A sketch of a Kripkean theory of consciousness

Federico Zilio

In this paper, I will propose a provisional blueprint of the notion of consciousness. I will start an analysis of the notion from the way we generally use the term "consciousness" in our ordinary language. In this regard, I will use Saul Kripke's direct reference theory to define the term "consciousness" in a non-descriptive way, that is, interpreting it as a rigid designator. Then, I will critically discuss the idea of a necessary a posteriori relationship between consciousness and brain activity, arguing instead that consciousness is intrinsically related to the concept of subjectivity.

1. Consciousness is said "in many ways"

The concept of consciousness is ambiguous and complex, composed of variable meanings but devoid of precise definitions; nevertheless, at the same time, we have a natural and pragmatic acquaintance with it. Apart from the use of the notion as an umbrella concept in the philosophy of mind and, more recently, in neuroscience, "consciousness" refers *prima facie* to the simple «states of sentience or awareness that typically begin when we wake up in the morning from a dreamless sleep and continue throughout the day until we fall asleep again»[1]. The word "consciousness" represents both the most familiar sensation of being aware of the surrounding world and one of the most evanescent and opaque conceptual issues of contemporary philosophy and science. In the words of David Chalmers: «[t]here is nothing that we know more intimately than conscious experience, but there is nothing that is

[1] J. R. SEARLE, *Mind, language and society: philosophy in the real world*, Weidenfeld & Nicolson, London 1999, pp. 40-41.

harder to explain»[2]. That recalls the problem of time presented by Augustine: «What then is time? If no one asks me, I know; if I want to explain it to a questioner, I do not know»[3].

Consciousness does not have a univocal definition, indeed the notion refers to a broad set of meanings[4]: sentience, i.e. the ability to sense and respond to the environment; wakefulness, i.e. the awake and alert state in contrast to sleep, coma or other abnormal states; interaction with the external and inner world; self-consciousness, i.e. the epistemic state of being aware that we are aware; the "what it is like to be" of the experience and the phenomenal content, i.e. those specific subjective characters of being in a particular conscious state[5]; intentionality or transitive consciousness, i.e. the fact that consciousness is always and necessarily "consciousness of something"; awareness of surrounding objects and reportable contents. These are only some of the meanings of this notion. Ram Vimal describes and lists forty distinguishable and non-exhaustive meanings attributed to "consciousness" from the literature and various cultural traditions[6].

For a long time, the study of consciousness has been the exclusive concern of philosophy; however, what had long been an exclusive matter for the theoretical and conceptual analysis of philosophy has progressively become an object of experimentation and empirical consideration. For this reason, consciousness seems to be also a potential subject of scientific study, thanks to new methodologies, disciplines, and technologies. The enormous increase in neuroscientific research about consciousness can be taken to support this[7]; in the last decades, besides focusing on perceptual and cognitive abilities – and impairments – of the brain, neuroscientists also began to deal with the concept of consciousness, trying new quantitative measurements within the empirical and experimental domain of cognitive sciences. Consequently, new concepts have emerged, such as visual consciousness, perceptual consciousness, level and content of consciousness, arousal and cortical

---

[2] D. J. CHALMERS, *Facing up to the problem of consciousness*, «Journal of Consciousness Studies», 2, 3/1995, p. 200.

[3] AUGUSTINE, *The confessions*, in *Masterpieces of Philosophical Literature*, edited by T. L. Cooksey, Greenwood Press, Westport 2006, p. 242.

[4] R. VAN GULICK, *Consciousness*, in E. N. ZALTA (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University 2018.

[5] T. NAGEL, *What Is It Like to Be a Bat?*, «The Philosophical Review», 83, 4, 1974, p. 435. N. Block, *On a confusion about a function of consciousness*, «Behavioral and Brain Sciences», 28, 2/1995, pp. 227-247.

[6] R. VIMAL, *Meanings Attributed to the Term «Consciousness»: An Overview*, «Journal of Consciousness Studies», 26, 5/2009, pp. 9-27.

[7] Medline trend for the term "Consciousness" (year: number of citations): 1998: 814; 2008: 1426; 2018: 2510.

awareness, altered forms of waking consciousness (trance, absorption, hypnosis, dissociation, meditative states, drug states, and out-of-body experiences), REM and dreaming versus slow-wave/deep sleep states, etc.[8]

1.1 What starting definition for consciousness?

Is there a definition that can be used as a starting point for a theory of consciousness? It seems quite problematic to give a univocal answer to this question, for many reasons. First, there are various perspectives (first-, second-, third-person perspectives) and approaches (methodological, ontological, epistemological, phenomenological, etc.), on which the theories of consciousness are based. Second, each theory of consciousness refers to an ontological and metaphysical commitment about what consciousness is, which can compromise any attempt to offer a complete description of its nature; at the same time, it seems impossible to avoid any – at least provisional – ontological commitment. Third, even within any single approach, there are sub-categories of definition, depending on the kind of method used (as in cognitive sciences) or on the variability of the meanings (as in phenomenology, e.g. for-me-ness, me-ishness, me-ness, myness or mineness, etc.[9]). Fourth, almost all of the definitions above-mentioned can be considered analytic a priori propositions of the concept of consciousness, i.e. they give us some detailed descriptions of how the term "consciousness" is used, without however offering any information not already contained in the term, producing a sort of circularity[10].

So, how can we build a theory of consciousness without too many presuppositions? Of course, this does not mean that we cannot start with a description or a set of descriptions of what we believe consciousness is, we have to start somewhere; nevertheless, any possible description does not work as the synonym of the word "consciousness", that is, the meaning of "consciousness" does not seem to lie in a set of descriptions. So far, I have talked about consciousness in many ways, without giving a description that fixes univocally the reference of the term. In our daily life, we competently

---

[8] B. FAW, *Cutting «Consciousness» at its Joints*, «Journal of Consciousness Studies», 26, 5/2009, pp. 54-67. S. LAUREYS, G. TONONI, *The Neurology of Consciousness*, Academic Press, San Diego 2009. G. NORTHOFF, V. LAMME, *Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight?*, «Neuroscience & Biobehavioral Reviews», 118, 2020, pp. 568-587.

