Integrated information theory and the metaphysics of consciousness

Abstract: In recent years, many philosophical and neuroscientific theories of consciousness were proposed with the aim of solving the hard problem of consciousness. Giulio Tononi’s Integrated Information Theory (IIT) defines consciousness as integrated information and provides an account of its quantitative and qualitative aspects based on cognitive neuroscience research (Tononi, 2004, 2008). Even if IIT covers the phenomenal aspects of consciousness, until now this theory has not been object of a philosophical analysis. In this paper I propose a categorization of IIT within the main metaphysical positions about consciousness. To do that I model a taxonomy of positions based on Chalmers (2002) descriptions of three possible strategies to face up to the hard problem: materialism (type-A and B), dualism (type-D and E) and non-reductive monism (type-F). I show that these positions and their subtypes are distinguished by respect of four basic metaphysical assumptions, that are: naturalism, the causal closure of the physical world, the existence of an explanatory gap and the existence of an ontological gap. I analyze which of these four assumptions IIT accepts and which denies and finally I conclude that IIT is compatible with type-B materialism and type-F monism. Since this positions are different under many aspects, I argue that IIT's metaphysical assumptions have to be discussed further for IIT to represent an exhaustive theory of consciousness.

Introduction

In the current debate in philosophy of mind, philosophers and cognitive neuroscientists proposed several theories as a solution to the so-called hard problem of consciousness. Among them, the Integrated Information Theory (IIT) proposed by Giulio Tononi (2004, 2008) is gaining experimental results on the study of altered states of consciousness (such as sleep) (Massimini et al., 2009) and disorders of consciousness (such as coma and vegetative state) (Rosanova et al., 2012).

IIT constitutes a model of consciousness, because it pinpoints explanatory links between phenomenal and qualitative aspects of consciousness and its neural substrate, besides describing only the neural correlates of consciousness itself (Seth, 2009). Nevertheless, IIT has never been subject to philosophical analysis until now. In this paper I will conduct a philosophical analysis of IIT's metaphysical assumptions, in order to show that IIT is consistent with two philosophical positions about the metaphysics of consciousness. For this purpose my argumentation will take the following steps. Firstly, I will build a taxonomy of the main metaphysical positions about
consciousness. This taxonomy will be based on Chalmers' (2002) distinction among materialism (type-A, B, and C), dualism (type-D and E), and non-reductive monism (type-F). I will show that these positions can be classified with respect to four base assumptions: naturalism, the causal closure of the physical world, the existence of an explanatory and of an ontological gap which distinguish phenomenal and physical properties. Then, I will show that IIT accepts for sure three out of four of these assumptions, with a doubt about the last one. For this reason, with respect to the taxonomy of positions finally I will conclude that IIT is not a form of substance dualism or type-A materialism, instead it is compatible both with type-B materialism and type-F monism.

Consciousness as Integrated Information

IIT identifies consciousness with the capacity of a system to integrate information. In order to do that, a system has to be characterized by (1) a large and differentiated repertoire of possible internal states and (2) an integrated internal structure which determines the causal dependence of each different element's state. For IIT, consciousness is integrated information and each conscious state is highly informative because the occurrence of each different state rules out the occurrence of all the alternative states of the system's repertoire (information), and every conscious state is a unique activation configuration of the whole system (integration) (Tononi, 2004, 2008).

IIT proposes a measure of consciousness generated by a system. Consciousness in a system is defined as the minimum quantity of causally effective information integrated across the weakest connection between two mid-partitions of the system itself, and it is identified by the letter Φ (phi). For IIT, the quantity of Φ is a continuum, and every system able to generate sufficient amount of Φ is conscious, being it a biological brain or an artificial system (Tononi, 2004).

Besides a quantitative theory of consciousness, IIT faces up to the problem of qualitative experience, therefore replying to the hard problem of consciousness. Tononi proposes an explanation of qualia, based on the concept of activation states' repertoire of each system. IIT defines the quality of experience as the set of informational relationships generated within integrated systems. The states' repertoire of a system
can be represented as a qualia space (Q), of which points represent probability distribution of each state, and arrows between points represent informational relationships among connected elements. Qualia are defined as shapes in Q that univocally specify each and every experience because of the set of informational relationships which characterizes the specific Q subspace (Tononi, 2008; Balduzzi & Tononi, 2009).

