

Implementation, Formalization and Representation: Challenges for the Integrated Information Theory

Authors:

Carlos Montemayor*
Department of Philosophy, San Francisco State University, San Francisco, CA 94132
cmontema@sfsu.edu

J. Acacio de Barros
School of Humanities and Liberal Studies, San Francisco State University, San Francisco, CA 94132
barros@sfsu.edu

Leonardo Guimarães De Assis
Suppes Brain Lab, Center for the Study of Language and Information, Stanford University, Stanford, CA 94305
lpgassis@stanford.edu

*Corresponding Author

Implementation, Formalization and Representation: Challenges for the Integrated Information Theory

Abstract:

Any theory of information needs to comply with what we call the *implementation, formalization, and representation* constraints. These constraints are justified by basic considerations concerning scientific theories and the philosophy of science. We argue that the implementation and formalization constraints are better satisfied by Shannon information, and that IIT requires more specificity to determine why it is superior to Shannon's account. The second part of the paper focuses on the representation constraint. We argue that IIT cannot succeed in satisfying this constraint for semantic contents without offering a way of formalizing and representing concepts, conceptual roles, and context sensitivity, among other problems. Further complications are raised with respect to the distinction between consciousness and attention, which apply more specifically to the representation constraint. We conclude with some recommendations as to how IIT may succeed in solving these problems, highlighting the advantages and enormous potential of IIT as a scientific theory.

Implementation, Formalization and Representation: Challenges for the Integrated Information Theory

Introduction

The Integrated Information Theory (IIT) has two key advantages when compared to other contemporary theories of consciousness. The first is that it explicitly confronts the hard problem of consciousness. Many other views of consciousness, for instance the Global Workspace Theory (GWT), identify consciousness through reductive methods, such as verbal report or broad neural activation. The canonical data used to confirm these theories, because of this reductive assumption, leave open the possibility that they measure or verify at best what philosophers call “access consciousness” rather than “phenomenal consciousness,” which is the type of consciousness associated with the hard problem (Block, 1995; Chalmers, 1995, 1996).

How can IIT avoid the pitfall (and potential contradiction) of measuring consciousness with reductive methods and at the same time provide some measure of consciousness that acknowledges the non-reductive nature of phenomenal consciousness? IIT postulates axiomatic truths about the existence of qualitative states that characterize and integrate phenomenal consciousness. From these axioms about experiences, it deduces postulates concerning the type of physical substrate that realizes or instantiates consciousness (Oizumi et al., 2014; Tononi et al., 2016). This is a “Euclidean” or a priori way of proceeding because, like Euclid’s axiomatic geometry, it proceeds from fundamental truths (axioms and their entailments) to empirical measurement, rather than from empirical measurement to truth-confirmation. This a priori methodology, as explained below, presents unique challenges to IIT as a scientific theory of consciousness. At the same time, it also makes it a unique and interesting approach to an extremely puzzling scientific problem that resists any type of reductive approach.

A second advantage of IIT is that it provides a measure of information in terms of the phenomenal *quality* and semantic content of a conscious state, as well as the *quantity* of information in that conscious state. It is a “measure” in the sense that the physical substrate of consciousness (PSC) is constrained by both phenomenal character and semantic content, in a way that is mathematically understood and, at least in principle, measurable through experiment and inductive methods. Experiences are isomorphic to a cause-effect structure that constrains past and future states in a unique, maximally informative manner. Experiences are also isomorphic, in deeper way, to conceptual structures. In fact, experiences are claimed to be *identical* to conceptual structures (Tononi et al., 2016, 451), which means that they are not just identical in structure but also in every other respect, including their content. This establishes an identity between composite phenomenal states and a conceptual structure.

These isomorphism and identity axioms explain the uniqueness and maximally determinate information of phenomenal consciousness. $\Phi_{\max}$ represents the degree or measure of interrelation, causal efficacy and conceptual interconnection of a phenomenally conscious state. In particular, it measures the degree to which the total state is irreducible to its component parts. This central idea is examined below in detail. What needs emphasis now is that these axioms of IIT are strong commitments to mappings that establish identities in structure (an isomorphism) and in content. These mappings generate the challenge of how might total information in a given state, including the probability of prior and posterior events, accommodate the contextual dependence of semantic and metric information (see Reference Omitted). They also create challenges, which we examine here, with respect to the status of IIT as a scientific theory.

The main challenge we present here is that IIT remains incomplete as a theory, at least as long as the axioms remain at the postulation level. More precisely, the challenge is that concrete *models* need to be provided for causation-mappings that are relevant for the PSC in order to understand how the theory can be implemented in physical systems, as well as how it is actually implemented in the brain. Modeling is an important addition to the theory because, for instance, if all PSC models that satisfy IIT are also models that satisfy other extant notions of information (such as Shannon information) then the advantage of IIT over other theories disappears. The axiomatic principles of phenomenology, for instance, would be irrelevant because models that satisfy them would also satisfy an account of information that dispenses with these axioms. Contextual elements that depend on how these models are implemented are critical at the causal and semantic levels. So the models should also account for how context sensitivity shapes $\Phi_{\max}$. Likewise, concrete models for semantic structure need to be mapped to qualia spaces (or some kind of parthood structure for experiences that are maximally specific) in order to determine how exactly semantic representation operates within the IIT framework. Although some models have been explored in the context of interpreting IIT's empirical consequences, we argue that none of them are completely satisfactory yet.