[9] M. GUILLOT, *I Me Mine: on a Confusion Concerning the Subjective Character of Experience*, «Review of Philosophy and Psychology», 8, 2017, pp. 23-53.

[10] E. SCHWITZGEBEL, *Do you have constant tactile experience of your feet in your shoes? Or is experience limited to what's in attention?*, «Journal of Consciousness Studies», 24, 3/2007, pp. 5-35.

mention and understand a lot of words without the complete knowledge of their conditions or with only a partial description of the referent[11]. Moreover, it seems difficult to describe the meaning of consciousness and avoid any circularity, e.g. consciousness = phenomenal = qualia = what it is like = experience = consciousness. Epistemically speaking, we do not need the knowledge of a complete descriptive content of consciousness to understand the word *per se*. Therefore – and this may seem quite obvious – in order to be acquainted with the referent of the word "consciousness", to refer to consciousness itself, or to be competent users of the word "consciousness" and consequently start an analysis of its conditions and features, we do not need to provide any set of descriptions beforehand, nor we need a descriptive identifying content associated with the word.

Accordingly, rather than introducing an empirical or phenomenological account of consciousness, I would begin with our ordinary language as a provisional starting point for the discussion about its nature. In order to do this, I need to make use of Saul Kripke's direct reference theory, taken from *Naming and Necessity*.

2. Kripke's direct reference theory

Kripke argues that names and natural kinds are not synonymous with the descriptions associated with them. A description can initially fix the reference of a name or kind, but it does not give us the meaning of the term and we do not need to know it to understand that meaning (as often happens within our everyday vocabulary); for example, the description "the clear, thirst-quenching liquid" does not indicate the meaning of the term "water". Instead, some terms designate an object without the need for a mediating meaning between the language and the object. According to Kripke, a term designates x rigidly if the term designates x in every possible world in which x exists and does not designate anything else with respect to worlds in which x does not exist.[12] Kripke principally holds that proper nouns are rigid designators, for example, "Aristotle" designates the same person in every possible world, while the description "Aristotle is the most important Greek philosopher" is not an analytic statement related to Aristotle – as the descriptivists would sustain (Frege, Russell, Wittgenstein) – rather it is a

[11] See, for instance, Kripke's example of Cicero and Feynman. S. KRIPKE, *Naming and Necessity*, Harvard University Press, Cambridge 1980.

[12] N. SALMON, *Are General Terms Rigid?*, «Linguistics and Philosophy», 28, 1/2005, pp. 117-134.

contingent one. Therefore, "the most important Greek philosopher" is not a synonym of Aristotle, because it is not necessary in every possible world, e.g. in some possible world, Aristotle could have been a farmer.

Also, certain general terms – including natural-kind terms, like ''water'', phenomenon terms, like ''heat'', and biological taxa, like "tiger" – are rigid designators, e.g. the propositions "water = H_2O" and "heat = molecular motion" are necessary in every possible world[13]. These identifications may be mistaken, e.g. perhaps scientists will discover that water is made of a new chemical element, in fact, the statement "water is H_2O" cannot be determined a priori; however, this would not affect Kripke's theory, which states that *if* "water is H_2O" is true, then water is necessarily H_2O. Thus, although our knowledge about water may not be complete, this will not change the fact that, if it is true, water is necessarily H_2O.

Kripke also gives an example with gold: if it will turn out that gold – perhaps due to a sort of hallucination – is not actually yellow but blue, this will not mean that gold never existed, but simply that gold has turned out not to be yellow, but blue. This is because the description "being yellow" is not rigidly applied to gold, e.g. in another possible world it might not apply to gold but to other things, therefore, the property "being yellow" for gold is not necessary in every possible world, so it can only represent a contingent identity in this world[14]. Instead, Kripke argues that "gold is the element with atomic number 79" represents a theoretical – i.e. not merely contingent – identity, involving two rigid designators of two different kinds, respectively the term "gold" from the common-sense and "atomic number 79" from the scientific natural kind. This means that the statement "gold = element with atomic number 79", if true, is necessary in every possible world, i.e. we cannot conceive the term "gold" referring to something different from that metal with atomic number 79, just as "water" with "H_2O" or "heat" with "molecular motion". Hence, if we find a metal that looks exactly like gold but is a different substance – e.g. iron pyrite – it would not represent a variance of gold, but a completely different thing.[15]

Now, we can consider another possible world similar to the actual one but different in respect of how some things are, like gold or heat. In this possible world, we can imagine that gold turned out to have an atomic number different from 79 (or that heat turned out not to be molecular motion).

---

[13] Note that here the metaphysical concept of "necessity" is not the same as the gnoseological concept of "a priori". Stating that "*x* is necessarily true" means that x is true in every possible world, while stating that "*x* is true a priori" means that x is true independently from experience. See S. KRIPKE, *Naming and Necessity*, cit., p. 38.

[14] S. KRIPKE, *Naming and Necessity*, cit., p. 118.

[15] Ivi, pp. 118-119, 124-125.

However, according to Kripke, these theoretical identities, if they are true, are always necessarily true, even a posteriori, i.e. derived from empirical discovery. Therefore, a proposition that seemed to be conceivable – e.g. "gold is a compound rather than an element with atomic number 79" – turns out to be impossible on closer inspection. As Kripke claims, it «should be replaced (roughly) by the statement that it is logically possible that there should have been a compound with all the properties originally known to hold of gold»[16]. The term "gold" has been initially baptised as "the metal with the atomic number 79" and is a rigid designation, that is, it designates the same object in all possible worlds, just like the proper noun "Aristotle" refers to the same person in every possible world, regardless of the possibility that Aristotle might have been a farmer, a king or a philosopher. This implies the existence of "necessary a posteriori" statements[17]; for example, if true, the proposition "gold has the atomic number 79" is a necessary statement and it is also a posteriori, as it comes from the empirical findings concerning the substance we initially baptised as "gold".