1. A taxonomy of positions in the metaphysics of consciousness

In his *Consciousness and its place in nature* from 2002, Chalmers claims that functional and reductive explanations work for the easy problems (such as explaining attention, behavior control, mental states report), but not for the hard problem of explaining why conscious experience originates in the first place. On this basis, Chalmers states that there is an explanatory gap between phenomenal and physical properties, that is to say: there is not an epistemic entailment between physical truths (P) and phenomenal truths (Q). But, he concludes, if there is no epistemic entailment, then there is no ontological entailment between P and Q. Moreover, if there is no ontological entailment, then there is an ontological gap that distinguish consciousness and the physical world, therefore materialism is false (Chalmers, 2002).

On this basis, Chalmers suggests three main positions that one can adopt to reply to these arguments: materialism, dualism, and non-reductive monism. I will classify the different positions and their respective subtypes with respect to four basic assumptions.

1.1 The four basic assumptions

The first assumption is naturalism (N), and states the following:

(N) Whatever exists is susceptible to explanation through methods which are continuous from domain to domain and which are paradigmatically exemplified in the natural sciences (Danto, 1967)

This perspective is not necessarily related to a form of reductionism of physicalism, because it doesn't state that every object is physical.

The second assumption is the causal closure of the physical world (CC), that is:
(CC) Every physical event has physical causes and there aren't non-physical causes of physical events (Kim, 1998)

This assumption posits that there is no interaction between physical and non-physical, but it doesn't exclude the existence of non-physical objects.

The third assumption concerns the subsistence of an explanatory gap (EG):

(EG) No physical account of structure and function can explain conscious experience (Levine 1983; Chalmers, 2002)

This assumption is the conclusion of many philosophical arguments and thought experiments (notably the knowledge argument presented by Jackson (1986)).

The fourth and last assumption accepts the subsistence of an ontological gap (OG):

(OG) Phenomenal properties are not reducible to physical properties (Chalmers, 2002)

This assumption is the conclusion of many philosophical arguments such as Chalmers’ thought experiment about philosophical zombies. Following this argument, if a system physically and functionally identical to a human being but devoid of conscious experience (namely a zombie) is conceivable, then it is logically and metaphysically possible. If zombies are metaphysically possible, therefore phenomenal properties are not physical, as conscious experience itself (Chalmers, 1995, 2002).

Resting upon these four assumptions, it is possible to shape a taxonomy of the different metaphysical positions about consciousness. Following Chalmers’ (2002) categorization, three main ways of replying to the hard problem of consciousness are materialism, dualism and non-reductive monism. Now, let's move to categorize each position, as well as their subtypes, with respect to the four basic assumptions.

1.2 Materialism and its subtypes

Firstly, materialist strategies accept both the assumption of naturalism (N) and of causal closure (CC) denying that there are non-physical causes of physical events. In addition, materialists claim that every object and property is physical or is reducible to physical objects and properties, therefore they deny the subsistence of an ontological gap and
reject OG. Nevertheless, there are two subtypes of materialism that take a different position with respect to EG.

Type-A materialists claim that there is no explanatory gap because a functional explanation of conscious experience is possible (if not now at least in the future).

On the contrary, type-B materialists buy EG even rejecting OG. Type-B materialists like Kripke (1980), for instance, claim that the ontological entailment between physical properties and phenomenal properties (P necessitates Q) is necessarily true, even if it is not knowable as true a priori. In fact, for type-B materialism the entailment between P and Q is a necessary a posteriori truth like “water is H₂O”, that is necessarily true (in every possible world) without being knowable a priori (Chalmers, 2002)¹.

With respect to the four basic assumptions, materialist strategies can be categorized in the following table (Tab. 1):

<table>
<thead>
<tr>
<th>Materialism</th>
<th>N</th>
<th>CC</th>
<th>EG</th>
<th>OG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type-A</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Type-B</td>
<td>√</td>
<td>√</td>
<td>‐</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 1: Materialism subtypes with respect to the four basic assumptions: naturalism (N), causal closure of the physical world (CC), explanatory gap (EG), and ontological gap (OG).

1.3 Substance dualism and its subtypes

Besides materialism, Chalmers describes the dualistic strategy, with its two subtypes. In general, dualism refuses a naturalistic view of the world (N). The substance dualism claims the existence of two different substances that distinguish mental and physical objects and therefore accepts both EG and OG. Nevertheless, Chalmers identify two subtypes of dualism: type-D and type-E.

Type-D dualism is the substance dualism proposed by Descartes himself, which distinguishes between res cogitans (characterized by inextension, freedom, and consciousness) and res extensa (characterized by extension, constraint, and

¹ Chalmers includes also type-C materialism, which claims that there is an explanatory gap but that it will be filled up in the future (Churchland, 1997), or, even if fillable in principle, it will remain due to the scientific or cognitive limitations of human beings (Nagel, 1974; Mc Ginn, 1989). Since this two positions are reducible respectively to type-A and type-B materialism, I will not include them in my taxonomy.
unconsciousness) (Descartes 1641/1996). This position denies CC because it claims that physical events have both physical and non-physical causes (e.g. mental states can interact with physical states), in fact it is also called interactionism\(^2\).

