CONSCIOUSNESS AND COMPLEXITY: NEUROBIOLOGICAL NATURALISM AND INTEGRATED INFORMATION THEORY

[Final Preprint, published at https://doi.org/10.1016/j.concog.2022.103281]

Francesco Ellia*[1] & Robert Chis-Ciure[1,2]

* Corresponding author
[1] School of Medicine, Department of Psychiatry, University of Wisconsin-Madison.
[2] Faculty of Philosophy, Department of Theoretical Philosophy, University of Bucharest.

Abstract

In this paper we take a meta-theoretical stance and aim to compare and assess two conceptual frameworks that endeavor to explain phenomenal experience. In particular, we compare Feinberg & Mallatt's Neurobiological Naturalism (NN) and Tononi's and colleagues Integrated Information Theory (IIT), given that the former pointed out some similarities between the two theories (Feinberg & Mallatt 2016c-d). To probe their similarity, we first give a general introduction into both frameworks. Next, we expound a ground-plan for carrying out our analysis. We move on to articulate a philosophical profile of NN and IIT, addressing their ontological commitments and epistemological foundations. Finally, we compare the two point-by-point, also discussing how they stand on the issue of artificial consciousness.

Keywords: Consciousness; Neurobiological Naturalism; Integrated Information Theory; artificial consciousness; neuroscientific theories of consciousness.

1. Introduction[1]

Over the past 30 years a new interdisciplinary field of study has emerged, namely the *science of consciousness*. While this new field grew rapidly and enthusiastically, with plenty of ideas coming from both science and philosophy, it still lacks an organizing set of principles that can turn objectively measured brain data into proper knowledge of subjective experience[2]. On one hand, traditional philosophical questions, such as whether consciousness is exclusively a human affair, or instead is spread across the living world (and perhaps beyond that), are now more open than ever. On the other hand, new problems have been formulated, e.g., the minimal set of neural mechanisms jointly necessary and sufficient for an experience in general.

In this paper we compare and evaluate two prominent scientific models of consciousness, *Neurobiological Naturalism* (NN) and *Integrated Information Theory* (IIT). In general, a model of consciousness is a theoretical description that relates physical properties of the brain to phenomenal properties of consciousness (Seth 2007). One motive for the present analysis stems from Wiese's (2020) call for "minimal unifying models" (MUM). In his view, a MUM of consciousness: (i) will specify only necessary properties of consciousness, but not stronger, sufficient ones; (ii) it has determinable descriptions that can be further sharpened and made more specific; and (iii) it unifies to models of consciousness by revealing their shared assumptions. Every individual model of consciousness has an important conceptual component that operationalizes its *explanandum*, namely the way in which it empirically defines phenomenal experience—a component that is necessary because experience is eminently subjective and not immediately amenable to a translation into physical terms. Let us call this component the *framework assumption*

[1] Both authors contributed equally to this work.
[2] A similar claim is made by Sporns (2015: 95).

of the theory. Our aim in this paper is to seek common framework assumptions in NN and IIT, with a prospect for unifying these models in Wiese's minimal sense. To this end, we consider how compatible these two approaches are on their ontological and epistemological commitments, and the subsequent implications for the science of consciousness.

Moreover, our contribution fits in a broader and emerging debate about models of consciousness, where scholars have either proposed comparison between different models (Del Pin et al. 2020, 2021; Sattin et al. 2021; Signorelli et al. 2021), convergence of elements across different features (Northoff & Lamme 2020; Sarasso et al. 2021; Rorot 2021) or comparison of different paradigms on empirical grounds (Doerig et al. 2020; Melloni et al. 2021).

A further rationale for this project came from the NN authors themselves, Feinberg & Mallatt (2016c: 27) pointing out some similarities between their model and IIT:

> "Similarly, Giulio Tononi says that the amount of consciousness in a system is the quantity of *integrated information* generated by the system's elements and their interactions beyond the quantity generated by the individual parts of the system. More of this complex information equals more consciousness. Along with Christof Koch, Tononi especially emphasizes that organized interactions and feedback between neuronal centers are important for consciousness, with which we agree." (Feinberg & Mallatt 2016c: 27)

What is more, in Feinberg & Mallatt (2016d: 124), the authors point out that NN shares some tenets with "theories that focus on recurrent neuronal interactions and feedback loops, *information integration* [our emphasis], oscillatory binding, neural coding strategies, or other brain processes that contribute to the creation of consciousness".

Because Feinberg & Mallatt found these similarities, we decided to look further, to see if major differences exist at a deeper level. Following our initiative, Mallatt proposes his own analysis of the differences and similarities between IIT and NN (Mallatt 2021).

To kickstart the discussion, we give a general description of the theories in Section 2. Next, in Section 3, we outline our blueprint for comparing the two theories. Then, in Section 4, we put forward our meta-theoretical analysis. In the first part of the section, we revisit the two theories from another point of view, by emphasizing three ontological and two epistemological dimensions on which we base our analysis. Second, we provide a point-by-point comparison and critical evaluation of the models' ontological-epistemological profiles. In Section 5, we go on to contrast the theories on the issue of machine consciousness, which is a topic of possible disagreement and a locus of empirical discriminability. The result of our analysis is that the two theories differ in their frameworks and in their epistemological and ontological assumptions.

2. Theoretical Background

This section contains a general introduction to Neurobiological Naturalism and Integrated Information Theory. Far from being a comprehensive account, it is meant to give an overview of the theoretical aims and main conceptual resources used in the two frameworks.

2.1. Neurobiological Naturalism (NN)

NN aims to explain phenomenal experience[3] (Revonsuo 2006) in its exteroceptive, interoceptive, and affective aspects (Feinberg & Mallatt 2016c). In the view of NN proponents, primary

[3] Also called phenomenal consciousness, primary consciousness, and sensory consciousness. In this paper we treat phenomenal experience, phenomenal consciousness, primary consciousness and sensory consciousness as synonyms. However, nothing substantial in the argument depends on this terminological choice.

consciousness has multiple features needing to be explained (explanatory gaps), namely referral, mental unity, mental causation, and qualia[4], and poses a Hard Problem for science to solve (Levine 1983; Chalmers 1995). Besides these features, the great inter-species diversity in brain anatomy makes it unlikely that one single mechanism can be sufficient for consciousness. Thus, the authors combine three explanatory domains: neurobiological, neuroevolutionary, and neurophilosophical, invoking several biological mechanisms, that span multiple levels of physical organization (Feinberg & Mallatt 2016a).