Section 1 explains the importance of models for the justification of theories, particularly those that are presented in axiomatic terms, such as IIT. Section 2 addresses the issue of how a theory of information might be implemented and formalized, and raises some difficulties concerning the best known theory of information, Shannon's account of information. Section 3 examines the axiom that conceptual mappings provide the semantic specificity of conscious states with phenomenal character, or qualia. The main challenge in this section is that a theory of concepts is required to interpret and model the identity between concepts and qualia. Finally, section 4 presents further problems regarding the distinction between consciousness and attention that may affect the mappings addressed in the previous sections. The conclusion emphasizes the advantages of IIT as a potential scientific explanation of phenomenal consciousness and presents basic considerations about the kind of interdisciplinary work that lies ahead.

1. Models, measurements and verification

Measurement and modeling are central themes in philosophy of science. Measurement theory is concerned with the central philosophical questions of what is a measurement, what kind of invariances it requires, what is the relationship between axioms, coordinative definitions and data? Modeling addresses some of these issues by presenting approximate ways of satisfying the predictions and postulates of a theory. Some of the mappings between mathematical structures and concrete magnitudes and physical properties will be homomorphic, in the sense that critical aspects of the physical structure are invariantly and systematically captured in the mathematical structure. Other coordinative definitions might be even stronger, or isomorphic, in the sense that the structures are supposedly identical. These are intricate issues and we are not going to explore them in detail. But since IIT postulates such strong mappings between phenomenology, conceptual content and causal structure, it is important to examine these general considerations about measurements and models to some degree, in order to evaluate IIT.

Modeling qualia in terms of conceptual and causal-structural mappings is an ambitious goal of IIT. As we shall argue below, it remains unclear how concrete measurements and models might provide enough structure or invariance to verify IIT as a theory of information. There are two aspects of this problem, which we explore in detail in the following sections. There is a physical realizer of the theory, the PSC, which is measured in terms of $\Phi_{\max}$ as a metric for cause-effect structure. This metric has a second aspect, because it measures the degree of integration between concepts, which are identified with experiences at the holistic or structural level. If the brain is only one model that satisfies these two aspects of IIT's metric, and $\Phi_{\max}$ can be used to measure many other realizers of IIT, the causal constraint seems to be in jeopardy, because there seems to be little by way of invariance between coordinate definitions and concrete physical properties. Consequently, if IIT can have an infinite number of models that satisfy it, then the theory seems to be either unverifiable or trivially true because it entails an unacceptably large number of realizations and models. This is the type of difficulty that concerns us here.

The axiomatization of a theory is only one half of the theory. The other half is its empirical interpretation of data through models. As mentioned, this is a complex issue in philosophy of science. Here, we want to briefly illustrate the kind of problematic IIT confronts as an axiomatic theory with no clear models (at least so far) that would satisfy it in terms of specific experimental paradigms and canonical interpretations of data. A central problem with respect to this issue is that the axiomatic part of the theory can never be "in direct contact" with data. Models, coordinative definitions, and further empirical investigation is required. As Suppes says:

The concrete experience that scientists label an experiment cannot itself be connected to a theory in any complete sense. That experience must be put through a conceptual grinder that in many cases is excessively coarse. What emerges are the experimental data in canonical form. These canonical data

constitute a model of the results of the experiment, and direct coordinating definitions are provided for this model rather than for a model of the theory. (Suppes, 2002, 7)

In order to produce a measurement that verifies a theory, the data need to be interpreted in a way that satisfies the theory as one of its models. Then further generalization and inductive reasoning are required. This scientific task cannot be performed axiomatically. Measuring and modeling require empirical coordinative definitions, namely, definitions that relate invariantly a set of measurements, canonically interpreted, to the predictions or consequences of a theory. In other words, the $\Phi^{\max}$ metric provides only a postulate for possible measurements of information, which needs to be matched to canonical interpretations of data. In this sense, it is not really a *coordinative definition* between IIT and data. The appropriate coordinate definitions must be provided independently. It is with respect to this issue that we raise the challenges below.

As Suppes emphasizes in the quote, a concrete experience provides the context in which a theory is verified at various levels. The formalization of the theory, in its axiomatic and mathematical form, must be contextualized in terms of physical measurement and coordinative definitions of data. IIT has a second challenge in modeling conscious experiences with conceptual structures. We explain why this mapping also requires concrete contextualization, in the sense that semantic models verify conceptual cognition. These two mappings, from phenomenal consciousness to conceptual structures, and from conceptual structures to cause-effect repertoires, make IIT a particularly challenging theory to model and measure.

