Abstract: I claim that the integrated information theory of consciousness (IIT) is a complexity science approach to consciousness. In general, complexity science investigates phenomena comprised of numerous non-linearly interacting parts, where global-scale behaviour and structures are irreducible to activities and properties of its constituent parts at local scales. I draw attention to some key features of IIT (i.e. cause–effect power and information integration) and complexity science (i.e. the search for and application of principles and interaction dominance) in order to defend the claim that IIT is properly understood within the broader theoretical framework of complexity science. Doing so has the advantage of making IIT an even more compelling theory of consciousness, which has the added benefit of strengthening its ability to respond to some common criticisms.

1. Introduction

Consciousness research in one form or another goes back to at least the Ancient Greeks and Buddhists around 500 BCE. However, the empirical and quantitative sciences of consciousness are relatively young. It was not until 2016 that the advisory committee thought it was time to change the name of the world’s largest consciousness conference from ‘Toward a Science of Consciousness’ to ‘The Correspondence:
Luis H. Favela, Department of Philosophy and Cognitive Sciences Program, University of Central Florida, USA. Email: luis.favela@ucf.edu

Journal of Consciousness Studies, 26, No. 1–2, 2019, pp. 21–47
Science of Consciousness’ (Arizona Board of Regents, 2017). Apparently, this move was motivated by the belief that a science of consciousness was not only possible but was already happening. The ever-increasing and legitimized empirical literature on consciousness surely lends support to that conclusion. With that said, it is clear that no single framework guides consciousness research even within an individual discipline such as neuroscience, let alone across disciplines such as artificial intelligence, cognitive science, and psychology. One framework, however, is garnering more and more attention within the sciences of mind and philosophy: the integrated information theory of consciousness (IIT).¹

IIT is a scientific theory of consciousness that attempts to capture the essential properties of consciousness and then infer the necessary features of physical systems that are needed to give rise to those properties (Tononi and Koch, 2015). The increasing attention that IIT receives is arguably due to two primary reasons. First, IIT is metaphysically provocative. Specifically, by equating consciousness with integrated information, IIT allows for a very broad range of systems to be conscious (e.g. Tononi, 2008; cf. Schwitzgebel, 2012). Second, IIT offers an epistemically innovative theory of consciousness. Specifically, IIT allows for the quantification of consciousness, which some argue allows it to be the closest means available to measure the amount of consciousness in a system (e.g. Balduzzi and Tononi, 2009; Gamez, 2010). These two strengths, however, are also viewed by some as two of its most fundamental weaknesses. First, it is argued that equating consciousness with integrated information results in metaphysically undesirable ‘consciousness bloat’, or the presence of consciousness everywhere from rocks to nations (e.g. Searle, 2013). Second, it is argued that the calculations required to assess integrated information in systems such as brains makes consciousness computationally intractable and, thus, of little epistemic use in scientific practice (e.g. Cerullo, 2015). While these critiques (as well as others; e.g. Mindt, 2017) are not unwarranted, they do not demonstrate that IIT should be abandoned. Instead, they motivate the need to situate

¹ IIT is the central topic of articles across a range of subfields in the mind sciences, for example in cognitive science (Maguire et al., 2014), computational biology (Cerullo, 2015), neuroscience (Krohn and Ostwald, 2017), and psychology (Barrett, 2014). IIT is also discussed broadly outside of the mind sciences, for example in philosophy (Schwitzgebel, 2015), popular science (Koch, 2017), and even religious studies (Biernacki, 2016).
IIT within a broader theoretical framework in order to overcome the limitations highlighted by those critiques.

I claim that IIT is a complexity science approach to consciousness. Even though there is no single ‘complexity science’, there are typical concepts, methods, and theories that fall under the heading. In general, complexity science investigates phenomena comprised of numerous non-linearly interacting parts, where global-scale behaviour and structures are irreducible to activities and properties of its constituent parts at local scales. By drawing attention to key features of IIT and complexity science, I am then able to defend the claim that IIT is properly understood within the broader theoretical framework of complexity science. Doing so has the advantage of making IIT an even more compelling theory of consciousness, which has the added benefit of strengthening its ability to respond to some common criticisms. I begin with a brief introduction to IIT. Next, I provide a brief introduction to complexity science. Then I elucidate how to understand IIT as being a part of complexity science. I conclude with potential avenues for responses to common criticisms of IIT by way of IIT cast as complexity science.

2. An Introduction to IIT

The inception of the integrated information theory of consciousness (IIT) can be traced back to work by Gerald Edelman, Olaf Sporns, and Giulio Tononi (e.g. Edelman and Tononi, 2000; Tononi, Edelman and Sporns, 1998; Tononi and Sporns, 2003). IIT began to take its current shape in Tononi’s (2004) ‘An Information Integration Theory of Consciousness’. Since then, it has been refined (Oizumi, Albantakis and Tononi, 2014; Tononi, 2008; 2012a; Tononi et al., 2016) and garnered the support of some of the biggest names in consciousness research, most notably Christof Koch (2012). My goal here is not to provide a full explication of IIT, which would require going far beyond the scope of the current work. Instead, I aim to provide a general outline and highlight some of the features that will be most important during later discussion of how IIT falls within a complexity science framework. For detailed introductions to IIT, I recommend Massimini and Tononi (2018), Tononi (2012b; 2015), and Tononi and Koch (2015).