In its attempt to account for phenomenal consciousness (raw "feelings"), NN builds on complex systems theory (Salthe 1985) and a sequence of explanatory steps, going from life properties to neurons, reflexes, core brain functions, ending with special neurobiological features that are characteristic of consciousness (Feinberg & Mallatt 2019). As such, the theory covers three main levels. The first level concerns *general biological features*, which are not conscious but are necessary for consciousness to appear at a higher level. These biological features are characteristic of all living beings and include life as an embodied process; that is, an organism separated from its environment by a boundary, allowing it ultimately to have a first-person perspective. Moreover, life's fundamental unit of organization is the cell, with cellular interactions that occur both at one level and across several hierarchical levels of organization. From a structural and functional point of view, the concepts of system, process, and hierarchy are central to the first-level requirements for primary consciousness. The body that will become the subject of experience is systemically organized, meaning that the concerted interactions between its parts are of the essence. Furthermore, it is inherently a collection of processes, since its individual parts are mechanisms that perform specific actions. Being a biological agent, the conscious organism displays goal-directed, adaptive behaviors, meant to ensure its survival.

The second level adds some new features to the hierarchy, features that do not achieve consciousness but without which phenomenal experience cannot emerge at the next level. These level-two features are a multicellular body, neurons, neuronal reflex arcs, and then simple, core brains. The neurons communicate through action potentials and synapses for fast signaling. In the more advanced animals, at this level, the neural connectivity is extensive enough to allow complex reflexes and basic motor programs, as well as core-brain functions like homeostasis and arousal (Feinberg & Mallatt 2016d, 2018a: ch. 6).

The *special neurobiological features* at the third level add the final requirements for experience. These are Feinberg and Mallatt's version of the neural correlates of consciousness, which are traits that almost every theory of consciousness seeks to identify (Blackmore & Troscianko 2018; Crick & Koch 1990). In NN, these features begin with increased neural complexity: a large number of neurons (min. ~100,000) diversified in type and connectivity. As the next neurobiological feature, the animal must have elaborated sensory organs tuned to its specific econiche for: vision, olfaction, hearing and balance, taste, touch detection, and proprioception. A particular characteristic of the neural hierarchies at this third level is that they need not be physically nested (like, e.g., tissues of a kidney), meaning different groups of neurons can be segregated and far apart, yet have an extreme degree of axonal and synaptic interconnection to allow information integration. The non-nested neural-neural synaptic interactions allow very fast and coordinated processing across multiple hierarchical levels, with plenty of feedforward, intra-level, and recurrent communication. This allows for the convergence of sensory processing across different modalities (binding together what is seen, heard, and touched) and prediction of future sensory inputs, enabled by the distributed yet integrated system architecture, which balances functional segregation and global coherence. These neural hierarchies create sensory images via topographic maps of the outside world and body structures; moreover, they make possible affective or emotional states through

[4] Called the neuroontologically subjective features of consciousness.

valence coding (i.e., marking sensory inputs as good or bad), and output to pre-motor regions to determine movements in space. All these features are complemented by increased selective attention and memory capacity (Feinberg & Mallatt 2018: ch. 6, 2019).

In their earlier work, Feinberg & Mallatt (2013, 2016a-d) seemed to be declaring their special neurobiological features by fiat, but these features were actually deduced from two basic assumptions, as their recent publications have made clear (Feinberg & Mallatt 2018b, 2019; Mallatt et al. 2020). These core assumptions are: (i) an organism experiences mapped mental images if it has demonstrably mapped neural representations; and (ii) an organism has affective or emotional consciousness if it has the capacity for complex operant learning from rewards and punishments. The animal clades that fit these two criteria were then examined to find additional shared features that relate to consciousness (Feinberg & Mallatt 2019: 3). This approach of building from assumptions is at least superficially similar to that of IIT, which was built from axioms of phenomenology (see next section).

Applying these assumptions and deductions, Feinberg and Mallatt reasoned that primary consciousness emerged during the Cambrian explosion, roughly 560-520 million years ago, independently in several animal phyla, including all vertebrates and arthropods, and the cephalopod mollusks (Feinberg & Mallatt 2013). By identifying the special neurobiological features of consciousness and explaining how they arose during brain evolution, NN worked to eliminate the explanatory gaps. However, the unique nature of consciousness creates *experiential gaps*, which, unlike the explanatory ones, cannot be closed but only bridged through scientific explanation (Feinberg & Mallatt 2016d). Recently this point was clarified in the following way: consciousness can be *explained* through its neural mechanisms, but can only be *experienced* by the subject (mind-reading by outside observers being impossible), and that this subjective/objective divide does not violate any physical law (Feinberg & Mallatt 2020). NN considers consciousness as a "unique multi-determined system feature of life and complex brains" (Feinberg & Mallatt 2019).

2.2. Integrated Information Theory (IIT)

IIT is a contemporary neuroscientific theory that aims to explain consciousness in terms of the *integrated information* present in a physical substrate (Oizumi et al. 2014; Tononi 2015; Tononi et al. 2016). A substrate is defined as a system of connected units in a state (e.g., a set of active or inactive neurons in a brain). Simply put, only information specified by a whole over and above that specified by its part can be called integrated information. For a given physical system, its integrated information is assessed by unfolding its *cause-effect structure*, which in turn captures how the elements of the system constrain its past and future states. IIT takes a different methodological approach than other scientific theories of consciousness: a *phenomenology-first approach*. So rather than starting bottom-up from neural correlates and neural mechanisms and then proceeding to explain experience thusly, IIT begins with phenomenology and then arrives at the mechanisms of consciousness.

