaning and Mind*, in: Perspectives on Quine, R. Gibson, R. B. Barrett (eds.)., Oxford: Blackwell, 1990; reprinted in: G. Harman, *Reasoning, Meaning, and Mind*, Oxford: Oxford University Press, 1999.

_____ , *Can Science Understand the Mind?*, in: Conceptions of the Human Mind: Essays in Honor of George A. Miller, edited by G. Harman. Hillsdale, NJ: Lawrence Erlbaum, 1993, vol. 4, Action Theory and Philosophy of Mind (1990), pp. 31–52.

E. Husserl, *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy* 1913—First Book: General Introduction to a Pure Phenomenology, trans. F. Kersten. The Hague: Nijhoff 1982.

D. Kahneman, A. Tversky, *Choices, Values, and Frames*, American Psychologist, 39 (4), 1984, 341–350; https://doi.org/10.1037/0003-066X.39.4.341

D. Kelley, *Preliminary Results and Analysis of an Independent Core Observer Model (ICOM) Cognitive Architecture in a Mediated Artificial Super Intelligence (mASI) System*, Updated: AGIL v10, BICA Preconference Proceedings, 2019; https://www.springer.com/us/book/97830 30257187.

R. W. Lurz, *Advancing the Debate Between HOT and FO Accounts of Consciousness*, Journal of Philosophical Research, 28, 2003, pp. 23–44.

M. Mori, *The Uncanny Valley*, IEEESpectrum, 12, JUN, 2012; https://spectrum.ieee.org/the-uncanny-valley.

T. Nagel, *The View from Nowhere*, Oxford University Press, 1986.

_____ , *Mortal Questions*, Cambridge University Press, 1979.

D. Parfit, *Reasons and Persons*, Oxford University Press, 1986.

K. O'Regan, *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*, OUP 2011.

D. M. Rosenthal, *Explaining Consciousness*, in: D. Chalmers PHILOSOPHY OFMIND Classical and Contemporary Readings, Oxford University Press, New York–Oxford 2002, pp. 406–421.

B. Russell, *Analysis of Mind*, George Allen and Unwin, London; The Macmillan Company, New York 1921.

B. Russell, *Analysis of Matter*, Kegan Paul, London; Trench, Trubner, Harcourt, Brace, New York 1927.

T. Schlich, A. Newen, *Kant and Cognitive Science Revisited*, History of Philosophy & Logical Analysis 18 (1), 2015, pp. 87–113; DOI: 10.30965/26664275-01801008.

J. R. Searle, *Intentionality: An Essay in the Philosophy of Mind*, Cambridge University Press, 1983.

A. Shalom, *Body/Mind Conceptual Framework and the Problem of Personal Identity: Some Theories in Philosophy, Psychoanalysis and Neurology*, Atlantic Highlands, 1985.

A. Sloman, *Varieties of Evolved Forms of Consciousness, Including Mathematical Consciousness*, *Entropy*, 22 (6), 2020, 615; https://doi.org/10.3390/e22060615

S. L Sorgner, *Transhumanism: The Best Minds of Our Generation Are Needed for Shaping Our Future*, The American Philosophical Association Newsletter on Philosophy and Computers, 18 (2), 2019, pp. 15–18; https://cdn.ymaws.com/www.apaonline.org/resource/collection/EADE8D52-8D02-413 6-9A2A-729368501E43/ComputersV18n2.pdf

P. F Strawson, *Individuals. An Essay in Descriptive Metaphysics*. Routledge, 1959.

A. M. Turing, *Computing Machinery and Intelligence*, Mind, 59 (236), 1950.

M. Velmans, *Preconscious Free Will*, Journal of Consciousness Studies, 10 (12), 2003, pp. 42–61.

ABOUT THE AUTHOR — Professor, Double Doctor (Bowling Green State University (Applied Ethics); The University of Warsaw (Philosophy of Person), Professor at the University of Illinois at Springfield, USA (Philosophy; Computer Science); The Warsaw School of Economics (Management Theory).

Email: pboltu@sgh.waw.pl