Because of these reasons, any theory of information needs to comply with what we call the *implementation, formalization, and representation* constraints. The implementation constraint concerns how the theory measures information flow in concrete physical systems. The formalization constraint concerns how the theory is mathematically modeled, such that representation theorems between signals and probabilities can be used to predict outcomes. Finally, the representation constraint concerns how information flow, as measured and formalized by the theory, maps onto contents, including semantic ones. IIT postulates mappings that attempt to address these three constraints, with the ambitious goal of providing a new theory of information capable of explaining phenomenal consciousness. In the first part of the paper, we argue that the two first constraints are better satisfied by Shannon information, and that IIT requires more specificity to determine why it is superior in terms of implementation and formalization. The second part of the paper focuses on the representation constraint. We argue that IIT cannot succeed in satisfying this constraint for semantic contents without offering a way of formalizing and representing conceptual structures and the context sensitivity of meaning. We then argue that the distinction between phenomenal consciousness and attention (as well as access consciousness) makes this difficulty even more intractable.

2. Implementation and Formalization

How are measurements going to verify the conceptual and causal structures postulated by IIT? As mentioned, there is an isomorphism between causal structures at a specific scale (not the micro scale) and $\Phi^{\max}$ information, in terms of the elements that constitute the PSC. These elements, which are highly interrelated and integrated, must also have cause-effect powers within the PSC (Tononi et al., 2016). This implementation of the causal structure is scale-dependent in the sense that it is irreducible to component parts. It is an emergent structure (Hoel et al., 2013) and, therefore, the mapping for identifying $\Phi^{\max}$ is at the macro-level. This entails a strong commitment to top-down causation that some authors consider as incompatible with physicalism (Kim, 2000). If one assumes physicalism as a basic constraint on any view of the mind (the view that all causal powers must be explained by physical powers, and that, because of the causal closure of the universe and the problem of overdetermination, this entails that causation must only occur at the micro-level), then IIT would be unacceptable because of its commitment to emergentism. This constitutes an “in principle” objection to IIT.

Let us grant, however, that this emergent causal structure is an adequate and scientifically sound implementation of IIT because top-down causation is compatible with physical causation. Two criteria must be met for the PSC to truly count as a causal structure and not merely as some convenient descriptive device. First, the relation between components and the emergent causal structure cannot be simply one of multiple realization, in which macro-level property types are multiply realized by micro-level physical tokens (Fodor, 1974) because this possibility is compatible with causation occurring only at the micro-level (Kim, 2000). Since IIT is incompatible with micro-level (reductive) causation, then the relation between the components and the emergent structure cannot be simply one of multiple realization. In fact, because of the problem of overdetermination (Kim, 1992), any non-reductive claim about macro properties should be understood only in terms of the metaphysical importance of the property, and not in terms of causal powers, which are fully determined and completely described at the micro-level. In other words, according to this version of physicalism, the emergent PSC has no causal powers “of its own” even if it might have emergent properties. This means that the PSC must itself be a *type* (rather than a token realizer) that only special macro structures can instantiate. Second, and because of this reason, the relation between the macro PSC and its micro-components cannot be merely supervenience, precisely because the relation of supervenience is compatible with a reductive account of causation (Kim, 2000).

Suppose these two criteria are met by IIT. The emergent causal structure of the PSC allows for degrees of $\Phi^{\max}$ that causally modify the micro level, without overdetermination (i.e., these are causally powerful structures that do not fully depend on component parts, such that the PSC has its own, and irreducible, causal powers). For IIT to be a complete theory, a model of such structures is needed. Marshall et al., (2018) offer “black boxes” as models (input-output devices in which the micro elements are hidden; see also Hoel et al., 2013). There is no space here to address this interesting proposal in detail, but it is sufficient to say that this model of the PSC is fully compatible with an alternative theory of information, Shannon’s account of channel

capacity (Hoel, 2017). So even if one accepts that black-boxes are adequate models of the cause-effect structure of a PSC, it is possible that Shannon's theory, rather than IIT, is the correct description.

This brings us to a related issue concerning the completeness of IIT as a scientific theory: its mathematical formalization. The implementation constraint concerns how the theory measures information flow in concrete physical systems, such as the emergent casual structure of the PSC. The formalization constraint concerns how the theory is mathematically modeled, such that representation theorems between signals and probabilities can be used to predict outcomes. Here a basic simplicity criterion speaks against adopting the a priori postulates that ground the mathematical formalization of IIT: Shannon's theory has been extremely well studied, it is very well understood and has been remarkably successful (Adriaans and van Benthem, 2008). If the preferred models for the PSC cause-effect structure (black boxes) are fully described by Shannon's channel capacity account (Hoel, 2017), then why appeal to IIT? It seems that basic considerations about simplicity and scientific methodology strongly favor adopting Shannon's approach. Thus, for IIT to be a complete theory, its models need to be explained only by the IIT metric, as a unique measure that cannot be explained in more reductive and simpler terms. This has not been achieved yet.