As such, the theory puts forward five *axioms* derived from reflection on our consciousness, meant to capture the essential properties of every possible experience. The axioms describe the fabric of consciousness, so no experience can fail to satisfy these properties. In IIT, these essential properties are: *intrinsicality*, *composition*, *information*, *integration*, and *exclusion*. By the axioms, a conscious experience is: (i) intrinsic = exists for its own subject, and it cannot be experienced by an external observer[5]; (ii) structured = is composed by phenomenal distinctions bound by

[5] In IIT, the concept of intrinsicality captures the idea of a first-person, subjective perspective. This is different from the way in which NN authors use the concept, insofar as they connect it with life as embodied process, which functions

relations; (iii) informative = is the particular way it is, meaning that it is specific; (iv) integrated = is unitary and, thus, not reducible to any of its parts (distinctions and relations); and (v) definite = has borders, in the sense that it has definite content (contains what it contains, neither less nor more) (Oizumi et al. 2014; Tononi 2015; Tononi et al. 2016; Haun & Tononi 2019).

To each axiom corresponds a *postulate*, which in conjunction describe the ontological (causal) properties of the physical substrate of consciousness. Briefly, the postulates provide the causal reasons why an experience is as the axioms describe it. Thus, according to the postulates, for a physical substrate to underlie experience, it must have *intrinsic*, *compositional*, *specific*, *integrated* and *maximal cause-effect power*. A proper substrate specifies a Maximally Irreducible Cause-Effect Structure (Haun & Tononi 2019: Section 2.4). IIT goes on to posit a fundamental *identity* between an experience and the Cause-Effect Structure of the physical substrate: a particular conscious experience is a particular Maximally Irreducible Cause-Effect Structure (CES[6]). The CES is described in causal information terms, with integrated information Φ quantifying its irreducibility. Notably, candidate physical substrates of consciousness are characterized from a topological rather than functional point of view, and this allows for two functionally identical systems to be phenomenologically distinct. This means that two systems given an identical set of inputs can provide the same set of outputs, and yet present radically different phenomenological properties (Oizumi et al. 2014; Grasso et al. 2021)[7]. Moreover, as long as their causal structures are identical, this allows for the same experience to be multiply realizable by different substrates, a point also emphasized by Feinberg & Mallatt (2020).

It is worth mentioning that, while IIT *per se* has not yet dealt directly with the evolutionarily adaptive value of consciousness, computational models suggest that, for an organism, the exposure to environments richer in complexity may lead to an increase in internal connectivity and richer intrinsic Cause-Effect Structures (Albantakis et al. 2014, Albantakis & Tononi 2015, Albantakis et al. 2020; Albantakis 2018, 2020; Juel et al. 2019; Grasso et al. 2021). IIT makes a vast and diversified set of predictions that can in principle falsify the theory if disconfirmed by empirical evidence (Tononi et al. 2016; Tsuchiya et al. 2020; Ellia et al. 2021). Finally, some preliminary empirical confirmations of IIT are given by several measures of brain complexity derived from the theory's formalism and their effectiveness in predicting recovery in patients with disorders of consciousness (Casali et al 2013; Massimini & Tononi 2018: ch. 6).

3. A Blueprint for Comparing NN and IIT

We try to find common grounds between NN and IIT based on Feinberg and Mallatt's claim that they are similar. We proceed on philosophical and conceptual rather than technical grounds for two reasons. First, the two theories are so divergent in the theoretical resources they employ that it is very difficult to see how someone could contrast and evaluate them on specific technical details. For instance, both theories give an important role to the principle of segregation/differentiation and integration/global coherence coexisting in the brain as a system-design feature. However, NN makes this principle one of its special neurobiological features

as individuating condition for the organism. However, they would not presumably identify these notions, since subjectivity requires all levels of biological explanation, with embodiment as necessary, but not sufficient, condition.

[6] A clarification on IIT's terminology is in order. Every Maximally Irreducible Cause-Effect Structure is a Cause-Effect Structure, but not all Cause-Effect Structures are maximally irreducible. In fact, the latter refers specifically to those Cause-Effect Structures that specify maxima of Φ, thus satisfying the postulate of exclusion and the requirements that IIT poses for the physical substrate of consciousness. In this paper, for ease of exposition, "CES" stands for a Maximally Irreducible Cause-Effect Structure.

[7] Due to this last point, behavioral and functional evidence are not to be taken sufficient for consciousness (Tsuchiya et al. 2020; Ellia et al. 2021).

necessary for primary experience, made possible by extended neural hierarchies and by many neuron-neuron interactions[8]. In contrast, IIT takes a computational, network stance, abstracting beyond the neural substrate and measuring integration in terms of its own information formalism, using partitions to assess the existence of joint cause-effect constraints on the state space of the system posed by a candidate mechanism over and above its parts[9]. In short, NN is more concrete and IIT is more abstract in their treatment of system-design, making a technical comparison difficult.

Second, we believe that an alleged unification is more dependent on common philosophical assumptions about the *explanandum* than on specification of the technical details. This coheres well with Wiese's (2020) conception of minimal unifying models, which aims to unearth general commonalities between theories. For example, if the two theories have radically different metaphysical positions about consciousness, then their unification becomes immediately problematic, because there is an incompatibility in framework assumptions rather than in theory construction. This being said, we grant that there is a continuous range of positions one can take when explaining experience, with some positions more closely related than others, but this is just a further reason to focus on spelling out the (dis)similarities of the extremes, which is the approach we take in this paper. What is more, we often see shared constraints between the ontology and the epistemology of a theory, even though ontology and epistemology are separable to a point. To exemplify, if a theory holds as an ontological posit that consciousness is fundamental, then this constrains the kinds of epistemological possibilities by removing those options of reducing consciousness to other phenomena. Or, if a theory explains consciousness as epiphenomenal, it cannot at the same time maintain that it constrains its realizing parts through mental causation. In conclusion, comparing how NN and IIT explain phenomenal consciousness will be valuable for assessing their alleged compatibility.