In broad strokes, IIT is a scientific theory of consciousness that aims to capture the essential properties of consciousness and then infer the necessary features a physical system needs in order to give rise to those properties (Tononi, 2015; Tononi and Koch, 2015). The
essential properties of consciousness are referred to as *axioms*. The features of a physical system necessary to give rise to those properties are referred to as *postulates*. Thus, for every axiom there is a corresponding postulate (see Table 1). Before defining the axioms and postulates, it is worth noting at the outset that IIT begins from axioms that originate from phenomenological considerations. As Tononi and colleagues state:

IIT does not start from the brain and ask how it could give rise to experience; instead, it starts from the essential phenomenal properties of experience, or axioms, and infers postulates about the characteristics that are required of its physical substrate. (Tononi *et al.*, 2016, p. 450)

This is noteworthy because it draws attention to one of IIT’s foundational theoretical commitments: a scientific theory of consciousness must account for phenomenology first and properties of underlying physical substrates second. The relationship of axioms (phenomenology) and postulates (physical substrate) makes this approach clear.

<table>
<thead>
<tr>
<th>Axioms/Postulates</th>
<th>Feature of Phenomenology</th>
<th>Feature of Physical System</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intrinsic Existence</strong></td>
<td>An individual’s consciousness is real and from their own perspective</td>
<td>Consciousness has real cause–effect power upon itself</td>
</tr>
<tr>
<td><strong>Composition</strong></td>
<td>The features of conscious states have distinctive structure</td>
<td>Parts underlying structure have cause–effect power on system</td>
</tr>
<tr>
<td><strong>Information</strong></td>
<td>Conscious states are informative and distinct from other conscious states</td>
<td>Physical substrate of conscious state has particular cause–effect properties</td>
</tr>
<tr>
<td><strong>Integration</strong></td>
<td>Conscious states are unified wholes</td>
<td>Physical substrate has irreducible cause–effect structure</td>
</tr>
<tr>
<td><strong>Exclusion</strong></td>
<td>Conscious states are definite</td>
<td>Cause–effect structure of physical substrate is definite</td>
</tr>
</tbody>
</table>

*Table 1. Integrated information theory’s axioms and postulates, and their respective features. (Based on Tononi, 2015, and Tononi and Koch, 2015.)*

*Intrinsic Existence* is the first axiom and postulate. Phenomenologically, this axiom captures the commitment to the existence of consciousness. Specifically, phenomenal states are real and not illusions, and they exist from an intrinsic perspective, namely, independently for
the one having the experience. In terms of the physical substrate, this postulate captures the idea that if consciousness is real then it has physical cause–effect powers that can be measured. Composition is the second axiom and postulate. Phenomenologically, this axiom captures the notion that consciousness has various and distinctive phenomenal qualities, such as the experience of seeing an apple being comprised of distinct redness, smoothness, roundness, etc. In terms of the physical substrate, this postulate captures the idea that those distinct phenomenological qualities are comprised of distinct combinations of elements that have cause–effect powers within that system. Information is the third axiom and postulate. Phenomenologically, this axiom captures the notion that each conscious state is specific and unique from every other conscious state. In terms of the physical substrate, this postulate captures the idea that the underlying elements within the system are also unique and have distinct cause–effect powers. Integration is the fourth axiom and postulate. Phenomenologically, this axiom captures the notion that each conscious state is a unified, global whole that is irreducible to the seemingly individual and distinct qualities of that experience. For example, the experience of seeing an apple cannot be reduced to the mere additive collection of redness, smoothness, roundness, etc. In terms of the physical substrate, this postulate captures the idea that the system’s state is comprised of parts that have unified, holistic cause–effect powers not reducible to the merely additive combination of the cause–effect powers of the parts in isolation. In other words, when a system is integrated, then the cause–effect powers it has globally are different than the sum of the cause–effect powers of its parts. Exclusion is the fifth axiom and postulate. Phenomenologically, this axiom captures the notion that conscious states are definite in their being constant and consistent over space and time. In terms of the physical substrate, this postulate captures the idea that the cause–effect structure of a system is definite such that there is a single physical substrate for each system state.

As the name highlights, the integration and information axioms and postulates are key to the integrated information theory. Consciousness within a system can be quantified via the amount of integrated information, and that value is referred to as Φ (‘phi’). A conscious state is a unified whole that is informative in a definite and unique way (Tononi, 2012a; 2015; Tononi and Koch, 2015). To be a unified whole, the information of that state must be integrated, namely, the state is an irreducible global property that is what it is only in so far as
all of the parts of the whole effect each other maximally, such that no subset or superset of elements can have stronger causal constraints. In order to facilitate understanding the nature of integrated information, I present a version of Tononi’s photodiode thought experiment (2004).

Consider the following three systems: (A) photodiode; (B) digital camera; and (C) a dog’s visual system. System A has one bit of information and no integration; B has one million bits of information and no integration; and C has one billion bits of information that are integrated. According to IIT, systems A and B have the same amount of consciousness, or one bit, which is the smallest amount of consciousness a system could have. System C has a high amount of consciousness. The reason that dogs have more visual consciousness than a photodiode and digital camera is not because it has more bits of information. The reason dogs have more visual consciousness is because their visual information is integrated whereas the information of photodiodes and digital cameras is not. The reason a digital camera with its one million bits of information has the same amount of consciousness as a photodiode with its one bit of information is because the digital camera’s information does not integrate. A photodiode has one state: if exposed to light or not, it is either on or off. A digital camera is like a collection of one million photodiodes: if exposed to light, each part of the lens is either on or off. The reason the digital camera does not sustain a high conscious state is because each of its one million ‘photodiodes’ do not have any cause–effect relationship with each other. Specifically, if one part of the lens is activated or not, no other part of the lens is affected. Dog visual systems, however, are not like digital cameras. If any part of the physical substrate of a conscious state is affected then there are cause–effect ramifications throughout the whole system, which in turn affect the nature of the conscious state. Consider the following: if one part of a digital camera is knocked out, then that would merely result in a single inactive pixel, but the rest of the image would remain the same. If one part of a dog’s visual system is knocked out, then that would result in an entirely different conscious state. In that way, highly conscious systems such as those had by mammals are not just informationally rich; they also have high amounts of integration among that information. Consequently, dogs are high Φ conscious systems whereas digital cameras are not. One way to get a handle on how conscious states are single, unified states of integrated information is via Gestalt psychology.