To sum up, Kripke argues that the semantic content of a term is nothing more than its referent. This does not imply that any set of descriptions of the term is useless or wrong, but that it does not indicate the meaning of the term. On the contrary, the term *per se* designates the object, and the semantic content of the object is revealed by the referring term or expression, without the need for any supplementary description.

2.1 A Kripkean-like definition of consciousness

What about the term "consciousness"? Can we apply the direct reference theory to this notion? Consciousness is not a proper noun, but someone, like Searle, might consider it as a natural kind or a biological phenomenon, like digestion, photosynthesis, mitosis, etc.[18] For the aim of this paper, it seems reasonable to interpret it as a general term[19]. As said, it is problematic to start the analysis of consciousness through a set of descriptions from the empirical or conceptual domains, so the point here is: can we say something about consciousness in a Kripkean-like way? I say "Kripkean-like" because I do not mean to apply *verbatim* Kripke's direct reference theory, but to use it as a heuristic tool for avoiding the murky waters of the ontological and epistemological debates between dualism and monism, first- and third-

[16] Ivi, pp. 142-143.

[17] Ivi, p. 140.

[18] J. R. SEARLE, *The Rediscovery of the Mind*, MIT Press, Cambridge 1992. J. R. SEARLE, *Il mistero della realtà*, Raffaello Cortina, Milano 2019.

[19] N. SALMON, *Are General Terms Rigid?*, cit.

person perspective, a priori and a posteriori, etc. Moreover, instead of interpreting the direct reference theory as a rigorous foundation of scientific essentialism[20], I rather use it as an attempt to define the concept of reference in an intuitive, pre-theoretic way; in other words, as a way to explain how names work according to common-sense. This would be a useful starting point from which one can go through with the investigation of the concept of consciousness, refining it with various approaches (phenomenological, neuroscientific, etc.).

The first questions are: Is the term "consciousness" a rigid designator? Can this term rigidly designate something with respect to every possible world? Generally, we use "consciousness" to refer to some behavioural and psychological state; for example, taking the expression from Searle, we refer to those states «of sentience or awareness that typically begin when we wake up in the morning from a dreamless sleep and continue throughout the day until we fall asleep again»[21]. We can determine if someone is conscious or has lost consciousness and we can do it pragmatically, without possessing specific philosophical, psychological, or medical knowledge. Of course, we can be wrong in some of our determinations, e.g. in case someone is pretending to be asleep, or in severe situations in which it is necessary to use clinical scales to assess the level of consciousness (coma, unresponsive wakefulness syndrome, minimally conscious state, etc.). Nevertheless, it seems impossible to be deceived about our own conscious state when I refer to myself as "conscious" at a given moment; I can be deceived about the notebook in front of me which I am conscious of, but not about the fact that I am conscious of something, even in virtual reality or during a hallucination[22]. Moreover, some situations superficially resemble a conscious-like state but they are not related to consciousness, and, conversely, other situations superficially resemble a non-conscious state, but they effectively represent a conscious state. So, following the analogy with natural kinds, consciousness may present some superficial features that are contingent, just like gold being yellow (as well as iron pyrite) or water being colourless (as well as vodka).

So far, I have discussed the notion of consciousness in a very general way; now, we need to go deeper into what the term consciousness means to us. In our daily life, we do not deal with an empty and abstract concept of consciousness, but with a full-fledged and lived experience of the world.

---

[20] The position that some internal structures are the ontological and necessary conditions for an object to be part of a given natural kind.

[21] J. R. SEARLE, *Mind, language and society*, cit., pp. 40-41.

[22] D. J. CHALMERS, *The virtual and the real*, «Disputatio», 9, 46/2017, pp. 309-352.

First of all, there is never consciousness as such, rather consciousness is always and intrinsically "consciousness of something"; if I am aware, I must be aware of something, and this "something" can possess different epistemic forms, e.g. it can be an object of perception or imagination, a pure sensation or an abstract thought, etc. This intrinsic aboutness or directedness towards an object that characterises consciousness is called "intentionality" (from the Latin *tendere*: directedness towards, attending to, referring to something[23]). Thus, when we perceive, believe, think, imagine, etc. we are perceiving, believing, thinking, imagining "something".

There is a very large debate about this concept, in particular, whether intentionality could be considered the essential "mark of the mental" or only a necessary condition for mental states/events[24], whether conscious states are intrinsically intentional[25] or some of them can be non-intentional (like nervousness or anxiety[26]), or even whether intentionality as the typical element of mental states[27] or as a feature of phenomenal consciousness[28] should be rejected in contemporary philosophy of mind. I cannot address here these issues about the nature and legitimacy of the concept of intentionality. For the moment, it should be sufficient to acknowledge that our conscious states are characterised by different experienced objects, so that it seems reasonable to say that our consciousness, in a broad sense, consists of a variety of "consciousness of something", at least in the way we commonly intend the term consciousness, e.g., "I have consciousness of this chair", "I am (consciously) perceiving an apple", "I am conscious of the consequences of my act", etc. To use an analogy, we may argue that in our phenomenal life we instantiate forms of "consciousness-of-something" that fall under the class of consciousness, just as the terms "Siberian tiger", "white Bengal tiger" and even the "Bali tiger" (extinct) represent subspecies that refer to the same species of the tiger (*Felis Tigris*).

---

[23] P. JACOB, *Intentionality*, in E. N. ZALTA (ed.), *Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University 2019.