Besides the implementation and formalization problems, there is a further question regarding the adequacy of black boxes as models that could potentially verify the isomorphism between causal structures and conceptual structures, which is a major commitment of IIT. If the micro-components are hidden, how are semantic-conceptual structures and cause-effect structures going to be mapped? Perhaps *functionally*, in the sense that computations that invariantly take the same inputs and produce the same outputs could explain an isomorphism of a causal and conceptual kind. But this would be problematic for IIT, because functional accounts are reductive and insufficient to address the hard problem (Chalmers, 1995). This would, therefore, contravene the axioms of IIT. So further details need to be provided about how black boxes model these mappings.

Finally, a different aspect of the formalization of IIT requires specifications concerning context sensitivity and a mathematical understanding of contextuality (for details, see Reference Omitted). Maximally-specific integrated information, based on the axiom of the phenomenal character of consciousness, includes what it is like to experience contents *subjectively*. Semantically speaking, this includes the self-referential, irreducibly contextual, first-person perspective. Here one potential difficulty is the apparently unconstrained multiple realization of the PSC, for instance, modelled in terms of black boxes, which could be incompatible with the maximally specific information of subjectivity. One possibility is that the black boxes are going to be mapped (in terms of homomorphism, presumably) to cognitive skills. Cognitive skills of all kinds, however, are best understood in terms of attention, not necessarily phenomenal consciousness (References Omitted). We further explore this difficulty in section 4.

In sum, no canonical interpretation of data in neuroscience that could verify IIT is possible without a specific solution to the implementation and formalization problems. No measurement will be uniformly correlated to predictions of IIT unless there are models that satisfy IIT and only IIT, and no coordinative definitions could be established without such uniqueness. As long as these issues remain unresolved, IIT is incomplete as a scientific theory. The postulation of models for measuring information, specifically black boxes, is an important step forward. However, if Shannon's theory is sufficient to describe the kind of emergent causal structure associated with the PSC, then IIT is not necessary as a theory of information, including its alleged explanation of phenomenally conscious information.

3. Conceptual Mappings

A crucial commitment of IIT is to model subjective experiences in terms of maximally informative conceptual structures. This mapping between experiences and concepts, as mentioned, has to be isomorphic because experiences are *identified* with conceptual structures (Tononi et al., 2016). Each concept of the conceptual structure has a cause-effect repertoire that constrains past and future states. The mapping between the PSC and the conceptual structure is, presumably, an isomorphism as well. This way of relating concepts to their cause-effect repertoires actually determines the $\Phi^{\max}$ of the entire conceptual structure that constitutes an experience. The $\varphi^{\max}$ measure tracks the degree of irreducibility of a concept (Tononi et al., 2016, 451). It is this combination of isomorphism and irreducible content that explains, according to IIT, the phenomenal character of conscious experiences.

In what sense are concepts irreducible? According to IIT, concepts are irreducible partly because their cause-effect repertoires are also irreducible to isolated components. They are also irreducible in a different sense, because each concept has an irreducible level of interrelation with other concepts. But then it has to be the conceptual content itself that is irreducible once it is integrated into the conceptual structure (independently of the PSC). So an independent mapping between concepts, their contents and experiences is needed to explain conceptual irreducibility. This requires a choice of semantics because one needs a model of how conceptual structures are integrated. In the case of the PSC one needs concrete models for the formalization and implementation of IIT, such as black boxes. Here one needs concrete semantic models of the highly integrated conceptual structures postulated by IIT. Otherwise, the mappings of IIT remain highly underdetermined.

This is the representation challenge for IIT: how to represent conceptual contents (semantic, truth-evaluable contents) in a way that captures their interrelatedness, integration and causal-effect repertoires? Conceptual role semantics seems to be the perfect candidate to address this representation problem because the functional and inferential role a concept plays in the conceptual structure can be identified with its semantic content, which according to IIT, is mapped uniquely to its cause-effect

repertoire.¹ Thus, the isomorphic mapping between conceptual structures and experiences would be determined by a conceptual role semantics, while the isomorphism between the conceptual structure and the PSC is determined by their cause-effect repertoires. The previous section presented several difficulties regarding this kind of mapping between the conceptual structure and the PSC. Here we focus on difficulties concerning the semantic isomorphism between conceptual structures and subjective experiences.

A conceptual role semantics could be a model of IIT's conceptual structure in the sense that predictions of IIT could be satisfied by some type of inferential network with contents that determine conceptual roles. The states with qualitative character would then supervene on (more accurately, "emerge" from) states with such informational conceptual-roles. This would complement IIT with a well-known approach to the nature of concepts and their contents. Highly integrated inferential networks for concepts, for example, would correlate with higher rankings of $\Phi^{\max}$ and a holistic account of meaning would explain how meaning correlates with whole, irreducible experiences, that are highly integrated.

But so conceived, such conceptual networks are, at least in principle, susceptible of being computationally understood and implemented, in terms of algorithms. This would be a model of IIT that would contradict the axiom that phenomenal consciousness is not reducible to computational functions, which can always be understood in terms of component parts. More specifically, since one can give a functionalist account of conceptual role semantics, and since functionalism seems insufficient to account for phenomenal consciousness (Block, 2000; Chalmers, 2002; Nagel, 1974), then conceptual role semantics would be an inadequate model of IIT. But then, what is the right semantic model?

Suppose there is such a model. Because of the high degrees of integration required by IIT, the holism entailed by such a model would be even more robust than the integration of standard conceptual role semantics. But this super-integrated semantics would still need to account for the content of concepts and provide some kind of explanation about standard topics related to this issue (e.g. conditions for concept possession, concept acquisition, *a priori* and *a posteriori* conceptual relations). This is no easy task and given that the level of integration between the concepts of the conceptual structure of experience cannot be related functionally (because of the problem mentioned in the previous paragraph) then the task is considerably more difficult than just providing a semantics for concepts. IIT requires a semantics for concepts that scales up in terms of integration, implying a degree of holism that is more robust than that of standard holistic theories. It is, however, hard to determine what model could satisfy this requirement.

¹ There are many versions of conceptual role semantics, some more amenable to strictly computational approaches than others. But they all share the structural commitment of inferential relations among concepts, which need not be related at all to phenomenal qualitative states. See for instance Block (1986) and Harman (1999) for naturalistic approaches; see Brandom (1994) for a normativist approach.

It would also seem that modular models of cognition and content (Fodor, 1983) would be in principle excluded because of this commitment to causal and semantic holism. Or perhaps non-conscious cognition could be modelled modularly but not conscious content. Both of these alternatives are problematic. The first because the cognitive efficiency-advantages provided by modularity would be absent in the holistic models of the conceptual structure and the PSC. The second because a mapping between unconscious contents and conscious ones would be needed, and this mapping cannot be isomorphic. So this generates yet another problem for the causal and conceptual mappings. Since the conceptual structure is not reducible to information-partitions or to components, it would be difficult to account for conceptual contents operating at a modular, and yet efficient, level.

It seems clear that a maximally specific mapping of conceptual information, not reducible to partitions, would need to be modeled in a holistic way. Traditional objections to semantic holism would need to be addressed (see Fodor, 1998), and a specific conceptual semantic model would need to be offered, in a way that solves the difficulties mentioned so far. But even if such a semantic model is provided, thorny issues remain. One of them is the problem of epistemic access. If the choice is a conceptual role semantics, it is unclear how the inferences relating concepts, or the causal functions that determine their roles, should be modelled holistically. In particular, it is unclear how a holistic architecture would address issues like the stability and cognitive impenetrability of the contents of early perception (Reference Omitted). Even if one adopts a fully inter-penetrable Bayesian framework, for instance, it still remains unclear how to model cognitive *access* to contents and their epistemic status (see Siegel, 2017, chapters 6-7). The mappings alone would not suffice. For them to play a conceptual role they must be somehow accessible for epistemic evaluation.

Another problem is the contextuality of meaning in relation to the uniqueness of the first-person point of view. Concepts can be understood as inferential rules, or definitions, or some kind of disposition to refer successfully to the set that satisfies a certain categorical type. But meaning requires salience, relevance and context sensitivity. Presumably, this would depend on the PSC being finely in tune with relevant social and external environments. But IIT is committed to a strong form of *holistic internalism* because, based on its axioms, identical conceptual structures mapped to PSC must have the same cause-effect powers and the same measure of $\Phi^{\max}$, independently of external factors. The question is how is the specific perspective of a cognitive agent and her conscious experiences related to this internalist structure given that, at least some of the contents of those experiences, are determined by the external environment?

The issue of internalism versus externalism about meaning may present problems for IIT. But suppose these are satisfactorily addressed with an appropriate conceptual role semantics. There is a kind of internalism endorsed by IIT that goes beyond semantic or even epistemic internalism (the views that all meaning and epistemic justification supervene on internal mental states of subjects). As mentioned, the fact that a *type* of conceptual structure is isomorphic to a *type* of cause-effect structure indicates that the

relation between the conceptual structure (identical to subjective phenomenal experiences) and the PSC is not merely one of physical realization. This is because realization is a type-token relation (this is what allows for multiple realization). Thus, it seems that, according to IIT, the causal powers of a mind are uniquely related to subjective experience, rather than physical properties as such. Some might take this to be compatible with dualism, epiphenomenalism or panpsychism (Chalmers, 2002). These are views, however, that cannot be easily reconciled with scientific methodology. If IIT entails one of these views, that means that its status as a scientific theory might be compromised because the constraints reviewed in this and the previous section could not be easily satisfied.