4. Discussion

We endeavor to analyze the theories on two dimensions. *Ontology* is the first dimension on which we compare NN and IIT, asking three questions:

> What is the nature of consciousness according to the theory?
> In what way does consciousness exist according to the theory?
> What is the relation between experience and its substrate within the theory?

Epistemology is the second dimension of comparison, where we ask two questions:

> How does the theory distinguish between conscious and nonconscious systems?
> How does the theory explain the character of consciousness in a system?

Sections 4.1 and 4.2 construct ontologically- and epistemologically-guided profiles of the two theories from a neutral perspective. Then section 4.3 assesses the claims of similarity on principled, conceptual grounds, a task that requires such a meta-theoretical analysis.

4.1. IIT's Ontology and Epistemology

Ontology. As stated above, IIT proponents take consciousness to be ontologically basic, in the sense that it exists fundamentally. Moreover, IIT holds that every experience is identical with a CES or

[8] See Feinberg & Mallatt (2018, ch. 6; 2019, Table 2).
[9] See Oizumi et al. (2014, Models section).

quale[10] specified by a system in a state. Note from the start that the identity is not between experience and the physical substrate of consciousness, so IIT does not collapse into standard identity theories of consciousness (e.g., Place 1956). Rather, the identity is between experience and the CES specified by that system. Importantly, the irreducibility of the CES should not be taken as our epistemic inability to further partition the system into subsystems without loss of information, but as a genuine feature of reality. Hence, integrated information reflects the ontological nature of a system in a state rather than just our ability to learn about it. The CES is structured according to the causal distinctions and relations that compose it[11].

Nevertheless, as pointed out by Tononi (2015, 2017), even though there is a distinction between the substrate and the CES, the substrate does *not* exist independently from the cause-effect structure, but rather it exists *as* that structure. Indeed, in IIT, what truly exists is the *unfolded substrate*, which is discovered through the algorithmic procedure of assessing its cause-effect power. The unfolded substrate just *is* the CES. Moreover, the cause-effect structure itself is *physical*[12], since it has a particular causal nature, i.e., certain cause-effect properties. Physicalism is an operational principle of IIT, which states that something is said to exists in a physical sense only if it has cause-effect power that can be assessed through observation and manipulation ("make/take a difference"). Cause-effect power is the *criterion* of physical existence. Given this definition, since a Cause-Effect Structure is essentially causal, it means that it is also physical.

To understand this claim it is important to distinguish between *being* and *describing*; a distinction drawn in a similar way by NN proponents also. In IIT, consciousness is a "way of being". On the other hand, description can provide an understanding of whether there is any experience and in what way in a certain system, but description is not identical with experience (i.e., by describing an experience we are not feeling it). This means that, when an observer wants to know what an entity experiences, that observer cannot *have* those experiences from an *intrinsic perspective*, yet she might be able to *describe* them from an *extrinsic perspective*. The main aim of IIT is to give a methodical and mathematically well-defined way to achieve this description (Tononi 2015).

Once the ontological identity between an experience and its corresponding Cause-Effect Structure and the distinction between being and describing are in place, IIT claims that we find experience wherever there is a Φ^{Max} complex; i.e., one that specifies a maxima of integrated information, relative to all the possible overlaps. IIT allows consciousness to be realized in multiple substrates, in fact the substrate *per se* is irrelevant, as long as it specifies the integrated information described by the mathematical formalism of the theory.

Epistemology. IIT's epistemology flows rather smoothly from its ontological posits, since it is based on the properties of the CES. The following quote shows this by saying that the ontologically "real" quality and quantity of experience are neatly specified, epistemologically, by:

> "A [cause-effect][13] structure completely specifies both the quantity and the quality of experience: *how much* the system exists—the quantity or level of consciousness—is measure by its Φ^{max} value—the intrinsic irreducibility

[10] More specifically, there are two senses in which one can use "quale" within the context of IIT. In the broader sense, a quale is the entire experience at a moment (CES). In a narrower sense, it is a particular content of an experience (a sub-structure of the CES) (Oizumi et al. 2014). In this article we use only the former when discussing IIT.

[11] It is useful to use this terminology because IIT does not employ the standard Shannon notion of information as bits coded in a message from outside a system, but rather proposes its own notion of information as a system's internal, causal power (Barbosa et al. 2020, 2021).

[12] Note that, historically, the well-accepted notion of physicalism had been very problematic to define (Goff 2019: ch. 3; Stoljar 2021). Because IIT explains the phenomenal in physical in terms, its characterization of physical is non-problematic: the physical is defined in terms of having cause-effect power as it can be assessed within one's own experience. It is physical anything that can affect and be affected by something else.

[13] We replace every "conceptual" by "cause-effect" in the quote from Tononi & Koch (2015: 9) to keep consistent with the newer terminology introduced in Haun & Tononi (2019) and Ellia et al. (2021).

> of the [cause-effect] structure; *which way* it exists—the quality of content of consciousness—is specified by the shape of the [cause-effect] structure." (Tononi & Koch 2015: 9)

Notably, consciousness is *fundamental* in an epistemological as well as an ontological sense. This is because our epistemological inquiry into the physical world (which can be carved at its joints by assessing the cause-effect structures that compose it) started through consciousness, so we are not reducing the phenomenal to the physical. As noted by Tononi (2015), "[t]his is because the existence of one's consciousness and its other essential properties is certain, whereas the existence and properties of the physical world are conjectures, though very good ones, made from within our own consciousness". Because IIT does not need to reduce the phenomenal to the physical, it can face up to the Hard Problem of Consciousness (see Chis-Ciure & Ellia 2021).

To answer the ontological questions that we asked at the start of Section 4, IIT holds that consciousness is fundamental and cannot be reduced to anything else. However, within our own experience, any system can be studied in physical terms (i.e., in cause-effect power terms). IIT then posits that only systems that specify a Maximally Irreducible Cause-Effect Structure, captured by Φ^{Max}, exist as conscious entities; hence the theory offers a way to carve nature at its joints.