[24] T. CRANE, *Intentionalism*, in A. BECKERMANN, B. P. MCLAUGHLIN, S. WALTER (eds.), *The Oxford Handbook of Philosophy of Mind*, Oxford University Press, New York 2009, pp. 474-493.

[25] J.-P. SARTRE, *Une idée fondamentale de la phénoménologie de Husserl: l'intentionalité*, «La Nouvelle Revue française», 304/1939, pp. 129-131.

[26] J. R. SEARLE, *Intentionality: An Essay in the Philosophy of Mind*, Cambridge University Press, Cambridge 1983.

[27] A. VOLTOLINI, *The Mark of the Mental*, «Phenomenology and Mind», IV, 2013, pp. 124-136.

[28] H. ROBINSON, *Why phenomenal content is not intentional*, «European Journal of Analytic Philosophy», V, 2/2009, pp. 79-93.

This last example may be helpful to analyze another issue concerning the potential use of the word "consciousness" as a rigid designator. The term "tiger" refers to different species and instances of tiger, regardless of some contingent variations, e.g. colours, anatomical defects, different sizes and weights, etc. For instance, we could see a white pet tiger, deprived of the tail or a leg, but none of these characteristics would induce us not to refer to it as a tiger. This because the term "tiger" rigidly refers to its internal structure as the mark of the species of tigers, e.g. the DNA, and not merely to the appearances of the tiger as a big cat. Interestingly, we also are keen to say that an animal that apparently differs from a tiger, but that has the same internal structure of a tiger, actually is a tiger, while a seeming tiger with the internal structure of a reptile is likely to be considered a reptile.[29] This does not mean that the old appearance-based concept of a tiger has to be replaced by the scientific one, but we may suppose that the internal structure of the tiger can determine the criterion for the natural kind of "tiger"; also, the determination of the natural kind by scientific knowledge is an a posteriori investigation. So, "tiger" is a rigid designator, which means that we do not understand it based on some identifying definition or a cluster of descriptions, rather on the internal structure of the tiger. Thus, we can track tigers across counterfactual situations, completely independent of any contingent characteristics, e.g. tigers without legs or tail, or strangely coloured tigers. Moreover, some other species could have had the external appearance of tigers, e.g. some other cat-like animal with orange fur and black stripes, etc., however, our term "tiger" picks animals out as they are in the actual world, and this is what fixes the species across possible worlds.

If the analogy holds, we understand the meaning of "consciousness" not necessarily through any cluster of descriptions; rather we can use "consciousness" across extremely different situations (consciousness of pain, of pleasure, of a red apple) without fixing any particular definition of "being conscious". When we say "consciousness" we directly refer to the act of intentionality towards an object, i.e. directedness towards something. Moreover, it seems that our general term "consciousness" can pick instances of consciousness in the actual world, working as a referent in every possible world, but we can even conceive a possible world in which, for example, consciousness is revealed through a different cluster of behaviours (closed eyes, little movements, no observable reactions) or different physical activities (neural activity in the cerebellum, without any significant activity in the cerebrum, activity in other organs, like the liver, or even activity in an electronic circuit board), in which, nevertheless, we would still necessarily

[29] S. KRIPKE, *Naming and Necessity*, cit., p. 121.

refer to consciousness as that process of intentionality/directedness towards something. Following the analogy, as "tiger" refers to those animals with the DNA of the *Felis Tigris* species (with its intraspecies differentiation), similarly "consciousness" refers to those acts of intentionality towards objects, i.e. to bear a relationship with some object (with its phenomenal differentiation, e.g. perceived, imagined, hallucinated objects, etc.). Taken together, I would consider "consciousness" as a rigid designator because, in every possible world where consciousness exists, it necessarily refers to the act of being directed towards – or being about – something. In other words, if true, the proposition "consciousness = being directed towards something" is necessary a priori in every possible world.

To prove this point, we can imagine a situation similar to that presented by Kripke concerning the difference between gold and iron pyrite. Suppose that it turns out that our experience is nothing but a deceptive illusion, as some illusionists argue[30], that our phenomenal world is not full of sounds, colours, sensations, and fluxes of thought, and also that our conscious memories are completely fake a posteriori reconstructions; does it mean that there is no consciousness? We would say that, due to our deceptive intuitions, consciousness is not what we think it is, but we would not say that consciousness *per se* does not exist; indeed, consciousness as intentionality, is preserved despite the illusory content of consciousness. Now, on the contrary, suppose that in a possible world there is a subject who, though having all external appearances of a conscious being (e.g. behaviour), has a completely different structure from the human one (e.g. a synthetic skeleton with synthetic organs, or even better a metal skeleton with gears, pumps and electronic devices). In this case, the assignment of consciousness might be questioned; however, if the android was actually conscious (in some way), then we would refer to it as a being with intentionality (as we would refer to a reptile in front of that seeming tiger with the internal structure of a reptile). Thus, in both imaginary scenarios, the term "consciousness" rigidly refers to the intentional act towards something.

3. A Necessary a posteriori Statement for Consciousness?

So far, I have presented a Kripkean-like direct reference theory, understanding the very general term "consciousness" as a general term, and I claimed that it directly refers in every possible world to the act of

[30] K. FRANKISH, *Illusionism as a Theory of Consciousness*, «Journal of Consciousness Studies», 23, 11-12/2016, pp. 11-39.

being directed towards something, i.e. intentionality. At this point, one may argue that this is only a synonym that simply paraphrases the supposed description of consciousness, while I should offer a reliable statement about consciousness that is at the same time necessary and a posteriori (like H_2O for the notion "water"). This criticism partially misses the point, given that the preliminary understanding of "being directed towards something" is only the semantic extension of the direct reference of consciousness; indeed, this extension is composed of concrete instances of conscious acts towards objects (in the world or imagined), in the same way as the term "water" does not merely refer to the abstract H_2O chemical definition, but to "this" or "that" instance of H_2O, e.g. a glass of water or the water of a river, etc. In other words, saying that "consciousness" refers to being directed towards something does not mean anything if we do not complete it – in an ostensive way – with instances of experience. For example, pain-consciousness directly refers to "that" pain state, as the consciousness of red directly refers to "that" red object (imagined or presented). Thus, "consciousness" does not refer to some general consciousness act, as well as when we mention "tiger", "gold", "water", we refer to instances of tiger, gold and water. So, our general understanding of the meaning of "consciousness" is not directly due to a set of descriptions of the phenomenon, rather it is pragmatically given by the reference to our experiential examples[31].