Although Gestalt psychology has fallen out of favour as a general theory of psychology, many of its lessons concerning the nature of
perception remain influential (e.g. Favali, Citti and Sarti, 2017; Huberle and Karnath, 2012; Lehar, 2003). One of those lessons is that perception is of *Gestalten*, or structured wholes of experience, which are different from the sum of their parts (Koffka, 1935, p. 176). This follows from the general Gestalt psychology claim that minds are primarily constituted by *Gestalten*, or structured wholes (Wagemans et al., 2012). Consider the well-known example of the black and white image of a Dalmatian dog (Figure 1). According to Gestalt theory, the act of one’s perception of the image switching from a meaningless collection of seemingly random black shapes on a white background to a meaningful image of a Dalmatian occurs because of one of the laws of *pragnanz*, specifically, the law of closure. *Pragnanz* are the presumed laws of form that govern the structure of the mind. *Closure* refers to the perceptual process whereby disparate elements group together such that fragmented pieces are perceived as unified wholes. One consequence of this Gestalt interpretation of the Dalmatian image is that perception is of wholes (i.e. global gestalts) that are irreducible to constituent parts. Moreover, such wholes result not from merely adding together individual elements, for the elements taken together have perceptual meaning different than the sum of its parts (Koffka, 1935). Understanding perception in terms of global gestalts can illuminate how IIT conceives of conscious states as integrated information.

Consider the image of the Dalmatian as comprised of distinct pixels (Figure 1). Each pixel is *information* when in an active state, that is, when it is black instead of white. Now think of the first time you saw that image as it is being presented to you in a non-integrated way. When not integrated, the image appears to be a meaningless collection of black dots. At some point, those black dots become integrated; that is, they take shape via perceptual closure. When that occurs, the information (black dots) becomes integrated (closure) such that you now perceive a Dalmatian where you once only perceived meaningless scattered black shapes on a white background. Consistent with Gestalt psychology, the shift from meaningless disparate black dots to a meaningful unified image of a dog is a shift from non-integrated information to integrated information. Moreover, the shift is from isolated bits that have no cause–effect relationship to a global structure where the bits of information take on a different meaning than they would both in isolation and if the whole resulted from merely summing the parts. If the whole is just a sum of its parts, then the image should have gradually ‘come into focus’ as a dog via a process
of sensory primitives being added together. Yet, the image is meaningful as a whole that is different than the mere sum of its parts. IIT can be understood as treating all conscious states as being similarly structured.

Figure 1. A Dalmatian dog in a field. This is an example of the Gestalt perceptual law of closure, or the visual effect when elements of a form group together such that fragmented pieces can be perceived as unified wholes. In this case, fragmented black shapes of various sizes become unified to reveal an image of a Dalmatian dog. (Image based on James, 1965.)

Take the example of conscious face perception (Figure 2). While there is evidence that the human brain, for example, detects visual features via hierarchically organized neural networks (Figure 2A), the conscious perception of a face is irreducible to those features in isolation (Figure 2B). A face can be understood to have features such as eyes, nose, and a mouth. The physical substrate of a system can be organized to detect those features in isolation. But when those features combine to contribute to a conscious perception, they participate in the global state that is a unified face. Put another way, the information that constitutes a face can be detected separately, but the conscious perception of a face only occurs via the integration of that information into a meaningful unified whole. Moreover, conscious states are not the mere collection of primitives associated via additive interactions. A conscious state — such as face perception or the perception of the Dalmatian image — is a global whole of phenomenal information.
integrated via mutual cause–effect relationships. The parts exhibit the cause–effect relationships they do in so far as they are integrated in particular wholes that establish a context whereby the interactions of those parts override what those parts do in isolation. The interactions of the parts of the physical substrate of a conscious state can be said to override what the parts do in isolation or when integrated in different ways.

*Figure 2.* Conscious face perception. (A) The conscious perception of a face (Φ) via higher-order feature detectors, hierarchically organized in a neural network. (B) Various detected features, such as edges, mouth, etc., that contribute to the global state that is the conscious perception of a face (Φ). Although global conscious states (Φ) can be constituted by collections of detected features (A), the whole conscious experience is irreducible to those features in isolation (B). (Adapted from Balduzzi and Tononi, 2009, CC-BY.)

Understanding the physical substrate of conscious states as being constituted by parts with cause–effect relationships that obtain via interactions is not what puts IIT within the realm of complexity science. IIT is properly placed within complexity science because both provide theoretical and methodological means for conceptualizing and evaluating systems defined by, for example, emergence, global-scale activities and structures, non-linearity, and self-organization. Complexity science can be regarded as a set of concepts, methods, and theories for discovering principles of activity and organization that explain global-scale activities that arise via the interactions of numerous parts. As discussed above, it appears that much of the cause–effect work of information integration seems to arise due to the interactions of parts within particular wholes. That is a major motivation for placing IIT within the broader scientific framework of complexity science, which I introduce in the next section.
3. An Introduction to Complexity Science

Although there is no single ‘complexity science’, there are typical concepts, methods, and theories that fall under that heading (e.g. Érdi, 2008; Favela, 2015; Fuchs, 2013; Hooker, 2011; Mainzer, 2007; Mitchell, 2009). In general, complexity science investigates phenomena comprised of numerous non-linearly interacting parts, where global-scale behaviour and structures are irreducible to activities and properties of its constituent parts at local scales. Such complex systems tend to display many degrees of freedom, emergence, non-linearity, scale-free structures, self-organization, strong feedback with circular or mutual causality, and unpredictability. In this brief introduction, I focus on one epistemic and one metaphysical feature that I think play the most significant roles in later demonstrating that IIT’s proper home is complexity science. The epistemic feature concerns the general methodological approach of complexity science research, namely, the search for and application of principles. The metaphysical feature concerns a key characteristic of many complex systems, namely, interaction dominance.