Mapping the conceptual structure, associated with phenomenal experiences to the unique first-person point of view is yet another challenge. Presumably, this mapping cannot be merely an issue of representation. But if the conceptual structure is standing in representation relations to semantic contents, it is not clear how such structure is related to the first-person point of view. Therefore, one could consider the problematic mapping between the first-person perspective and the conceptual structure as part of the difficulties presented by the representation constraint.

The problems presented in this section concern a different type of mapping relation that must be modelled with: a choice of semantics (presumably a conceptual role semantics); the unique, first-person point of view of a subject (a highly contextual issue, not explicitly addressed by IIT); and the epistemic access the subject has to those contents. Given that, in addition to this representation constraints, the implementation and formalization constraints are not fully specified, one must conclude that IIT remains an incomplete scientific theory, still at the postulation-level.

In the next section, we explore a different question: what kind of evidence produced in psychology or neuroscience could confirm IIT? We show that, even if all the above problems are addressed concerning the constraints examined in this and the previous section, IIT may turn out to be an extremely important theory of intelligence or access consciousness, rather than a theory about phenomenal consciousness.

4. The Distinction Between Consciousness and Attention

A critical advantage of IIT over other scientific theories of consciousness is the independence of its verifiability from behavioral data and introspective report. IIT is the only theory that has an explicitly formulated measure that tracks states with maximally integrated information, associated with phenomenal consciousness, without depending at all on behavioral or verbal outputs. The benefits of this way of approaching consciousness are very significant. For instance, in intensive care units, where patients may be in no condition to respond either behaviorally or verbally, IIT might offer critical ways of detecting whether they are conscious, with the potential of improving standard diagnostic criteria.

There are many other advantages of IIT, as we highlight in the conclusion. Here we present a set of difficulties concerning the interpretation of scientific data that might allegedly verify IIT. According to IIT, the only kind of consciousness it measure is *phenomenal* consciousness. As opposed to other theories, which are functionally defined, IIT might be the only theory that is about phenomenal consciousness (the subjective character of experience), instead of access consciousness (the availability of information for thought, action, perception and decision-making (see Block, 1995). The difficulty we present here is that, even if one concedes that IIT can be implemented as a scientific theory, it might not have this distinct advantage because it might only concern access consciousness. In other words, in spite of its independence from reportable or behavioral data, characteristic of functional approaches, IIT might still be exclusively about access consciousness.

As mentioned, according to IIT, subjective experiences are identical to conceptual structures. It is important, in studying the kind of structure concepts have (e.g., inferential, prototypical, compositional), to explain how the application conditions of a concept are shared by thinkers or experiencers. Concepts, regardless of how they are defined, require a semantic structure in order for them to apply to the same referents (the same extension) while having the same meaning or “aboutness.” More important, the contents of concepts, what they are about, need not be defined in terms of conscious experience. The same is true of both the conceptual structure or the possession conditions of a concept. So in what sense is the conceptual structure identical to experiences?

Although IIT does not provide a clear answer to this question, it is plausible to interpret it as claiming that it concerns a specific family of concepts: phenomenal concepts (roughly speaking, concepts that only apply to *your* experiences, rather than to anything external or environmental). This is because the only type of concepts that *necessitate* phenomenology are phenomenal concepts. But this is a problematic commitment. First, possessing a phenomenal concept does not entail that the possession of that concept must depend only on having an experience. For instance, besides having the experience, a linguistic category and an indexical or contextual way of representing it may both be necessary, and these latter conditions might necessitate publically constrained representations or structures, rather than privately determined experiences. In fact, there are good reasons to think that, given that we converge in our application of phenomenal concepts such as ‘pain’ or ‘red’, we must at least publically and linguistically share some of the conditions for their application. But this means, then, that the conceptual structure and its cause-effect powers alone are insufficient for concept learning, concept possession or concept application (at the very least, the conceptual structure is not sufficient for subjective experience because it requires some publically shared information; but the sufficiency of the conceptual structure is a deep commitment of IIT).

Second, the publically determined conditions for the application of any concept are very likely linguistically mediated in the case of humans. This opens a different set of difficulties regarding how to understand the conceptual structure assumed in IIT. Some

of these difficulties depend on the thorny issue of language acquisition and linguistic competence. Other issues depend on the nature of language itself (e.g., is it exclusively human, as Chomsky and Berwick (2016) think, or is it shared with other species). This set of issues highlights the central concern that IIT might, in spite of its axioms, leave the issues of phenomenal consciousness and the hard problem largely untouched. In other words, these difficulties show that even if IIT is satisfied by the models for causation and semantic content, and completed by the theories described in the previous sections, it might be a theory of *access* consciousness. If the conceptual structure is determined by access-relations to information, including public information, then phenomenology is not playing a central role. If so, then IIT's key advantage is lost.

Are there any theoretical or empirical reasons to support the claim that IIT might be a theory of access consciousness besides these general considerations about concepts? Although these general considerations are powerful enough, given the commitment of IIT to conceptual structures, they would be even more problematic if there are independent reasons for thinking that IIT is a theory of access consciousness. In the remainder of this section we present what might be the most important independent consideration of this kind: the distinction between consciousness and attention. We do not argue that these considerations are jointly decisive as a case against IIT. As we have done in the entire paper, our main goal is to present problems that IIT needs to resolve and clarify in order for it to be considered as a scientific theory of phenomenal consciousness with full credentials. We are not claiming these are decisive objections against IIT, and in that sense, we believe that presenting these difficulties can play a constructive role in completing IIT, as we explain in the conclusion.

The application-conditions for concepts (the conditions thinkers must satisfy in order to apply the concept to the same set of things, or extension) must be somehow shared publically (by common internal mechanisms or by common external conditions), because otherwise miscommunication ensues. What is the cognitive process responsible for providing or determining such conditions, if not the subjective experience by itself, which *by definition*, is not shared publically? One of the best candidates to fulfill this role is attention. Independently of choosing what theory of concepts best fits IIT, the cognitive act of categorizing a feature, event or object seems to be perfectly explained by attention routines. This is the case not only with respect to epistemic issues (e.g., justification, knowledge, testimony) but also with respect to mental representation issues (e.g., accuracy conditions for application, reliability, cognitive integration). In fact, one can provide a whole theory of epistemic agency and its normative implications based on attention, without necessarily appealing to consciousness (see Reference Omitted).

We have, therefore, the first part of an answer to the question: how to account for the conceptual capacities assumed by IIT without necessitating phenomenal consciousness? Attention routines are necessary and it seems, sufficient to explain these conceptual capacities. The second part of the question is to provide empirical and theoretical reasons for the dissociation between phenomenal consciousness and attention. The dissociation between consciousness and attention has been documented

empirically through several experiments and is assumed by most of the main theories of consciousness (Reference Omitted). Evolutionary considerations also speak strongly in favor of the dissociation between consciousness and attention (Reference Omitted). This does not mean that subjective experience is not necessary in many cases of conceptual attention. What it means is that, in general, attention *suffices* to explain conceptual routines that require external accuracy conditions, and since subjective experience alone cannot provide such conditions, the conceptual structure of IIT is best understood in terms of attention, without necessitating phenomenal consciousness. Thus, unless more specific arguments and models are provided, a very plausible way of interpreting IIT, given its current formulation, is as a theory of *access consciousness*, rather than phenomenal consciousness.

This is not good news for IIT. If the conceptual structure IIT identifies with experiences is fully captured in terms of attention routines, without necessitating awareness, then its key advantage over other theories vanishes, because all the information in the conceptual structure would be reportable and poised for behavioral manifestation and it could easily be explained by other theories based on reports and behaviors. Even if IIT succeeds, and becomes the best theory of access consciousness, an independent theory of phenomenal consciousness would still be needed. A related problem is that if the information in the conceptual structure is captured in exclusively epistemic terms, then the role of phenomenal consciousness is left unexplained because all these epistemic activities can be captured through attention and access consciousness.

If language is necessary for the type of conceptual representations humans have, then that would reinforce the claim that IIT is a theory of access consciousness (related to intelligence and rationality, but not necessarily phenomenal consciousness). Joint attention to linguistic, publically available conceptual referents, would be necessary for the conceptual structure to play its semantic role. The entire conceptual structure, with its cause-effect repertoires at the right macro-scale PSC might be, even if successfully measured, entirely driven by linguistically-oriented joint attention to features of external objects, intentions of other speakers and conversational backgrounds. All this publically available information should, in principle, increase the amount of information integration and causal efficacy. But as IIT is currently stated, it is not possible to truly account for this form of integration, because of its commitment to the PSC being, presumably, instantiated by areas of the brain that realize the conceptual structure at the right macro-scale. This is an interesting challenge for any theory of information that aims at capturing both causal structures and semantic relations, but it especially pressing for IIT.

To summarize, we presented several challenges to IIT as a scientific theory. Unless they are addressed properly, IIT cannot be considered as a complete scientific theory, in spite of its great appeal and advantages. We examined several proposals for the verification, implementation, modelling and formalization of IIT and argued that more details need to be provided. With respect to the issues of formalization and implementation, we argued that it is not clear what are the advantages of IIT over the best known theory of information, Claude Shannon's theory. A positive account of the

difference would allow scientists to use canonical interpretations of data that Shannon's theory cannot accommodate. So far, it is not clear that there is a substantial difference. With respect to representation and coordinative definitions, we articulated skeptical worries about the type of mapping between the conceptual structure and the PSC, and also about the nature of the conceptual structure and the IIT metric, based on basic considerations concerning the nature of concepts. We expanded on these problems by focusing on the dissociation between consciousness and attention. Although these arguments show that IIT remains at the initial or postulation level, we conclude with a more positive perspective on IIT.

Conclusion

In the previous sections, we presented several difficulties regarding the implementation, formalization and conceptual representation of IIT. Here we want to conclude by highlighting the substantial advantages of IIT. The main point of these concluding remarks is that looking forward, IIT should be studied from multiple disciplines because it is uniquely positioned to bring together all the relevant areas of research needed to confront the puzzle of consciousness.

We first want to emphasize that it shouldn't be surprising that IIT is still in its initial postulation-level. For a theory so ambitious to be fully confirmed (for any scientific theory of consciousness, for that matter) the models that need to be provided and the implementation, formalization and representation constraints that must be satisfied will involve innovative findings across several disciplines, fundamentally neuroscience, psychology, information theory and philosophy of mind. This will take an enormous amount of effort. Our main goal in this paper is not to object to IIT in general, which would actually be premature, given that the theory is still at its very initial stages. Rather, we wish to demonstrate that substantial work remains to be done before IIT is considered as a fully confirmed scientific theory of phenomenal consciousness. But in this regard, IIT is not in an inferior position to other theories and, on the contrary, has advantages that we shall briefly list here.

As mentioned, a key advantage of IIT is that it could be confirmed independently of any behavioral or report-based evidence. This is a great advantage for many reasons, including the importance of having less reflective and linguistically based paradigms for studying phenomenal consciousness. This is an improvement on previous theories of consciousness that transcends its academic study. It could, for instance, transform the methods for diagnosing disorders of consciousness and ameliorate the treatment of critical patients at hospitals. If this were the case, IIT could potentially have ethical consequences.

IIT is the only theory that explicitly addresses phenomenal consciousness. We argued that IIT might actually be a theory of access consciousness. This criticism, partly based on the lack of models for the conceptual structure assumed by IIT, should not overshadow the issue that even if IIT turns out to be a theory of access consciousness, that would still be a remarkable achievement. However, if IIT succeeds at explaining

phenomenal consciousness scientifically, it would also provide new scientific insights beyond consciousness studies. It is very hard to overestimate the deep scientific breakthrough this would be. With respect to its formalization, IIT would foster new insights on information theory and computability, by distinguishing integrated information from extant theories of information, in a way that explains, without appealing to explanatory gaps, what is unique about consciousness. IIT would certainly provide new ways of understanding causation, especially through the isomorphism of cause-effect repertoires and conceptual integration and shed light, theoretically and experimentally, on the metaphysics of consciousness. Finally, it would potentially provide a wholly new way of thinking about concepts and their relation to phenomenal consciousness. These potential findings would produce new ideas that could transform research in many disciplines.

To solve the problems presented in this paper, researchers working on IIT must first use the usual “divide and conquer” strategy. Each of the constraints behind these difficulties must be satisfied independently. But because IIT requires such unique mappings, researchers must then collaborate and work in unison, rather than only come up with results independently. This is an exciting project, but also a very challenging one. Any successful solution to any of these problems would signal substantial scientific advances. In this sense, any researcher in philosophy of mind, cognitive science or information theory must be grateful to the appearance of IIT in the scientific and philosophical landscape.

References

Adriaans, P. and van Benthem, J. (eds.) (2008). *Handbook of the Philosophy of Science: Philosophy of Information*. Oxford, UK: Elsevier.

Berwick, R.C., and Chomsky, N. (2016). *Why Only Us: Language and Evolution*. Cambridge, MA: MIT Press.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy* 10: 615-78.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18 (2):227-47.

Block, N. (1997). On a confusion about a function of consciousness. In *The Nature of Consciousness: Philosophical Debates*, edited by Ned Block, Owen J. Flanagan and Güven Güzeldere, 375-415. Cambridge, MA: MIT Press.

Block, N. (2000). Troubles With Functionalism. In *Readings in the Philosophy of Psychology, Volumes 1 and 2*, edited by Ned Block, 268-305. Cambridge, MA: Harvard University Press.

Brandom, R. (1994). *Making it Explicit*. Cambridge, Massachusetts: Harvard University Press.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2 (3):200-19.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Chalmers, D. J. (2002). Consciousness and Its Place in Nature. In *Philosophy of Mind: Classic and Contemporary Readings*, edited by David Chalmers, 247-272. New York: NY: Oxford University Press.

Fodor, J. A. (1974). Special Sciences: (Or: The Disunity of Science as a Working Hypothesis). *Synthese*, 28(2): 97-115.

Fodor, J. A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Fodor, J. A. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Harman, G. (1999). *Reasoning, Meaning and Mind*. Oxford: Oxford University Press.

Hoel, E. P. (2017). When the Map Is Better Than the Territory. *Entropy* 19(188): 3-15.

Hoel, E. P., Albantakis, L. and Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *PNAS* 110(49): 19790-19795.

Marshall, W., Albantakis, L. and Tononi, G. (2018). Black-boxing and cause-effect power. *PLOS: Computational Biology* 14(4): 1-21.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review* 83 (4):435-50.

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): e1003588.

Suppes, P. (2002). *Representation and Invariance of Scientific Structures*. Stanford, CA: CSLI Publications.

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