Now, it must be said that the level of knowledge we possess about water, tiger and gold is more detailed and deeper than about consciousness. Of course, we have a direct acquaintance with what it is like to be conscious, but I mean the set of knowledge about what Kripke calls the "internal structure"[32], e.g. H_2O for "water", the atomic number 79 for "gold", or molecular motion for "heat". I sustain that the internal structure of consciousness is intentionality, but we should also specify what this internal structure is concretely and

[31] «The crucial point is that the word "consciousness," as pretty much everyone uses it, is defined largely by reference to our own example. We don't have access to some separate, human-independent definition of consciousness, which would allow us even to frame the question of whether it's possible that toasters are conscious whereas humans are not. By analogy, imagine 19th-century scientists built a thermometer that delivered the result that boiling water was colder than ice. The possibility that that was true wouldn't even merit discussion—it would be immediately rejected in favor of an obvious alternative, that the thermometer was simply a bad thermometer, since it failed to capture our pre-theoretic notion of what temperature is even supposed to mean, which concept includes boiling water having a higher temperature than ice» S. AARONSON, *Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)*, retrieved November 20, 2020, from: https://www.scottaaronson.com/blog/?p=1799.

[32] S. KRIPKE, *Naming and Necessity*, cit., p. 120.

empirically made of. In other words, the question now is: can we find a necessary a posteriori statement for consciousness?

First of all, instead of "internal structure", which implicitly suggests the idea of something embedded inside, I would prefer using the terms "intrinsic" or "constitutive", that is, something necessary and fundamental, as opposed to something extrinsic, contingent and not essentially related to something. Therefore, I will use the "intrinsic or constitutive structure of consciousness" as those conditions that are necessary for consciousness in all possible worlds.

Following Kripke's theory, if we want to define the intrinsic structure of consciousness, we should not settle for a mere analytic definition of consciousness, but we should consider the possibility of a necessary a posteriori statement, like "water is H_2O". The statement "consciousness is being directed towards something" is not sufficient for this aim, and constantly needs to be fulfilled by instances of conscious states. In this respect, we would need to identify those elements that are necessary for the existence of these various instances of consciousness, in a similar way as H_2O is fundamental for the existence of any instance of water (the water in the cup, the water from the river, etc.). Now, the first question is whether it is possible to find a necessary a posteriori statement for consciousness. Then, the second question will be about the kind of necessary a posteriori truth we can find for consciousness.

3.1 The Consciousness-Brain Relation

The crucial question is: what is the intrinsic structure of consciousness? The proposition "consciousness is being directed towards something" is necessary a priori but is vague, poorly informative, and needs empirical and phenomenological improvement to be considered as a proper necessary a posteriori statement. One might claim that not the whole body but the brain or even a particular neural network can be considered as the rigid designator that is necessary and a posteriori related to consciousness; as "water" is "H_2O", then "consciousness" could be "brain activity *x*". The fundamental presupposition here is that the brain is the necessary and sufficient condition for the emergence of any experiential state, that is, the minimal neural activations that are sufficient for specific contents of consciousness, e.g. the so-called "neural correlate of consciousness", i.e. those neural correlates that mark the difference between presence, absence and alteration of consciousness[33]. However, the strong correlation between specific brain

[33] D. J. CHALMERS, *What is a neural correlate of consciousness*? in T. METZINGER (ed.), *Neural Correlates of Consciousness*, MIT Press, Cambridge 2000, pp. 17-39.

activity and the presence of consciousness does not necessarily imply an identity relation. For example, the statement "consciousness = brain activity *x*" could represent a case of "contingent identity", i.e. two terms differing in sense but identical in reference (we can imagine some possible world in which one side of the identity exists without the other)[34]. Kripke argues against the supposed validity of contingent identity[35], arguing that identity can be only a necessary relation, using the converse of Leibniz's Law, i.e. if a = b, then a and b must share all their properties, otherwise there would be a difference[36].

To explain this, Kripke uses the example of "pain = C-fibers stimulation". According to the physicalist view, this identity is supposed to work in the same way as the proposition "heat = molecular motion". All the terms – "pain", "C-fibers stimulation", "heat" and "molecular motion" – seem to be rigid designators, i.e. they refer to the same phenomenon in every possible world, so, if true, the identity of pain with C-fibers stimulation and heat with molecular motion must be necessary[37]; however, the two pairs of elements behave differently.

The example of the "heat-molecular motion" dyad is clear. Heat cannot exist without necessarily being molecular motion; nevertheless, the molecular motion can exist without producing the accidental property of "sensation of heat", which is a mediating element between heat and molecular motion, i.e. our consciousness of the heat[38]. In other words, the term "heat" directly refers to the molecular motion, nevertheless, the latter could exist without producing the sensation of heat, for example, when part of my skin is anesthetised; on the other side, there could be the sensation of heat without presence of heat/molecular motion, e.g. when we get freezer burns. Therefore, "heat" is theoretically (necessarily) identical with "molecular motion", while it is only

[34] A clear example of contingent identity is given by A. GIBBARD, *Contingent identity*, «Journal of Philosophical Logic», 4, 2/1975, pp. 191. «There is a statue, Goliath, and the clay, Lumpl, from which it is composed. If Lumpl and Goliath coincide spatiotemporally, it is reasonable to say that they are identical. But they might not have been. «For suppose I had brought Lumpl into existence as Goliath, just as I actually did, but before the clay had a chance to dry, I squeezed it into a ball. At that point, according to the persistence criteria I have given, the statue Goliath would have ceased to exist, but the piece of clay Lumpl would still exist in a new shape. Hence Lumpl would not be Goliath, even though both existed». Thus, the identity of Lumpl and Goliath seems merely contingent.