3.1. The epistemology of complexity science

In order to obtain knowledge about phenomena of interest, the contemporary life sciences have been dominated by broadly mechanistic and reductionistic approaches. These approaches tend to share methodological commitments to decomposition and localization. In short, in order to understand a phenomenon, an investigator breaks apart an entity or event into its constituent parts, identifies what the parts contribute to the phenomenon, and then attempts to explain the entity or event in a specific order. The ‘breaking apart’ is decomposition, which, as Bechtel and Richardson put it, allows for ‘the activity of a whole system [to be treated] as the product of a set of subordinate functions’ (1993/2010, p. 23). The ‘identify contributions’ is localization, which is the process whereby ‘the different activities proposed in a task decomposition [are identified] with the behavior or capacities of specific components’ (ibid., p. 26). The ‘specific order’ is treated in additive and linear terms (ibid., p. 23), that is, the sequence of operations that allow for an entity to obtain or event to occur. This currently popular mechanistic approach shares a scientific lineage with methodological forms of reductionism. Methodological reductionism is, in general, the idea that natural phenomena are best investigated by an approach that aims to get to the lowest possible
causally-relevant scale (Brigandt and Love, 2017). The discovery of the structure of DNA by James Watson and Francis Crick is a classic example of methodological reductionism applied to the phenomenon of life (Schaffner, 1976). Specifically, in order to explain life, Watson and Crick aimed at getting to the lowest possible, causally-relevant scale, which, they claimed, was the molecular structure of DNA. There is no doubt that the mechanistic and reductionist approaches have facilitated immensely successful scientific research.

Two types of phenomena are straightforwardly amenable to mechanistic and reductionist methodologies: first, those phenomena that are linear, that is, phenomena that produce regular changes from start or set-up to finish or termination conditions (Craver and Darden, 2013); second, those phenomena whose constitution are component dominant. Component dominance refers to phenomena whose spatial or temporal structures primarily result from additive organization. A toilet, for example, is component dominant: its structure is the result of parts simply added together (float ball, handle, siphon, etc.). Its activity occurs via effects localizable to the capacities of its parts that can be summed to reveal the whole behaviour. Moreover, those parts are not altered via their participation as components in the toilet system. Although phenomena such as toilets are amenable to mechanistic and reductionist methodologies, many other phenomena reveal the limitations of those approaches.

Systems comprised of and exhibiting many degrees of freedom, emergence, non-linearity, scale-free structures, self-organization, strong feedback with circular or mutual causality, and unpredictability resist being characterized as ‘component dominant’ and, therefore, are improperly investigated via mechanistic and reductionist methodologies (though, for a more nuanced perspective, see Baxendale, 2018). Complexity science investigates such phenomena via an approach that searches for and applies principles instead of mechanisms (Favela, 2014; Gurova, 2013; Stepp, Chemero and Turvey, 2011). The general idea is that complexity science posits a common set of principles (or laws, rules, etc.) that explain how simple causes give rise to complex effects (cf. Phelan, 2001). Moreover, these principles tend to occur at a wide range of spatial and temporal scales and are substrate neutral, which means they can occur in systems of varying material constitution. It is important to note that, although such principles are instantiated in various kinds of systems, those systems are all physical. The ‘substrate neutrality’ of complexity science does not include non-physical phenomena. An example of such a principle is self-organized
criticality (SOC). SOC is a principle postulated to explain the apparent ubiquity of power laws in nature (Bak, Tang and Wiesenfeld, 1987; Jensen, 1998; Pruessner, 2012). SOC is also postulated as a unifying theory for the many phenomena in nature and experimental settings that demonstrate spatial organization such as scale invariance and self-similar structure and temporal dynamics characterized by \(1/f\) signals. Power laws, scale invariance, and \(1/f\) signals are indicators of systems poised at critical states, that is, between order and disorder. The wide range of phenomena that exhibit SOC include rice and sand pile avalanches (Bak, 1996), earthquakes (Bak et al., 2002), networks of cortical neurons (Beggs and Plenz, 2003), neural network connections (Sporns, 2011), spontaneous neuron activity (Favela et al., 2016), and mental image rotation (Gilden, 2001), to name just a few.

By appealing to principles as sources of hypotheses for experimental investigation and as explanations of results, complexity science appears, epistemologically, much more like physics than biology. The reason is that many explanations in physics deal with systems that are spatially, temporally, and/or numerically too large to decompose and localize mechanisms or reduce to simple linear and additive processes. One modelling approach utilized to get an epistemic handle on such complex systems is to treat phenomena of interest as order parameters (Haken, 1988/2006). An order parameter is the global state of a system that is the target of investigation, for example, water transitioning from liquid to solid. These states are guided by control parameters, for example, in the case of water, pressure and temperature. Because complex systems involve numerous parts interacting in non-linear ways, order parameters are understood as occurring at a global scale such that the details of its constituent parts — such as specific location, material composition, or quantity — or control parameters are often abstracted when they are unnecessary to explain and understand the phenomenon of interest. The search for and application of principles is readily combined with this modelling practice. For example, a target of investigation may be a system exhibiting phase transitions across a wide range of starting points and parameter values. Here, SOC can be posited as the order parameter and the control parameter(s) are those values that drive phase transitions, for example energy dissipation and system memory (Cocchi et al., 2017). Such an approach is appropriate for highly complex, multiscale phenomena that resist finding specific component causes. Though mechanistic and reductionist approaches have
facilitated many successes in the life sciences (e.g. discovery of DNA), more and more biological phenomena are revealing the limits of such methods. Investigations of the human brain provide such examples. The spatial and temporal complexity of neural structure and function is driving more investigators towards frameworks that are not generally mechanistic or reductionistic in nature, such as coordination dynamics (Tognoli and Kelso, 2014), dynamical systems theory (Izhikevich, 2007), and synergetics (Haken, 2015); not to mention complexity science (Van Orden and Stephen, 2011).

With its focus on *principles* of activity and organization, the epistemology of complexity science differs greatly from more typical practices in the life sciences that are guided by mechanistic and reductionist methodologies. Be that as it may, the focus on principles is quite common in physics and is becoming increasingly utilized in the life sciences; in particular, in the cognitive, neural, and psychological sciences. By focusing on global-state activity and organization, research guided by the search for and application of principles has the benefit of making comprehensible staggeringly complicated phenomena such as the brain and cognition (Favela, 2014). With that said, it would be mistaken to utilize such an approach for all phenomena. It is proper to investigate component dominant and linear systems via mechanistic and reductionist means. In actuality, it could be counterproductive to investigate such systems via complexity science, or coordination dynamics, dynamical systems theory, and synergetics. Many of the tools used within the aforementioned frameworks are justifiably applied to phenomena both within and outside their purview. Differential equations, for example, are clearly appropriate for use by dynamicists and mechanists. However, the metaphysical commitments of such frameworks are not readily as equally applicable. If a system is treated as non-decomposable, then it is more likely to pose a challenge for mechanist and reductionist frameworks that require — at least methodologically — a system to be decomposable. One such metaphysical commitment of complexity science is to the investigation of phenomena that are interaction dominant.

### 3.2. The metaphysics of complexity science

Interaction dominance is a common metaphysical property of complex systems, which is a kind of dynamic and organizational property. A system is *interaction dominant* when the activities and/or organization of its parts supersede those that the parts would have separately from
each other (Favela and Martin, 2017; Holden, Van Orden and Turvey, 2009; Silberstein and Chemero, 2012). Moreover, it is the continuous interaction of the system’s parts that sustain the global-scale activity. Such global-scale activity and organization has a mutual effect on the parts, as there is a global-to-local as well as local-to-global cause–effect relationship that maintains a particular order. As a consequence, interaction-dominant systems are organizationally and dynamically non-decomposable and non-localizable — in short, they are not component dominant. Systems that are component dominant — namely, decomposable and localizable — exhibit activities that are additive. They are additive in that the nature of the individual constituent parts are primarily responsible for the cause–effect power of the system. On the other hand, the cause–effect powers of interaction-dominant systems are primarily due to the continuous interaction of the parts as contributors to a particular whole.

Locust swarms provide an illustrative example of interaction dominance. Interactions among relatively simple parts (i.e. individual locusts) give rise to complex actions (e.g. swarm migration). This occurs by way of the global-scale swarm dynamics resulting from and overriding the local-scale individual locust dynamics. Locust swarms result from strong feedback with circular or mutual causality from swarm to locust and locust to swarm. Of course, there would be no swarm without the contributions of the individual locusts. However, it is the interactions among the locusts that is more significant to the activities and organization of the swarm than the properties of individual locusts. As a result, the global-scale system that is the swarm is resilient enough for local-scale individual locusts to align or misalign with the swarm to wide and varying degrees without the swarm collapsing. Additionally, individual locusts that fall too far out of alignment with the swarm can be corrected by reintegrating into the swarm just enough to allow it to be affected by and contribute to the global-scale behaviour. Consequently, the swarm can accomplish goal-directed behaviours such as migration via a wide range of structural alterations as long as the interaction dynamics among the parts are maintained. Thus, contrary to toilets that are the sum of their parts, via interaction dominance locust swarms are different than the sum of their parts.

As I have emphasized, complexity science-based investigative frameworks highlight the idea that some natural activities are the result of global-scale properties overriding local-scale properties. It is in that way that complex systems are often considered to be more than
and/or different from the sum of their parts. Two concepts that make this point clear are component dominance and interaction dominance. Systems like toilets are prototypically component dominant: their output activities result from weak interactions (i.e. additive, linear, and local). Systems like locust swarms are prototypically interaction dominant: their output activities result from strong interactions (i.e. multiplicative, non-linear, and global). I am now positioned to make the case that IIT is properly understood within the broader theoretical framework of complexity science.

4. IIT as Complexity Science

IIT is viewed by some as one of the most promising scientific theories of consciousness currently available (e.g. Koch, 2017). I claim that IIT can provide an even more compelling theory of consciousness if it is situated within the broader framework of complexity science. Here I focus on two aspects of IIT that are not only congruent with complexity science but are also strengthened if situated within such a framework. The first set is complexity science’s methodology guided by the search for and application of principles and IIT’s focus on cause–effect power over the particular material constitution of consciousness. The second set is IIT’s claim that conscious states are irreducible integrated wholes and the complexity science concept of interaction dominance.

4.1. Principles, axioms, and postulates

One feature of complexity science’s epistemology is that it focuses on the search for and application of principles of activity and organization. Self-organized criticality (SOC) was presented above as an example of such a principle that has guided experiments and informed explanations. One reason principles such as SOC are so valuable to scientific practice is that they are instantiated in a wide range of system types. In other words, what matters most to an explanation is giving an account of a phenomenon’s activities and/or organization and not necessarily what that phenomenon is made of. Put in the vocabulary of complexity science: discovering the control parameters that guide order parameters does not require necessarily revealing the material constitution of either the control or order parameter. Such an approach has more in common with explanations in physics, which often appeal to principles (e.g. universality classes; Kadanoff, 1990), than with the life sciences, which often appeal to mechanisms (i.e.
specific material constitution). This emphasis on widely-applicable principles is the first reason I provide for situating IIT within complexity science.