To briefly answer the epistemological questions, according to IIT, detection of a Φ^{Max} complex means that the system is conscious. The value of Φ^{Max} indicates the degree or quantity of consciousness present and the particular form of the quale or CES, its composition of causal distinctions bound by relations, is how the experience feels like, i.e., its qualitative character.

4.2. NN's Ontology and Epistemology

As the name suggests, this theory explains consciousness as a natural phenomenon, without any appeal to "mysterious" (occult) or new "fundamental" physical processes (Feinberg & Mallatt 2016d). We already mentioned the multiple explanatory gaps, i.e., referral, mental unity, mental causation, qualia, gaps that the authors consider as contributing to the puzzling character of consciousness. After saying that subjective experience cannot be reduced to objective neural processes, they proceed to seek the evolutionary origins of the said explanatory gaps and account for them in terms of "conventional biological principles" (Feinberg & Mallatt 2016c: 14). They found that the four explanatory gaps can be successfully closed (explained away or filled) but that two *ontological* (real) gaps remain and can only be "bridged" (understood) but not closed. These indelible gaps are the *auto-ontological irreducibility*, meaning that one's subjective experience does not refer to the objective neurons that create it, and the *allo-ontological irreducibility*, meaning that an objective observer cannot directly access or measure a subject's experiences[14] (Feinberg & Mallatt 2016d, 2018: ch. 8).

NN proponents emphasize that consciousness is a very diverse phenomenon, both within-brain and inter-species (Feinberg & Mallatt 2016b). Inside a single brain, there are many mechanisms responsible for different features of sensory experience, such as mental images vs affects and qualia vs mental causation[15]. In the same vein, consciousness is widely spread in the animal kingdom, being realized in the very different brains of vertebrates, arthropods, and cephalopods, all of which

[14] Although these irreducibilities are *real* barriers and thus are ontological, they are also barriers to *learning* about consciousness and thus are epistemological. Because they are not purely ontological, that word was recently removed from their names, which were shortened to *allo-* and *auto-irreducibilities* (Feinberg & Mallatt 2018b, 2019). We kept the word in the main text to make the contrast to the explanatory gaps more transparent.

[15] Important for this diversity topic, but beyond the scope of the present article, it is still debated in the literature on brain connectivity if one can speak of a neuro-typical brain within a species. In fact, the immense variety of brain connectivity pathways between subjects represents the biggest challenge for the science of connectomes (Sporns 2011).

however share the general and special features for consciousness (we referred to them as levels 1-3 in Section 2.1).

NN authors dissociate between the explanation of qualia in terms of their neural mechanisms and the explanation of their subjective character. The neurobiological basis of qualia and the causes of subjectivity are not identical, meaning that only a combination of the (i) unique neurobiological features for qualia with the (ii) unique ontological features of subjectivity (the auto- and allo-irreducibilities) can account for the unique subjectivity of qualia. In this sense, the neural features in (i) answer the question of how qualia come into being, while the irreducibilities in (ii) answer the question about the subjective character of qualia. This is the reason why the authors appeal to both neurobiology and philosophy in explaining consciousness in an evolutionary context.

According to NN, life is a fundamental prerequisite of consciousness. The biological domain thus bounds the phenomenal one. Life is obviously not sufficient by itself, so the neural reflexes, core brain functions and special neurobiological features found in more complex brains must be added. Thus, in this framework, consciousness is a uniquely biological emergent system-feature, built on the foundations of life processes and on the scaffolding of complex brains.

To answer the ontological questions that we asked at the start of Section 4, NN holds that consciousness is physical by nature; that it is not fundamental but an emergent process in systems displaying certain features; and that it is confined to the biological realm of evolving brains whose complex neural networks are the substrate that generate the experiences. Note however that NN's concepts of 'complexity' and 'emergence' are not defined in a mathematically strict way, which suggests a possible avenue for future research.

Now let us address how NN answers the *epistemological* questions, the first of which was how it distinguishes between conscious and nonconscious systems. In the NN framework, any normally functioning brain that satisfies the general life properties, has reflexes and also the special neurobiological features (see Section 2.1), is capable of primary consciousness:

> "[T]he combination of life and reflexes, the special features, and auto- and allo-ontological irreducibilities can account for how both subjectivity and the unique phenomenon of consciousness are naturally created." (Feinberg & Mallatt 2018: 120.)

Anything that does not satisfy these conditions is not conscious.

The next epistemological question is how NN explains the character of consciousness. The quality of phenomenal experience or its character lies in the differentiation of neural states themselves. The feeling of "red" is both mechanistically unique and subjective, so it is once again explained by the convergence of all factors, i.e., of subjectivity and brain networks.

4.3. Ontological and Epistemological Divergence

The stark contrast between the two theories' metaphysical assumptions is by now clear. Their commonalities notwithstanding[16], they differ greatly on how they take consciousness to exist, and in how they spell the relation between consciousness and its substrate. While IIT treats experience as fundamental in its existence, NN takes it to be an emergent feature arising at some point when certain complexity criteria are met. While IIT claims the only things that matter for experience are the architectural principles and topological features of a dynamical system constraining its state

[16] As noted in the previous sections, these commonalities include: commitment to known physical processes and not the occult; emphasis on organized feedback interactions; the use of axioms in IIT and the two assumptions in NN; the distinction between being (i.e., having an experience) and describing (i.e., characterizing an experience); and that both entail the multiple realizability of consciousness.

space, NN considers life processes necessary and thus only a neurobiological substrate as adequate for realizing consciousness.

The contrast persists insofar as their epistemologies are concerned. Both theories attempt to explain phenomenal consciousness, so they both have the same *explanandum*, yet they differ greatly in their *explanans*. While IIT explains the presence, level and quality of consciousness by starting from phenomenological axioms and deriving the physical requirements for the substrate of consciousness, NN uses two basic assumptions to deduce a list of biological principles and special neural features to account for primary experience and its character, while retaining some irreducibilities for which it gives a philosophical account.