[35] S. KRIPKE, *Naming and Necessity*, cit., p. 4, pp. 97-105.

[36] S. KRIPKE, Identity and Necessity, in S. KRIPKE, Philosophical Troubles: Collected Papers, Volume 1, Oxford University Press, New York 2011, pp. 1-26.

[37] Ivi, pp. 148-149.

[38] Similarly, "water" can also be picked out either by its internal structure H_2O or by the combined sensations of wetness, transparency, tastelessness, etc.; however, only "water = H_2O" is necessarily true, while the combination of sensations is an accidental property.

contingently identical with the sensation of heat. This is not the case for "pain" and "C-fibers stimulation". Indeed, there is no accidental intermediate between the two, i.e. there is no difference between "pain" and "sensation of pain"; to be in the situation of having pain means only that I have a pain, i.e. the sensation of pain is a pain as such[39]. Moreover, the relationship between "pain" and "C-fibers stimulation" (or, more generally, "brain activity *x*") is not the same as the "heat-molecular motion" dyad; indeed, we can imagine a situation in which a particular pain sensation exists without neural activity *x* and vice versa. Note that Kripke is not arguing for some ontological mind-body dualism between pain sensations and physical-biological activities of the body; he is just showing how "C-fibers stimulation" and "pain" cannot be a theoretical identity, i.e. a necessary identity, but rather a correlation, correspondence or co-occurrence. Thus, following Kripke, the "C-fibers stimulation-pain" relationship is not an identity at all, while for others like Gibbard or Schwarz is a case of contingent identity[40].

3.2 The intrinsic relationship between consciousness and subjectivity

Now, we need to focus again on the relationship between "consciousness" and "brain activity *x*" to see whether it represents a theoretical identity – as "heat = molecular motion" – an empirical covariation/correlation or, at best, a contingent identity – as "pain" and "C-fibers stimulation". First, I introduce the terms "consciousness" and "brain activity x" respectively for "heat" and "molecular motion", for "pain" and "C-fibers stimulation", and also for "water" and "H_2O". Then, I would introduce the subjective character as the intermediate element for the "consciousness-brain activity

[39] «In the case of the identity of heat with molecular motion the important consideration was that although "heat" is a rigid designator, the reference of that designator was determined by an accidental property of the referent, namely the property of producing in us the sensation S. It is thus possible that a phenomenon should have been rigidly designated in the same way as a phenomenon of heat, with its reference also picked out by means of the sensation S, without that phenomenon being heat and therefore without its being molecular motion. Pain, on the other hand, is not picked out by one of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality. Thus pain, unlike heat, is not only rigidly designated by "pain" but the reference of the designator is determined by an essential property of the referent» S. KRIPKE, *Naming and Necessity*, cit., p. 4, 152-153. Interestingly, this is quite similar to what Sartre means with "consciousness (of) pleasure", which also works with the notion of pain: «Pleasure can not be distinguished – even logically – from consciousness of pleasure. Consciousness (of) pleasure is constitutive of the pleasure as the very mode of its own existence, as the material of which it is made, and not as a form which is imposed by a blow upon a hedonistic material» J.-P. SARTRE, *Being and Nothingness*, Philosophical Library, New York 1956, p. liv.

[40] A. GIBBARD, *Contingent identity*, cit. W. SCHWARZ, *Contingent Identity*, «Philosophy Compass», 8, 5/2013, pp. 486-495.

x" dyad. "Subjectivity" means that when I am conscious of something, there is something it is like for me to be conscious; in other words, when I am conscious of some particular content, I am also conscious of myself being in that particular condition. This subjective character works here as the analogue term for "sensation of heat", "sensation of pain" and "colourless liquid". Now, we have four triads of terms: "heat-sensation of heat-molecular motion", "pain-sensation of pain-C-fibers stimulation", "water-colourless liquid-H_2O" and "consciousness-subjectivity-brain activity *x*" (see figure below).

"Heat"	"Water"	"Pain"	"Consciousness"
Sensation of heat	Colourless liquid	Sensation of pain	Subjectivity
Molecular motion	H_2O	C-fibers stimulation	Brain activity *x*

According to Kripke, the intermediate element named "sensation of heat" is an accidental property of the "heat-molecular motion" identity and this is even more evident with the intermediate element "colourless liquid" concerning the "water-H_2O" necessary a posteriori identity. On the other side, pain can only be picked out by the sensation of pain, because being in "that" pain is not an accidental property of pain, rather an essential, intrinsic one. Regarding consciousness and the brain, given a particular brain activation – just for the sake of argument – we need to understand whether the "consciousness-subjectivity-brain activation" triad is similar to "pain-sensation of pain-C-fiber stimulation" one or to "heat-sensation of heat-molecular motion" and "water-colourless liquid-H_2O" ones. One might state that since the feeling of pain is an instance of a conscious state, the relation between consciousness and that brain activity is contingent as well. This is true, however, I will test it also in another way, first through the Cartesian and Zombie intuitions[41], then by analysing the nature of subjectivity.

The so-called "Cartesian intuition" refers to Part IV of the *Discourse of the Method*, where Descartes addresses the possibility that our experience might not cohere with the physical world as it actually is, as in the movie *The Matrix*. The "Zombie intuition", instead, derives from Robert Kirk's argument in 1974, later developed by Chalmers' thought-experiment about the so-called "philosophical zombie"[42]; the intuition suggests that, in some possible

[41] E. DIETRICH, V. G. HARDCASTLE, *Sisyphus's Boulder*, John Benjamins Publishing Company, Amsterdam 2005.