By focusing on axioms and postulates, IIT is a radical departure from methods central to most sciences of consciousness. Christof Koch and Francis Crick were early contributors to the legitimization of the scientific study of consciousness (Crick and Koch, 1990). The method they advocated for was the search for neural correlates of consciousness (NCCs), which has since become the central approach within the sciences of consciousness. The NCCs are ‘the minimal neuronal mechanisms jointly sufficient for any one specific conscious percept’ (Mormann and Koch, 2007). Accordingly, and put roughly, consciousness research often centres on finding mappings between conscious states and brain states. In this way, the sciences of consciousness have typically been mechanistic and reductionist, with methodological emphases on decomposition and localization of functions (Uttal, 2011). IIT, on the other hand, does not centre on the search for neural mechanisms that correlate with conscious states. Instead, IIT begins with phenomenology (axioms) and then attempts to infer what processes (postulates) are necessary to give rise to those states. IIT does not start with brain mechanisms and ask how they give rise to consciousness (Tononi et al., 2016), nor does it build a theory based on brute correlations among brain activity and conscious states. In this way, IIT breaks from more traditional mechanist-reductionist approaches in neuroscience by not focusing on the search for the neurobiological substrates of consciousness. IIT is not a bottom-up approach from neurobiology to consciousness but is top-down from consciousness to its physical substrates.

It is reasonable to think that such a top-down approach implies that IIT is committed to consciousness being completely substrate-neutral; for example, thinking that IIT asks how a conscious state (axioms) could come to be regardless of material constitution (postulates). It is true that IIT widens the range of possible systems that can be conscious (e.g. machines; Koch and Tononi, 2008). However, IIT requires that conscious systems be physical, where ‘physical’ means the system has particular cause–effect power. Furthermore, the physical substrate of any conscious system is not reducible to its mechanisms but is defined by its maximally irreducible cause–effect structure. In this way, IIT does not propose that consciousness is substrate-neutral or best understood via mechanisms. Instead, it proposes that the best way to understand consciousness is via a principle of maximally
irreducible intrinsic cause–effect power. In this way, IIT is more like complexity science than mainstream neuroscience, for it prioritizes explanations and investigations via principles of activity and organization that attempt to shed light on the nature of the system and not mere brute correlations or specific material constitution (e.g. neurons).

IIT’s emphasis on principles of activity and organization can benefit from a complexity science modelling practice centred on order and control parameters. Here I present three ways this can be applied, in order from most concretely applicable to more speculative.

The first, and most concrete, way is to treat a system’s macro spatio-temporal scales as an order parameter and its coarse-grained micro elements as control parameters. For example, the macro scale captured by an order parameter could be a brain region and the micro elements could be specific neural networks. Though they do not specifically utilize the ‘order and control parameters’ concepts, such a method is consistent with work by Hoel, Albantakis and Tononi’s (2013) that attempts to quantify causal emergence as the result of a system’s macro causation superseding its micro causation.

A second way to apply this approach would be to treat a system’s maximally irreducible cause–effect powers as the order parameter and the physical constituents of its substrate as control parameters. Such an approach is in principle possible but would be computationally demanding as causally-relevant control parameters would need to be identified for each order parameter that captures maximally irreducible conscious states. However, the ability to compute $\Phi$ for systems as complex as the human brain is a common criticism of IIT, and is one I sketch a response to in Section 5.

The third way is the most speculative and would involve treating axioms as order parameters and postulates as control parameters. In this way, axioms are like order parameters in being the global-scale phenomena of interest (i.e. a system’s conscious state) and the investigator’s task is to reveal how the postulates (i.e. control parameters) guide that activity and organization. The challenge with this more speculative application of order and control parameters to axioms and postulates would be to make clear, both conceptually and empirically, the relationship between the two. Remember, control parameters are not the cause of order parameters, for the relationship is one of mutual causality. Accordingly, for such an approach to properly apply to investigations of axioms and postulates, their relationship would also have to be one of mutual causation. Specifically, it would need to be understood how the phenomena that postulates refer to cause and are
caused by axioms and how the phenomena that axioms refer to cause and are caused by postulates.

The points I have raised support the claim that IIT can benefit from some of the key epistemological features of complexity science. By utilizing an approach guided by order parameters and control parameters, IIT as a complexity science approach to consciousness is tasked with discovering principles of activity and organization of conscious states and the physical systems that define them. Such an approach is a radical departure from typical sciences of consciousness that focus on the search for NCCs via mechanistic and/or reductionist commitments. Next, I show how an aspect of the metaphysics of IIT benefits from complexity science.

4.2. Irreducible consciousness and interaction dominance

Interaction dominance is a property of many systems investigated by complexity science. Interaction-dominant dynamics, for example, is irreducible and resists methods such as decomposition and localization. This irreducibility is the result of the continuous interaction of parts to sustain global-scale activity and organization. Moreover, when contributing to an interaction-dominant system, parts exhibit cause–effect powers due to their interactions they would not have in isolation or if organized in different ways. Building on Favela (2017), I claim that interaction dominance is a proper concept to explain some of the key metaphysical claims of IIT. Central to IIT is the claim that conscious states are irreducible (Oizumi, Albantakis and Tononi, 2014; Tononi, 2012a; Tononi and Koch, 2015). The irreducibility is captured by the integration axiom and postulate. As a reminder, in terms of phenomenology, the integration axiom refers to the claim made by IIT that each conscious state is a unified, global whole. That global state is irreducible to the seemingly distinct qualities of that experience. Take the conscious perception of a face (Figure 2B): notwithstanding the fact that we can identify different features (e.g. eyes, lips, mouth, etc.), a face is experienced and is meaningful as a whole. In terms of the physical substrate, the integration postulate infers that even if a system is comprised of parts, those parts contribute cause–effect powers as a unified whole that is irreducible to the mere additive combination of those parts or if the parts were organized differently. Returning to the conscious perception of a face (Figure 2A): notwithstanding that a system could be hierarchically organized to detect features via, for example, a neural network, the
parts of that system have cause–effect powers by way of their particular interactions and organization, and not via their individual properties. This is where interaction dominance provides an appropriate concept to explain the irreducibility of integration.

Consider again the example of face perception (Figure 2). According to IIT, when a digital camera takes a picture of a face there is a low Φ value of one bit of information. The reason the digital camera does not sustain a high Φ value is because each of its lens’s one million parts have no cause–effect relationship with each other. Digital cameras are component dominant: the images they produce are the product of merely adding together the information from each individual part of the lens. For that reason, if one part of the lens is activated or not, then no other part of the image is affected. If the human brain worked like a digital camera, then face perception would be the result of adding together features, that is, a face would be the sum of eyes, mouth, nose, etc. The hierarchical and feature detection-based function of the human brain’s neural networks seem to lend support to such an account of face perception (Figure 2). However, the phenomenology of face perception is not like that. Our conscious states are the result of contributions from distinct processes (e.g. feature detection). But the end result of those processes is a unified whole that is irreducible to the individual contributions. Consider again Gestalt psychology and the Dalmatian image.

According to Gestalt psychology, perception is of Gestalten, or structured wholes of experience (global gestalts). These global gestalts result from laws of perception such as closure, or the tendency of our perceptual systems to group features together. The notion that we perceive unified wholes is supported by our phenomenological experience when viewing the Dalmatian image (Figure 1). When we first see the image, it appears as a random spattering of black shapes on a white background. Then, seemingly at random, the image becomes meaningful as we perceive a Dalmatian walking away and sniffing the ground. As argued above, if perception is the result of mere additive collections of primitive percepts, then the phenomenological experience of viewing the Dalmatian image would be quite different: instead of a sudden phase transition from meaningless random black shapes to meaningful Dalmatian, we would gradually see the Dalmatian come into being as features were detected and added together ‘before our eyes’. If the Gestalt conceptualization of this perceptual experience is correct — I think it is — then it fruitfully serves as an example that supports IIT’s claim that conscious states are irreducible due to
information integration. Nevertheless, that does not explain how the states are irreducible. That is where interaction dominance can help.

If a system is interaction dominant, then its global-state activity and/or organization is due to the interactions of its constituents and not the properties of the constituents merely added together or organized in different ways. The Dalmatian suddenly appears to us because the information (black shapes) integrates via interactions that give rise to a meaningful global state that is not merely the result of summing the information together. The parts have cause–effect powers (i.e. seeing a Dalmatian) by means of integration. Integration is able to enable cause–effect powers due to the interactions of the parts overriding what the parts do in isolation. Interaction dominance suitably explains how both the phenomenology and physical substrate of conscious states are irreducible. The phenomenology is irreducible because conscious states are meaningful as global states, or global gestalts. The physical substrate is irreducible in that their cause–effect powers result from and are constituted by their interactions. Along these lines, the irreducibility of conscious states as proposed by IIT is explained via interaction dominance, which places IIT within the scope of complexity science’s metaphysics.

5. Avenues for Replies to Critiques

If IIT is indeed a complexity science approach to consciousness, then I think it can be better positioned to respond to some of its common criticisms. Here I highlight two particular criticisms that could be addressed more fully if IIT is situated within a complexity science framework. The first criticism is epistemological and centres on the ability to compute Φ. Koch makes the point clear when he states that,

> Computing Φ is rather demanding because all possible ways the system can be divided have to be considered… For the 302 neurons that make up the nervous system of *C. elegans*, the number of ways that this network can be cut into parts is the hyperastronomical 10 followed by 467 zeroes. (Koch, 2012, p. 128)

Even if calculating Φ is in principle possible, given that the human brain is many orders of magnitude more complex than a worm’s brain, such a task seems impossible in practice. With that said, a number of computational techniques have been developed to supplement IIT (e.g. Oizumi *et al.*, 2016; Tegmark, 2016). Koch further adds that ‘heuristics, shortcuts, and approximations’ will be required because ‘[c]omputing Φ for any nervous system… is fiendishly difficult’
One potentially fruitful path to a solution to the ‘fiendishly difficult’ challenge of calculating $\Phi$ may be found in a combination of complexity science’s method of modelling via order and control parameters along with the search for and application of principles of activity and organization. Mathematical techniques that enable dimension reduction could bring into focus tangible order and control parameters of conscious systems. From there, commonalities among the ways in which control parameters guide order parameters could lead to the discovery of new principles of activity (for similar strategy, see Favela, 2014, and Cocchi et al., 2017). Those principles could serve as the heuristics, shortcuts, and approximations Koch claims are needed to make calculating $\Phi$ pragmatically possible.

One approach that can enable dimension reduction and serve as a ‘shortcut’ when calculating $\Phi$ is the application of the renormalization group (RG; Wilson, 1983). The RG is a method from statistical physics that is often applied to investigations of complex phenomena with numerous, multiscale interactions. It enables wide ranges of systems that exhibit discrete features to be grouped together (Boettcher and Brunson, 2011). Those groups can then be categorized into relatively few classes that exhibit shared principles of activity and organization that are determined by few characteristics and occur across multiple spatial and temporal scales (Batterman, 2000; Thouless, 1989). One such principle was mentioned above: self-organized criticality (SOC). SOC and other critical states (e.g. sub-criticality and supercriticality) have been utilized to characterize the dynamics and information flow of complex neural systems (e.g. Beggs and Plenz, 2003). When computing massively connected multiscale phenomena, it helps to make the calculations more tractable by fixing key aspects of the dynamics (e.g. as being subcritical or supercritical) and then assessing the connection patterns compatible with those kinds of dynamics (e.g. small-world networks; Sporns and Tononi, 2007). Instead of calculating every aspect of a system, computational resources can be saved by building assumptions about the system into the models. Those assumptions serve as auxiliary hypotheses that have already been empirically justified by previous research. Assumptions such as critical dynamics and small-world network connectivity can serve as shortcuts when computing $\Phi$, which would make calculating $\Phi$ pragmatically possible and not just possible in principle.