Type-E dualism on the contrary accepts CC because it claims that there is no interaction between mind and matter, even if they are different substances. This position is also called epiphenomenalism because it sees consciousness as an epiphenomenon, a by-product of cerebral activity without any causal power on the physical world, similar to the steam-whistle which accompanies the work of a locomotive engine (Huxley, in James, 1890).

With respect to the four basic assumptions, dualist theory and its subtypes can be categorized in the following table (Tab. 2):

<table>
<thead>
<tr>
<th>Dualism</th>
<th>N</th>
<th>CC</th>
<th>EG</th>
<th>OG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type-D</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Type-E</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

Table 2: Dualism subtypes with respect to the four basic assumptions: naturalism (N), causal closure of the physical world (CC), explanatory gap (EG), and ontological gap (OG).

1.4 Non-reductive monism

Despite the usual opposition between materialism and dualism, substance dualism is not the only alternative to a materialistic perspective. Non-reductive monism constitutes an alternative too, and it is the strategy of rejecting materialism without necessarily buying a substance dualism perspective. As described by Chalmers (2002), one more position can be identified: type-F monism, also called neutral monism or property dualism.

Type-F monism claims that phenomenal properties (i.e. conscious experience) are fundamental in the same way in which physical properties are; nevertheless, it doesn't state that phenomenal and physical properties pertain to different substances. On the contrary, for type-F monism there is only one substance, one metaphysical principle that constitutes reality, and both phenomenal and physical properties are fundamental.

\(^2\) An example of interaction dualism is the position of Eccles and Popper (1984).
aspects mutually irreducible of this one and only substance. This position, in fact, accepts a naturalistic view (N) and states that only physical objects have causal power (CC). In addition, since phenomenal and physical properties are fundamental and mutually irreducible, type-F monism states that there are both an ontological and an explanatory gap between the phenomenal and the physical realms. In this way, type-F monism constitutes a naturalistic and monistic position avoiding materialism, and at the same time it constitutes a non-reductive and non-materialistic position avoiding substance dualism.

With respect to the four basic assumptions, non-reductive monism can be categorized in the following table (Tab. 3):

<table>
<thead>
<tr>
<th>Non-reductive monism</th>
<th>N</th>
<th>CC</th>
<th>EG</th>
<th>OG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type-F</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

*Table 3: Non-reductive monism with respect to the four basic assumptions: naturalism (N), causal closure of the physical world (CC), explanatory gap (EG), and ontological gap (OG).*

1.5 A taxonomy of positions in the metaphysics of consciousness

As we can see, the three main metaphysical positions and their subtypes can be categorized with respect to the four basic assumptions. Now that our taxonomy is complete, we can synthesize it in the following table (Tab. 4) and move to the next step: the analysis of IIT’s basic assumptions and its categorization.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>CC</th>
<th>EG</th>
<th>OG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type-A</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Type-B</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>X</td>
</tr>
<tr>
<td>Type-D</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Type-E</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Type-F</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

*Table 4: Taxonomy of metaphysical positions about consciousness with respect to the four basic assumptions: naturalism (N), causal closure of the physical world (CC), explanatory gap (EG), and ontological gap (OG).*
2. IIT’s metaphysical assumptions on consciousness

At this stage of the analysis, it is sufficient to evaluate if IIT accepts each one of the four basic assumptions in order to classify IIT with respect to the built taxonomy. Despite its proposer Giulio Tononi claims IIT is a theory of consciousness as a fundamental entity (Tononi, 2004, 2008), type-F monism is not the only metaphysical perspective compatible with it. As we will see, IIT is compatible also with type-B materialism.

2.1 IIT and naturalism

IIT is by definition a neurobiological as well as a quantitative and mathematical theory of consciousness. It identifies consciousness as a measurable property (i.e. integrated information) analyzable through neuroscientific experiments and techniques (Tononi, 2003, 2004). For IIT, both quantitative and qualitative aspects of consciousness arise from a system’s capacity of generating high-levels of integrated information. Consciousness is measurable, in fact Tononi and colleagues are developing different methods to assess and measure its level in different altered states of consciousness such as deep and REM sleep (Massimini et al., 2009, 2010), anesthesia (Ferrarelli et al., 2010), vegetative and minimally conscious state (Rosanova et al., 2012). Given that, for IIT consciousness is analyzable by means of natural sciences, therefore IIT definitely embraces the assumption N.