[42] R. KIRK, *Zombies*, in E. N. ZALTA (ed.), *Stanford Encyclopedia of Philosophy*, Metaphysics

world, there might be creatures perfectly identical to us who behave in a perfectly normal way, but who completely lack subjective consciousness, i.e. there is "nothing it is like" to be a zombie, they do not have any phenomenal sensation. To sum up, the Zombie intuition claims that consciousness may vary while the world remains the same, while the Cartesian intuition claims that the world may vary while consciousness remains the same[43].

Transposing this mind-world relationship into our "consciousness-brain activity x" relation, we can conceive some possible world in which consciousness exists without the particular brain activation (Cartesian intuition) and, at the same time, we can conceive the activation of that neural network without the necessary presence of consciousness (Zombie intuition). Therefore, the relation between "consciousness" and "brain activity *x*" is not necessary. This is far from saying that consciousness and the brain are not related, but that we need to be careful about the distinction between necessary and contingent identities. For this reason, the intrinsic structure of consciousness is not identical to some part of the brain, and consequently "consciousness" does not rigidly refer to some brain activity, although it depends in part on the brain.

We need to discuss also the nature of the intermediate element named the "subjective character" of consciousness. The intentional content of consciousness ("I am conscious of something") also has a subjective character that implicitly refers to the intentional act itself ("I am conscious that I am conscious of something"). Is the subjective character an accidental property like "sensation of heat" for "heat" or "colourless liquid" for "water" or is it an intrinsic property like "feeling of pain", whereby "pain" is rigidly designated by the feeling of "that" pain? I would argue that consciousness is not picked out by some accidental property, but rather that subjectivity is an intrinsic property of consciousness. Indeed, we cannot conceive the separability of the act of "being conscious of something" and the act of "being conscious of

Research Lab, Stanford University 2019. Actually, this intuition also derives from a Cartesian argument we can find in the VI *Metaphysical Meditation* (it is not the same, though): «if I consider the body of a man as being a sort of machine so built up and composed of nerves, muscles, veins, blood and skin, that though there were no mind in it at all, it would not cease to have the same motions as at present, exception being made of those movements which are due to the direction of the will, and in consequence depend upon the mind (as opposed to those which operate by the disposition of its organs)» R. DESCARTES, *Discourse on Method and Meditations*, Dover Publications, Mineola, New York 2003, p. 117.

[43] E. DIETRICH, V. G. HARDCASTLE, *Sisyphus's Boulder*, cit., p. 28. Note that it is not my intention to use these as strong ontological arguments against any physicalist theory of consciousness; indeed there are some concerns about the validity of these intuitions, e.g. regarding the logico-ontological passage from "conceiving" zombie to the "possibility" of them. I am using the two intuitions just to put into question the alleged necessary identity between consciousness and some particular brain activity.

being conscious of something", i.e. the awareness of being in that intentional state.

In the same way, we cannot conceive the separability of pain and the feeling of pain. We cannot imagine the possibility of being intentionally directed towards something (being conscious of an object) without necessarily implying being conscious of this act. So, while the relationship between the sensation of heat and the "heat-molecular motion" dyad is contingent, i.e. "heat-molecular motion" can exist without anyone experiencing the sensation of heat, the relationship between "consciousness" and "subjective character" is necessary. Once we remove the latter, there is no consciousness at all; still, this intrinsic relationship between "consciousness" and subjectivity is necessary a priori, while the alleged necessary a posteriori relationship with some brain activity has proven to be false.

Conclusions

In this paper, I tried to identify the peculiar features of consciousness, starting from the very use of the notion of "consciousness". I based the analysis on a Kripkean-like framework, treating the term "consciousness" as a general term. I have argued that this term rigidly refers to the act of being directed towards something (intentionality) and, following Kripke's argument concerning "pain" and "C-fibers stimulation", I discussed and criticized the supposed identity between consciousness and brain activity. It does not seem possible to develop a necessary a posteriori statement that relates consciousness and brain activity, as happens in the proposition "heat is molecular motion"; nevertheless, neuroscientific findings show that the brain remains a fundamental element for consciousness, and, perhaps, it will be possible to define a necessary a posteriori statement by shifting the perspective from a brain-centred to a neuro-ecological approach, so that the brain will be conceived as a part of a broader and holistic body-environment relational system, which is related to the presence of consciousness[44]. For the moment, through this Kripkean approach, it seems possible to define subjectivity as the intrinsic property of the notion of consciousness; when we directly refer to the intentional act towards something, we cannot separate it from the subjective character of the act itself, that is, the fact that we are conscious of being conscious. Against this idea, one may argue that in our daily experience we are always aware of every detail and, in some case, we

[44] G. NORTHOFF, *The Spontaneous Brain: From the Mind–Body to the World–Brain Problem*, The MIT Press, Cambridge 2018.

could perceive something unconsciously that still has a significant impact on later behaviour and thought, such as during the so-called "unconscious perception"[45]. For example, one might say that people with blindsight – but also healthy subjects during visual masking – can perceive something without being conscious of it. This is an interesting issue that, I believe, rests on misconceptions of consciousness and awareness that are often interpreted as higher-order cognitive acts, instead of pre-reflective, non-cognitive acts[46]. Thus, to give a provisional answer, we may consider blindsight not as a genuine case of perception without subjectivity, but rather as an abnormal or degraded – nonetheless conscious – case of perception[47]. In order to avoid these misinterpretations, a further phenomenological investigation is necessary, which may start from the provisional concept of consciousness presented here[48].

References

S. AARONSON, *Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)*, Retrieved November 20, 2020, from: https://www.scottaaronson.com/blog/?p=1799

AUGUSTINE, *The confessions*, in *Masterpieces of Philosophical Literature*, edited by T. L. Cooksey, Greenwood Press, Westport 2006

N. BLOCK, *On a confusion about a function of consciousness*, «Behavioral and Brain Sciences», 28, 2/1995, pp. 227-247

D. J. CHALMERS, *Facing up to the problem of consciousness*, «Journal of Consciousness Studies», 2, 3/1995, pp. 200-219

D. J. CHALMERS, *What is a neural correlate of consciousness*? in T. METZINGER (ed.), *Neural Correlates of Consciousness*, MIT Press, Cambridge 2000, pp. 17-39

D. J. CHALMERS, *The virtual and the real*, «Disputatio», 9, 46/2017, pp. 309-352

[45] J. PRINZ, *Unconscious Perception*, in M. MATTHEN (ed.), *The Oxford Handbook of Philosophy of Perception*, Oxford University Press, New York 2015.

[46] F. ZILIO, *Consciousness and World. A Neurophilosophical and Neuroethical Account*, Edizioni ETS, Pisa 2020.

[47] M. PERSUH, *The Fata Morgana of Unconscious Perception*, «Frontiers in Human Neuroscience», 12, 120/2018. I. PHILLIPS, *Consciousness and Criterion: On Block's Case for Unconscious Seeing*, in «Philosophy and Phenomenological Research», 93, 2/2016, pp. 419-451.

[48] I thank the anonymous reviewer for useful suggestions that improved this manuscript.

T. CRANE, *Intentionalism*, in A. BECKERMANN, B. P. MCLAUGHLIN, S. WALTER (eds.), *The Oxford Handbook of Philosophy of Mind*, Oxford University Press, New York 2009, pp. 474-493

R. DESCARTES, *Discourse on Method and Meditations*, Dover Publications, Mineola, New York 2003

E. DIETRICH, V. G. HARDCASTLE, *Sisyphus's Boulder*, John Benjamins Publishing Company, Amsterdam 2005

B. FAW, *Cutting «Consciousness» at its Joints*, «Journal of Consciousness Studies», 26, 5/2009, pp. 54-67

K. FRANKISH, *Illusionism as a Theory of Consciousness*, «Journal of Consciousness Studies», 23, 11-12/2016, pp. 11-39

A. GIBBARD, *Contingent identity*, «Journal of Philosophical Logic», 4, 2/1975, pp. 187-222

M. GUILLOT, *I Me Mine: on a Confusion Concerning the Subjective Character of Experience*, «Review of Philosophy and Psychology», 8, 2017, pp. 23-53

P. JACOB, *Intentionality*, in E. N. ZALTA (ed.), *Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University 2019

R. KIRK, *Zombies*, in E. N. ZALTA (ed.), *Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University 2019

S. KRIPKE, *Naming and Necessity*, Harvard University Press, Cambridge 1980

S. KRIPKE, *Identity and Necessity*, in S. KRIPKE, *Philosophical Troubles: Collected Papers, Volume 1*, Oxford University Press, New York 2011, pp. 1-26

S. LAUREYS & G. TONONI, *The Neurology of Consciousness*, Academic Press, San Diego 2009

T. NAGEL, *What Is It Like to Be a Bat?*, «The Philosophical Review», 83, 4, 1974, pp. 435-450

G. NORTHOFF, *The Spontaneous Brain: From the Mind-Body to the World-Brain Problem*, The MIT Press, Cambridge, 2018

G. NORTHOFF, V. LAMME, *Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight?*, «Neuroscience & Biobehavioral Reviews», 118, 2020, pp. 568-587

H. ROBINSON, *Why phenomenal content is not intentional*, «European Journal of Analytic Philosophy», V, 2/2009, pp. 79-93

M. PERSUH, *The Fata Morgana of Unconscious Perception*, «Frontiers in Human Neuroscience», 12, 120/2018

I. PHILLIPS, *Consciousness and Criterion: On Block's Case for Unconscious Seeing*, in «Philosophy and Phenomenological Research», 93, 2/2016, pp. 419-451

J. PRINZ, *Unconscious Perception*, in M. MATTHEN (ed.), *The Oxford Handbook of Philosophy of Perception*, Oxford University Press, New York 2015

J.-P. SARTRE, *Une idée fondamentale de la phénoménologie de Husserl: l'intentionalité*, «La Nouvelle Revue française», 304/1939, pp. 129-131

J.-P. SARTRE, *Being and Nothingness*, Philosophical Library, New York 1956

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

J. R. SEARLE, *The Rediscovery of the Mind*, MIT Press, Cambridge 1992

J. R. SEARLE, *Mind, language and society: philosophy in the real world*, Weidenfeld & Nicolson, London 1999, pp. 40-41

J. R. SEARLE, *Il mistero della realtà*, Raffaello Cortina, Milano 2019

R. VAN GULICK, *Consciousness*, in E. N. ZALTA (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University 2018

R. VIMAL, *Meanings Attributed to the Term «Consciousness»: An Overview*, «Journal of Consciousness Studies», 26, 5/2009, pp. 9-27

A. VOLTOLINI, *The Mark of the Mental*, «Phenomenology and Mind», IV, 2013, pp. 124-136

N. SALMON, *Are General Terms Rigid?*, «Linguistics and Philosophy», 28, 1/2005, pp. 117-134

W. SCHWARZ, *Contingent Identity*, «Philosophy Compass», 8, 5/2013, pp. 486-495

E. SCHWITZGEBEL, *Do you have constant tactile experience of your feet in your shoes? Or is experience limited to what's in attention?*, «Journal of Consciousness Studies», 24, 3/2007, pp. 5-35

F. ZILIO, *Consciousness and World. A Neurophilosophical and Neuroethical Account*, Edizioni ETS, Pisa 2020