The second criticism is metaphysical and centres on the nesting of multiple consciousnesses within single systems. Eric Schwitzgebel (2012; 2015) has criticized IIT for not making a strong case as to why
a single system has one consciousness. If consciousness equals information integration, then, according to Schwitzgebel, anywhere there is integrated information within a single system there will be consciousness. Since there are multiple locations where information is integrating in single systems such as humans, then it seems that IIT is committed to humans having multiple consciousnesses at any one time. To combat that worry, Tononi argues that single systems have a single consciousness that is its maximum $\Phi$ ($\Phi^\text{max}$; Hoel et al., 2016). Other than considerations of parsimony, Tononi’s response is not particularly satisfying. Interaction dominance may provide a way to respond to this criticism.

If a system is component dominant, then its system-level properties result from additive combinations of its parts. Moreover, the capacities or organization of those parts are not altered when contributing to a component-dominant system; they merely contribute to a system as they are. If a system is interaction dominant, then its system-level properties result from non-additive combinations. Moreover, the capacities and/or organization of those parts are altered when contributing to an interaction-dominant system; they are different or more as a consequence of participating in that particular system and in particular ways. In other words, when contributing to a system, the properties of components of an interaction-dominant system are overridden and altered by the nature of their interactions within a larger system. Now consider IIT’s claim that each system will have a single consciousness, that is, $\Phi^\text{max}$. If each system has one $\Phi^\text{max}$, then it does not have nested subsystems that attain consciousness simultaneously when the system is in a $\Phi^\text{max}$ state. Nested subsystems could be conscious within larger systems if the system is component dominant. Components of a system that can have phenomenal properties in isolation are not altered by being part of a larger system. Thus, they could remain conscious even if the system as a whole has its own conscious state. However, if phenomenal states are interaction dominant (Favela, 2017), then it follows that the contributions of the parts of those states are overridden when they contribute to the interactions that give rise to that single, global, and unified consciousness. For example, it is likely that the features that contribute to conscious face perception do not have individual consciousnesses. There is no consciousness-of-eyes, consciousness-of-mouth, etc. nested within the consciousness-of-face (Figure 2B). Perhaps if consciousness resulted from component dominance, then nested consciousnesses could exist within each other, for example distinct consciousness-of-eyes, consciousness-of-
mouth, etc. But for the same reasons why a digital camera has only one bit of Φ and the image of the Dalmatian occurs holistically, by resulting from interaction dominance, a system can only have one global state at a time. Admittedly, these responses need further refinement. Nevertheless, I present them with the aim of motivating the claim that understanding IIT as a complexity science approach to consciousness opens up new possibilities for responding to some of its major criticisms.

6. Conclusion

The integrated information theory of consciousness (IIT) is currently one of the most compelling scientific theories of consciousness. It presents a provocative metaphysics that equates consciousness with integrated information. Moreover, it provides an epistemically innovative approach to quantifying consciousness (Φ). I have argued that IIT is a complexity science approach to consciousness. Complexity science investigates phenomena with irreducible activities and organization that result from numerous non-linearly interacting parts. Complexity science utilizes concepts, methods, and theories traditionally applied to phenomena investigated in physics to those in the life sciences. I drew attention to key features of IIT (i.e. cause–effect power and information integration) in order to demonstrate that they are congruent with and can benefit from epistemic and metaphysical features of complexity science (i.e. the search for and application of principles and interaction dominance). I then suggested some ways common criticisms can be responded to if IIT is cast as complexity science. Even cast as complexity science, IIT may not be able to overcome epistemic issues concerning the ability to compute Φ or metaphysical issues concerning the relationship of consciousness to information. With that said, even if it only serves as inspiration or a base for other theories, IIT is a step forwards in our scientific understanding of consciousness. Yet, as I have argued, it could do even more if treated as a complexity science approach to consciousness.

Acknowledgments

I would like to thank Garrett Mindt for inviting me to submit to this special issue and for his support of the project. I would also like to thank the two anonymous reviewers who provided many helpful suggestions. The paper is much improved thanks to their feedback.
References


dynamics, in Jirsa, V.K. & McIntosh, A. (eds.) *Handbook of Brain Connectivity*,
pp. 117–147, Berlin: Springer.

Stepp, N., Chemero, A. & Turvey, M.T. (2011) Philosophy for the rest of cog-

Tegmark, M. (2016) Improved measures of integrated information, *PLoS Compu-
tational Biology*, 12 (11), e1005123, pp. 1–34.

Thouless, D. (1989) Condensed matter physics in less than three dimensions, in
versity Press.


Neuroscience*, 5 (42), pp. 1–22.

Tononi, G. (2008) Consciousness as integrated information: A provisional mani-


Tononi, G. (2012b) *Phi: A Voyage from the Brain to the Soul*, New York:
Panthec Books.


474–484.


Tononi, G. & Koch, C. (2015) Consciousness: Here, there and everywhere?, *Phil-
osophical Transactions of the Royal Society B: Biological Sciences*, 370, art. 167,
pp. 1–18.

theory: From consciousness to physical substrate, *Nature Reviews Neuroscience*,
17, pp. 450–461.


Van Orden, G. & Stephen, D.G. (2012) Is cognitive science usefully cast as com-

Wagemans, J., Elder, J.H., Kubovy, M., Palmer, S.E., Peterson, M.A., Singh, M.,
& von der Heyd, R. (2012) A century of Gestalt psychology in visual per-
ception I. Perceptual grouping and figure-ground organization, *Psychological