2.2 IIT and the causal closure of the physical world

Denying the CC assumption means to accept that non-physical objects can cause (or interact with) physical events. For Tononi (2004, 2008), consciousness as integrated information is the result of brain activity. It is a property that emerges from the causal interactions of physical elements, and it is produced according to the laws of physics as it is the information generated by physical systems. As information, mind has no causal power on its physical substrate. As we will see, this is true either IIT endorses a materialist perspective (for which consciousness is a functional property generated by physical systems), or a non-reductive monistic perspective (for which consciousness is a fundamental property of the world).

Since IIT accepts both N and CC then it is not compatible with a form of substance
dualism, because accepting N is not compatible with type-E dualism and accepting N and CC is not compatible with type-D dualism (see Tab. 4). Therefore, IIT is not a form of substance dualism.

2.3 IIT and the explanatory gap

The three arguments presented by Chalmers (2002) suggest the impossibility to deduce truths about phenomenal properties (Q) from truths about physical properties (P). As claimed by Chalmers, all the truths about P do not exhaust all the truths about Q (from the explanatory argument), P is conceivable also in absence of Q (from the conceivability argument), and even knowing all truths about P one cannot deduce all truths about Q (from the knowledge argument); therefore, there is an explanatory gap that limits the knowledge about Q that we can deduce from our knowledge about P (Chalmers, 2002).

Tononi states clearly his position about the explanatory gap. In fact, he claims that a description in terms of informational relationships among different internal states of a system constitutes a full description of the content of conscious experience and of how it is generated. Nevertheless, this description cannot substitute for first-person experience, for being is different from describing (Tononi, 2008). Referring to the third argument presented by Chalmers, the knowledge argument about Mary the neuroscientist (Jackson, 1986), Tononi states his position. In the thought experiment, Mary knows all the truths pertaining to the physical aspects of seeing colors, but ignores every truth about phenomenal aspects, because she never saw colors first-hand. For Tononi, Mary actually learns something new when she sees colors, because describing only implies knowing from the outside, while being implies knowing from the inside, and consciousness is a way of being and not a way of knowing (Tononi, 2008). Therefore, IIT accepts the assumption of the explanatory gap (EG) between phenomenal and physical knowledge.

Given that IIT accepts EG, besides N and CC, we can conclude that this theory is not a form of type-A materialism, which instead denies the existence of an explanatory gap and refuses the corresponding assumption EG.
2.4 IIT and the ontological gap

As Chalmers claims, the three arguments against materialism bring to the conclusion that there is an explanatory gap. Moreover, from this conclusion Chalmers derives that there is not only an epistemic gap, but also an ontological one (Chalmers, 2002). In fact, referring to Kripke's (1980) argumentation, Chalmers claims that if P necessitates Q, then the entailment between P and Q is ontological and it is true in every possible world. If this is the case, then the entailment between P and Q should also be epistemic, that is to say that one can't know a truth about P without knowing *a priori* the corresponding truth about Q. But, given that the explanatory gap states the contrary, Chalmers concludes that there is a more fundamental gap, an ontological one.

The last two positions that remain both accept N, CC, and EG; nonetheless, they are distinguished with respect to the last assumption OG. Type-B materialism accepts a physicalist perspective and therefore denies the existence of an ontological gap. On the contrary, type-F monism claims that there is an ontological gap in the sense that it rejects physicalism, claiming that phenomenal and physical properties are not mutually reducible. Even if Tononi states that his theory describes consciousness as a fundamental entity, in the next paragraphs I will show that IIT results compatible with both type-F monism and type-B materialism.

2.4.1 IIT and type-F monism

Since IIT describes consciousness as a fundamental entity, it is useful to analyze deeper which sense of “fundamental” Tononi means. IIT in fact is not a form of substance dualism, therefore consciousness it is not fundamental in the sense of a metaphysical principle other from matter. IIT claims consciousness is fundamental in the sense that it characterizes reality in the same way of mass, charge and energy: if there is a functional mechanism with integrated structure that is in a certain state, then consciousness is generated and it exists *ipso facto* as integrated information (Tononi, 2008). In this sense, consciousness can be considered a widespread property, extended to every object whose structure generates integrated information.