Therefore, to sum up point-by-point, the two theories differ at their metaphysical and epistemological core in how they:

- take consciousness to exist: fundamental feature of reality (IIT) *vs* emergent as specified by a three-leveled set of criteria (NN);
- think of the relation between experience and its physical substrate: substrate not important *per se*, only its causal structure (IIT) *vs* only a neurobiological substrate (NN);
- explain the presence and character of consciousness: causal structures modeled in causal information terms (IIT) *vs* by biological principles and features in evolutionary and philosophical context (NN);
- starts with principles derived from phenomenal experience (IIT) *vs* starts with principles derived from the observation of the physical world (NN).

5. The artificial consciousness conundrum

The problem of artificial consciousness is another point of divergence and, for the purpose of this paper it is highly relevant. In particular, it provides a concrete case where the empirical discriminability of the two theories can be investigated. We have shown the difference between general assumptions of the two frameworks; there is also a significant contrast—and even complementarity—between the epistemic strengths and weaknesses of the two. On one hand, IIT has a strong mathematical formalism that, in principle, allows for methodical analysis of e.g., neural networks; yet it is mostly silent about the adaptive value of consciousness (Tononi & Koch 2015). On the other hand, NN provides a rich, detailed account for the evolutionary origins of consciousness, but so far without any mathematical quantification of its principles. A possible unification, which alas seems not currently achievable, would have yielded a theoretical framework of great explanatory power, which was already a sufficient reason to carry out the present analysis.

Be that as it may, the question of artificial consciousness attracts contradicting predictions from the two theories, and could thus serve as a criterion of theory choice in the future. Here by artificial consciousness, we mean experience instantiated by non-living, human-made system in its broadest acceptation[17]. In IIT, artificial consciousness is a real possibility because within this framework the only requirement for consciousness is a positive value of Φ^{Max} (Oizumi et al. 2014; Tononi et al. 2016; Haun & Tononi 2019). Therefore, non-living systems, not necessarily human-made ones only, can potentially sustain consciousness. On the contrary, in NN machine consciousness seems implausible due to the strong biological nature of phenomenal experience. As the authors put it:

> "Our thesis could be challenged in four ways. First, one might argue that isomorphic, topographic representation cannot be equated with consciousness because artificial sensors and computers can receive and

[17] Notice that our definition of artificial consciousness goes beyond virtual systems that simulate neural networks such as contemporary AI and includes different kind of computing systems of different architectures, including but not limited to von Neumann and neuromorphic ones.

> map out stimuli, yet these machines are not conscious. In response, we reiterate that our hypothesis states that sensory consciousness and isomorphic representations entail a highly specific '*kind*' of isomorphic representation, not just *any kind*. The brain possesses an entirely *unique* architecture that features – in addition to a huge 'computer-like' amount of complex processing – reciprocal communication between the levels of the neural hierarchy with integrated and novel emergent properties appearing with the addition of each level. Thus, the neural hierarchy represents a unique neurobiological substrate and organization quite different from that found in computers made of silicon chips and wires. (Feinberg & Mallatt 2013: 15)

Moreover, in an earlier paper, discussing an alleged relation between consciousness and inanimate matter, Feinberg writes:

> "[M]y analysis attempts to explain or derive consciousness from matter with no reference to any unknown physical laws or any new physics or the application of physics to emergence or reduction beyond that normally applied to *biology in general*" (Feinberg 2012: 20; our emphasis).

While both NN authors are skeptical about the possibility of artificial consciousness, they have slightly different views on this topic. Todd Feinberg flatly denies it, yet Jon Mallatt concedes that there could be an exception. In particular, according to Mallatt (personal communication), if it were the case that an artificial system far more advanced than those available today could replicate the complexity of biological systems, and integrate many types of sensory information in a way that allows the machine to move independently, find sustenance for its power, survive and reproduce in nature, then such machine could be considered conscious. However, such a machine is not achievable in the foreseeable future according to Mallatt. Therefore, for all practical purposes, the NN authors agree that the life requirement is the safest assumptions given present evidence (Mallatt, personal communication; Mallatt 2021).

Returning to IIT, the theory says that any physical substrate can underlie conscious experience as long as it has maximally irreducible, specific, compositional, intrinsic cause-effect power (Oizumi et al. 2014). Therefore, even very simple non-living systems such as an eight-node grid can, in principle, have an experience (Haun & Tononi 2019). However, it is important to reemphasize that the main requirement for consciousness according to IIT is true causal power. A *simulated* system, such as an AI software, cannot yield any cause-effect power just as a simulated ocean is not wet and a simulated black hole does not bend space-time (Tononi & Koch 2015). Therefore, when questioning the presence of consciousness, we should not consider the software but the hardware. And the main problem of current chipsets is that they are based on a von Neumann architecture with a strictly feed-forward modular organization so, under IIT's analysis, it fails to satisfy the integration postulate. However, if a different physical architecture could be implemented, such as, e.g., a silicon-based neuromorphic substrate with feedback connectivity able to satisfy the causal requirements of the postulates, it would have consciousness (Koch 2019). Finally, it is worth pointing out that IIT is agnostic about whether consciousness exists in a specific physical system until such system can be unfolded to determine whether there is a positive value of Φ^{Max}.

6. Conclusion

We opened the discussion with descriptions of both NN and IIT as models of consciousness that aim to explain phenomenal experience. Subsequently, we articulated a philosophical profile of both theories, trying to distill their ontological and epistemological tenets. We then proceeded to contrast them. Thus, on the ontological side, IIT takes experience to be a fundamental feature of the world. NN considers experience as an emergent phenomenon. In IIT, an experience is identical with a maximally irreducible cause-effect structure, hence prioritizing the causal structure of a substrate over any other of its specifics, e.g., the particular kind of its components. In NN, the emergence of consciousness depends upon the satisfaction of certain complexity criteria of

biological and neurobiological organization, hence restricting the possible substrates to the biological domain.

On the epistemological side, IIT detects the presence of consciousness in a system by finding if it has a Maximally Irreducible Cause-Effect Structure, quantified by Φ^{Max}. At the same time, the phenomenal character of that experience is explained in cause-effect language, which can be expressed in mathematical terms. NN demarcates conscious from nonconscious systems by a three-leveled list of general biological and special neurobiological criteria (e.g., cellular embodied life actively organized in hierarchical systems, with high neural complexity given by neuron number, types, and connectivity, enabling topographic maps and valence coding, arousal, attention, and memory). Then NN explains the character of consciousness both mechanistically and philosophically, accepting its irreducible subjectivity.

The last point of comparison was the issue of artificial consciousness. IIT predicts that this is a real possibility, for any system that presents the right causal structure, with no constraints on the kind of such a system, e.g., natural or artificial. By contrast, NN considers consciousness as a uniquely biological phenomenon, hence excluding artificial systems, at least on a practical level for the time being.

Our conclusion is that, while both IIT and NN present ambitious goals, they are metaphysically too dissimilar to unify, even in the minimal sense of Wiese (2020), because they don't share enough core assumptions. A unification seems not only unaccomplished but *unaccomplishable* at the moment, due to the radical differences in their framework assumptions. Our skepticism is strictly based on the results of our present analysis, but we do not reject a priori any possible developments that might alleviate the disparities. On the contrary, we look forward to future work—both on separate and common grounds.

References

Albantakis, L., Hintze, A., Koch, C., Adami, C., Tononi, G. (2014). Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. *PLoS Computational Biology*, 10 (12), 1-19.

Albantakis, L., Tononi, G. (2015). The Intrinsic Cause-Effect Power of Discrete Dynamical Systems – From Elementary Cellular Automata to Adapting Animats. *Entropy*, *17*, 5472-5502; doi:10.3390/e17085472.

Albantakis, L. (2018). "A Tale of Two Animats: What Does It Take to Have Goals?", in *Wandering Towards a Goal*, Aguirre, A., Foster, B., Merali, Z. (eds.), Berlin: Springer.

Albantakis, L. (2020). "Integrated Information Theory", in *Beyond Neural Correlates of Consciousness*, Overgaard, M. et al. (eds.), Taylor & Francis Group, ProQuest Ebook.

Albantakis, L., Massari, F., Beheler-Amass, M., Tononi, G. (2020). A macro agent and its actions, arXiv:2004.00058v1.

Barbosa, L. S., Marshall, W., Albantakis, L., & Tononi, G. (2021). Mechanism Integrated Information. *Entropy*, *23*(3), 362. https://doi.org/10.3390/e23030362

Barbosa, L. S., Marshall, W., Streipert, S., Albantakis, L., & Tononi, G. (2020). A measure for intrinsic information. *Scientific Reports*, *10*(1), 18803. https://doi.org/10.1038/s41598-020-75943-4

Blackmore, S., Troscianko, E. (2018). *Consciousness: An Introduction*, New York: Routledge.

Casali, A. et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5, 198ra105.

Chalmers, D. (1995/2010). "Facing Up to the Problem of Consciousness", in *The Character of Consciousness* (pp. 3-34), Chalmers, D. (auth.), New York: OUP.

Chalmers, D. (2009/2010). "The Two-Dimensional Argument Against Materialism", in *The Character of Consciousness* (pp. 141-205), Chalmers, D. (auth.), New York: OUP.

Chis-Ciure, Robert (2022). *The A Priori Foundations of Integrated Information Theory. Toward a Transcendental Science of Consciousness.* Doctoral Dissertation. University of Bucharest.

Chis-Ciure, R., Ellia, F. (2021). Facing up to the Hard Problem as an Integrated Information Theorist, *Foundations of Science*, Berlin: Springer. https://doi.org/10.1007/s10699-020-09724-7.

Crick, F., Koch, C. (1990). Towards a neurobiological theory of consciousness. *The Neurosciences*, 2, 263-275.

Del Pin, S., Skóra, Z., Sandberg, K., Overgaard, M., Wierzschon, M. (2020). Comparing theories of consciousness: Object position, not probe modality, reliably influences experience and accuracy in object recognition tasks, *Consciousness and Cognition*, 84, 102990, https://doi.org/10.1016/j.concog.2020.102990.

Del Pin, S., Skóra, Z., Sandberg, K., Overgaard, M., Wierzschon, M. (2021). Comparing theories of consciousness: why it matters and how to do it, *Neuroscience of Consciousness*, 7(2), 1-8, https://doi.org/10.1093/nc/niab019.

Doerig, A., Schurger, A., Herzog, M. (2021). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*, 12(2), 41–62, https://doi.org/10.1080/17588928.2020.1772214.

Ellia, F., (2021) *Integrated Information Theory: An Empirically Testable Solution to the Mind-Body Problem.* Doctoral Dissertation. Alma Mater Studiorum Università di Bologna.

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., Lang, J., Haun, A., Albantakis, L., Boly, M., Tononi, G. (2021). Consciousness and the fallacy of misplaced objectivity. *Neuroscience of Consciousness*, 2, niab032, https://doi.org/10.1093/nc/niab032.

Feinberg, T. (2012). Neuroontology, neurobiological naturalism, and consciousness: A challenge to scientific reduction and a solution. *Physics of Life Reviews* 9, 13-34.

Feinberg, T., Mallatt, J. (2013). The evolutionary and genetic origins of consciousness in the Cambrian Period over 500 million years ago. *Frontiers in Psychology* 667 (4), 1-27.

Feinberg, T., Mallatt, J. (2016a). "Neurobiological Naturalism", in *Biophysics of Consciousness: A Foundational Approach*, Poznanski, R., Tuszynski, J., Feinberg, T. (eds.), Singapore: World Scientific.

Feinberg, T., Mallatt, J. (2016b). "The evolutionary origins of consciousness", in *Biophysics of Consciousness: A Foundational Approach*, Poznanski, R., Tuszynski, J., Feinberg, T. (eds.), Singapore: World Scientific.

Feinberg, T., Mallatt, J. (2016c). *The Ancient Origins of Consciousness. How the Brain Created Experience*, Cambridge: MIT Press.

Feinberg, T., Mallatt, J. (2016d). The nature of primary consciousness. A new synthesis. *Consciousness and Cognition* 43, 113-127.

Feinberg, T., Mallatt, J. (2018a). *Consciousness demystified.* Cambridge, MA: MIT Press.

Feinberg, T., Mallatt, J. (2018b). Unlocking the "Mystery" of Consciousness, *Scientific American*, in press.

Feinberg, T., Mallatt, J. (2019). Subjectivity "Demystified": Neurobiology, Evolution, and the Explanatory Gap. *Frontiers in Psychology*, 1686 (10), 1-10.

Feinberg, T., Mallatt, J. (2020). Phenomenal Consciousness and Emergence: Eliminating the Explanatory Gap, *Frontiers in Psychology* 11:1041. doi: 10.3389/fpsyg.2020.01041.

Goff, P. (2019). *Galileo's Error*. Oxford: OUP.

Grasso, M., Haun, A., Tononi, G. (2021). Of Maps and Grids. *Neuroscience of Consciousness*, Volume 2021, Issue 2, niab022, https://doi.org/10.1093/nc/niab022.

Grasso, M., Albantakis, L., Lang, J., Tononi, G. (2021) Causal Reductionism and Causal Structures. *Nature Neuroscience*, https://doi.org/10.1038/s41593-021-00911-8.

Haun, A., Tononi, G. (2019). Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy*, 21, 1160.

Juel, B., Comolatti, R., Tononi, G., Albantakis, L. (2019). When is an action caused from within? Quantifying the causal chains leading to actions in simulated agents. arXiv:1904.02995v2.

Koch, C. (2019). *The Feeling of Life Itself: Why Consciousness is Widespread but Can't be Computed*, Cambridge: MIT Press.

Levine, J. (1983). Materialism and phenomenal properties: the explanatory gap. *Pacific Philosophical Quarterly* 64 (4), 354–361.

Mallatt, J. (2021). A Traditional Scientific Perspective on the Integrated Information Theory of Consciousness. *Entropy* 23(6), 650. http://dx.doi.org/10.3390/e23060650

Mallatt, J., Blatt, M.R., Draguhn, A. et al. (2020). Debunking a myth: plant consciousness. *Protoplasma*, http://dx.doi.org/10.1007/s00709-020-01579-w.

Massimini, M., Tononi, G., (2018). *Sizing Up Consciousness: Towards an Objective Measure of the Capacity for Experience*, transl. by Anderson, F., New York: OUP.

Melloni, L., Mudrik, L., Koch, C. (2021). Making the hard problem of consciousness easier, *Science*, 372 (6545), 911-912, 10.1126/science.abj3259.

Northoff, G., & Lamme, V. (2020). Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? Neuroscience & Biobehavioral Reviews, 118, 568–587, https://doi.org/10.1016/j.neubiorev.2020.07.019.

Oizumi, M., Albantakis, L., Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, 10(5).

Place, U. T. (1956). Is Consciousness a Brain Process? *British Journal of Psychology*, 47: 44–50.

Revonsuo, A. (2006). *Inner presence: Consciousness as a biological phenomenon*. Cambridge: MIT Press.

Rorot, W. (2021). Bayesian theories of consciousness: a review in search for a minimal unifying model, *Neuroscience of Consciousness*, 7(2), 1-14, https://doi.org/10.1093/nc/niab038.

Salthe, S. (1985). *Evolving Hierarchical Systems*. New York: Columbia University Press.

Sarasso, S., Casali, A., Casarotto, S., Rosanova, M., Sinigaglia, C, Massimini, M. (2021). Consciousness and complexity: a consilience of evidence, *Neuroscience of Consciousness*, 7(2), 1-24, https://doi.org/10.1093/nc/niab023.

Sattin, D., Magnani, F. G., Bartesaghi, L., Caputo, M., Fittipaldo, A. V., Cacciatore, M., Picozzi, M., Leonardi, M. (2021). Theoretical Models of Consciousness: A Scoping Review. *Brain Sciences*, 11(5), 535, https://doi.org/10.3390/brainsci11050535.

Searle, J. (1992). *The Rediscovery of the Mind*. Cambridge: MIT Press.

Searle, J. (2007). Dualism revisited. *Journal of Physiology – Paris*, 101, 169–178. doi:10.1016/j.jphysparis.2007.11.003.

Seth, A. (2007). Models of consciousness, *Scholarpedia*, 2(1): 1328, doi:10.4249/scholarpedia.1328.

Signorelli, C., Szczotka, J., Prentner, R. (2021). Explanatory profiles of models of consciousness – towards a systematic classification, *Neuroscience of Consciousness*, 7(2), 1-13, https://doi.org/10.1093/nc/niab021.

Sporns, O. (2011). *Networks of the Brain*. Cambridge: MIT Press.

Sporns, O. (2015). "Network Neuroscience", in *The future of the brain: essays by the world's leading neuroscientists*, Marcus, G., Freedman, J. (eds.), Oxford: PUP.

Stoljar, P. (2021). Physicalism. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/physicalism/>.

Tononi, G. (2015). Integrated information theory. *Scholarpedia*, 10(1).

Tononi, G. (2017). "Integrated Information Theory of Consciousness: Some Ontological Considerations", in *The Blackwell Companion to Consciousness*, Schneider, S., Velmans, M. (eds.), Chichester: Wiley Blackwell.

Tononi, G., Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 370 (1668), pp. 20140167.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450-461.

Tsuchiya, N. Andrillon, T., Haun A. (2020). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a truer science of causal structural theories of consciousness, *Consciousness and Cognition*, 79 102877, https://doi.org/10.1016/j.concog.2020.102877.

Wiese, W. (2020). The science of consciousness does not need another theory, it needs a minimal unifying model, *Neuroscience of Consciousness*, 1, niaa013, https://doi.org/10.1093/nc/niaa013.