Given that for IIT consciousness is not a separate substance, but it is a fundamental entity not reducible to the physical ones, IIT is compatible with a form of non-reductive
monism as type-F monism. In this perspective, phenomenal and physical properties are considered fundamental but mutually irreducible properties of the one and only substance that constitutes the monistic principle of all reality. Therefore, as a form of type-F monism, IIT denies physicalism and accepts the existence of an ontological gap between the phenomenal and the physical realms. As Chalmers claims, in fact, type-F monism accepts that the explanatory gap derives from a more profound ontological gap, but remains a monistic theory that denies the existence of two separate substances (Chalmers, 2002).

2.4.2 IIT and type-B materialism
Even if IIT is compatible with type-F monism, its position about OG is not clear. Moreover, since non-reductive monism is a position that refuses materialism, it is useful to underline that IIT is also compatible with a form of type-B materialism, if it refuses the existence of an ontological gap. In a certain sense, in fact, consciousness can be described as a “fundamental” property also from a purely materialistic perspective, even accepting the existence of an explanatory gap.

The first strategy that type-B materialists can adopt consists in claiming that the relationship between phenomenal and physical properties is identity. I will call this strategy type-BI materialism, where the “I” stands for identitarianism. The second strategy is to claim that physical and phenomenal properties are not identical, but that phenomenal properties supervene on physical properties (in the sense of Kim, 1990). I will call this strategy type-BS materialism, where the “S” stands for supervenience.

Type-BI materialists like Papineau (1998) and Levine (1998) claim that identity is a primitive relationship, a “brute fact” that needs no explaining. For instance, when we say: “water = H₂O”, it makes no sense to ask further why water is identical to H₂O, because if they are the same they are the same, and that's it. In the same way, if conscious and material properties are identical, then it makes no sense to ask, for instance, why pain is identical to the firing of C-fibers, because conscious and mental states don't arise from physical states, they are physical states, and that's it (Papineau, 1998). The only difference is that the identity between water and H₂O is more
comprehensible than the identity between pain and C-fibers firing, or between the experience of seeing red and the corresponding cerebral activations. For this reason, asking why cerebral and mental states are identical only seems to make sense, but our inability to understand it easily doesn't show that they aren't identical, because a gap in our understanding of nature doesn't demonstrate the existence of a gap in nature (Levine, 1998).

While type-BI materialism refuses OG and accepts EG, claiming that mental and physical properties are identical, type-BS materialism adopts a different strategy. Type-BS materialists claim that mental states (M) are not identical, but supervene on physical states (P) in the sense that there cannot be M-differences without P-differences (Kim, 1990). Therefore, in this perspective, there is no ontological gap between phenomenal and physical properties. Nonetheless, even if type-BS materialists refuse OG, they can still assume EG, claiming that supervenience is a necessary relationship, even if not knowable a priori. The relationship between phenomenal and physical properties can be considered as a necessary a posteriori truth in the kripkean sense (Kripke, 1980), such as the relationship between water and H₂O or between genes and DNA. In fact, as one can know truths about water without knowing a priori that water is necessarily identical to H₂O, one can know truths about phenomenal properties without knowing a priori that they necessarily supervene on physical properties.

**Conclusions**

IIT is a promising approach to the study of consciousness for it is based on rigorous neuroscientific data and faces up many philosophical issues about phenomenal consciousness. To the present day, this theory has not been object of philosophical analysis. In this paper I moved the first steps, answering to the basic and fundamental question about which philosophical position IIT is compatible with.

By modeling the taxonomy of philosophical positions upon four basic assumptions (naturalism, the causal closure of the physical world, the existence of an explanatory and an ontological gap) I showed that materialism, substance dualism, non-reductive monism, and their subtypes are clearly distinguishable. This method has revealed to be
fruitful because it was sufficient to evaluate if IIT accepts or refuses each one of the four basic assumptions to deduce which metaphysical position IIT is compatible with. Since IIT is compatible both with type-B materialism and type-F monism, it can follow two strategies to reply to the hard problem of consciousness. Each strategy has strengths and weaknesses, and a deeper analysis should be conducted in order to reveal the respective philosophical implications of the two different positions. Moreover, if IIT aspires to be considered as an exhaustive theory of consciousness, its metaphysical assumptions and its philosophical perspective have to be discussed and clarified further.

To the present day, many neurobiological and philosophical theories are competing to finally provide a solution to the hard problem of consciousness. If the quest for consciousness will see an end in the near or distant future is difficult to say; nonetheless, the research on consciousness is of paramount importance both for its theoretical aspects as well as for its consequences and applications. If there will be a solution at all, it is more likely that it will come from the synergy between philosophical and scientific research. As we can say: philosophy without science is empty, and science without philosophy is blind. For this reason, philosophical analyses of neuroscientific theories of consciousness are useful if not mandatory. There is a long way to go, but a journey of a thousand miles begins with a single step.
References:


