THE TRANSCENDENTAL DEDUCTION OF INTEGRATED INFORMATION THEORY: CONNECTING THE AXIOMS, POSTULATES, AND IDENTITY THROUGH CATEGORIES

Final preprint of https://link.springer.com/article/10.1007/s11229-022-03704-z, please cite the published article as:

Chis-Ciure, R. The transcendental deduction of Integrated Information Theory: connecting the axioms, postulates, and identity through categories. *Synthese* **200,** 236 (2022). https://doi.org/10.1007/s11229-022-03704-z

Abstract

This paper deals with a foundational aspect of Integrated Information Theory (IIT) of consciousness: the nature of the relation between the axioms of phenomenology and the postulates of cause-effect power. There has been a lack of clarity in the literature regarding this crucial issue, for which IIT has received much criticism of its axiomatic method and basic tenets. The present contribution elucidates the problem by means of a categorial analysis of the theory's foundations. Its main results are that: (i) IIT has a set of nine fundamental concepts of reason, called *categories*, which constitute its *categorial lexicon* and through which it formulates a *system of principles* incorporating the axioms, the postulates, and the central identity; and (ii) the connection between the axioms and postulates is grounded by their common root in this categorial lexicon, the categories of which find their justification by means of a phenomenological and *transcendental deduction*. Some further results are the unique origin of axioms and postulates in the categories; the distinction between conceptual and formalized postulates; a clarification of the uniqueness problem of categorial lexica in general; and an IIT account of objectivity by explicating how the physical is (re)defined by means of categories. All of this is put to use against various criticism targeting IIT's theoretical core. If successful, the proposed interpretation illuminates a central issue in the contemporary study of consciousness and contributes to an environment of mutual understanding between defenders and critics of the theory.

Keywords: consciousness; Integrated Information Theory; categories; transcendental deduction; a priori

1. Introduction

Consciousness study has become a burgeoning field of science and philosophy. Integrated Information Theory (IIT) is a major player in the game (Oizumi et al. 2014; Tononi et al. 2016; Tononi 2015, 2017a; Haun & Tononi 2019; Barbosa et al. 2021). However, there is persistent skepticism regarding the theory's axiomatic approach, the main issue being the "translation" of axioms of phenomenology into postulates of cause-effect power (see Bayne 2018; McQueen 2019; Doerig et al. 2019; Negro 2020; Merker et al. 2021; Kleiner & Hoel 2021 for different lines of criticism). The research community needs a clear account of the nature of the relation between the axioms and postulates, which has generated much confusion and debate.

Providing such an understanding and filling this literature gap is precisely the aim of this paper. I briefly introduce and deploy the method of *categorial analysis*, which involves a systematic inquiry into the philosophical core of concepts by means of which a theory articulates its foundations. This methodology has been characterized in more detail in Chis-Ciure (2022a: ch. 2, 2022b), work to which the present paper stands as a concrete application. My main results are that: (i) IIT has a set of nine fundamental concepts of reason, called *categories*, which constitute its *categorial lexicon* and through which it formulates a system of principles incorporating the axioms, the postulates, and the central identity; and (ii) the connection between the axioms and postulates is grounded by their common root in this categorial lexicon, the categories of which find their justification by means of a phenomenological and *transcendental deduction*.

In this sense, Sections 2 and 3 put forward the categorial analysis of IIT's foundations, spelling out in detail the axiomatic and operational categories and how they are used to articulate the axioms as principles of phenomenal existence, the postulates as principles of physical existence, the identity as principle of explanation, and other three operational assumptions. After a brief on transcendental arguments, Section 4 gives a transcendental deduction of the operational categories that makes intelligible the "translation" of the postulates from the axioms, differentiates between uniqueness *ab initio* and *in fine* of categorial lexica, and draws the important distinction between the conceptual and the formalized postulates, the latter being mathematical principles of physical existence. In Section 5, I use these interpretative innovations to reply to the criticism that has been advanced against IIT's axiomatic method quoted above. The conclusion summarizes the analytic results of the paper and suggests a possible avenue for future research. I end this introductory section by listing IIT's fundamental categories and principles, which are expounded in detail below.

IIT'S CATEGORIAL SYSTEM

Categorial Lexicon—elements

1) Axiomatic Categories (of Phenomenality)

Existence—Non-existence
Intrinsicality—Extrinsicality
Structuredness—Unstructuredness
Specificity—Genericity
Unity—Reducibility
Definiteness—Indefiniteness

2) Operational Categories (of Objectivation)

Reality
Substance (substrate)
Causality (cause-effect power)

3) Non-categorial Concepts (of Objectivity)—based on categories

The Physical
Atom

System of Principles—based on elements

1) Axioms (Principles of Phenomenal Existence)

Intrinsicality Axiom
Composition Axiom
Information Axiom
Integration Axiom
Exclusion Axiom

2) Operational Assumptions (Principles of Objectivation)

Realism
Physicalism
Atomism

3) Postulates (Principles of Physical Existence)

Intrinsicality Postulate
Composition Postulate
Information Postulate
Integration Postulate
Exclusion Postulate

4) Principle of Explanation

Central Identity

2. Categorial Analysis of IIT's Foundations I: Axiomatic Categories

Categorial Systems. Terminologically, all of a theory's *categories* constitute its *categorial lexicon*, each being a proper unit of the set. Based on these categories, one can formulate using logical instruments a *system of categorial principles* that frames an understanding of consciousness as a scientific phenomenon. To be clear, by "categorial lexicon" I refer strictly to a theory's table of categories; by "system of categorial principles" I refer strictly to those principles of a theory that are formulated using only logic and the categories; and by "categorial system" I refer to the entire foundation of a theory, comprising categories, categorial principles, but also derived, non-categorial concepts and principles (Chis-Ciure 2022a: ch. 2, 2022b).

Axiomatic vs Operational Categories. IIT adopts a phenomenology-first approach and aims to identify, by means of introspection and reasoning (phenomenological reflection), the essential properties of experience. Then, it infers the requirements that a physical system must satisfy in order to instantiate those properties. The contrast between the immediate character of phenomenological data and the mediated, postulational character of explanatory physical properties opens up a distinction between two kinds of categories: *axiomatic* and *operational*. The former are fundamental concepts of reason through which IIT captures the essence of consciousness. The latter are fundamental rational concepts through which IIT bootstraps a physical and, in the limit, scientific explanation of the essential structure of phenomenology. While all categories have a basis in phenomenology, as all theorizing takes place *within* consciousness, only the axiomatic categories can be justified ("deduced") by phenomenological proof, whereas the operational ones need also a transcendental deduction as an entitlement for their use. Moreover, the axiomatic categories come in mutually contradictory pairs, while the operational ones are standalone. None of the categories can be defined in more basic terms, such that the unpacking of IIT's categorial lexicon below is merely a characterization. Yet the categories can be used for defining other concepts, like the concepts of objectivity (e.g., the physical), which are derivative and thus non-categorial. The presentation of the axioms

and postulates relies mainly on Oizumi et al. (2014), Tononi et al. (2016), Tononi (2017a), Haun & Tononi (2019), Barbosa et al. (2021), and Tononi (personal communication).

Zeroth Axiom: Experience exists.

Existence–Non-existence. By the zeroth axiom, experience is existence. Thus, the first category of IIT is *existence*: I can be immediately and indubitably certain that I exist, or that consciousness exists. Experience is directly present to its subject, which is thereby assured that there is something rather than nothing. IIT is a theory of existence as discovered from within consciousness, with the latter itself understood *as* existence. On the contrary, *non-existence* entails that there is no experience.

First Axiom: Every experience is intrinsic.

Intrinsicality–Extrinsicality. Subjectivity is the first essential property of consciousness: every experience exists for its subject, from its own intrinsic perspective, rather than for something extrinsic. The category of *intrinsicality* expresses the subjective character of experience. By contrast, *extrinsicality* implies an entity exists only for something else (e.g., an observer), but not from its own perspective. Combining categories, an entity can both exist intrinsically (as a subject of experience) and extrinsically (as a content of another subject's experience). Yet, experience qua experience can only be intrinsic.

Second Axiom: Every experience is structured.

Structuredness–Unstructuredness. The second essential property of consciousness is that it is structured: every experience is composed of phenomenal distinctions bound by relations. The category of *structuredness* captures the compositional nature of experience, paired with *unstructuredness* as its complement. According to IIT, any possible experience displays some structure, irrespective of how rich or scrutable through introspection it is.

Third Axiom: Every experience is specific.

Specificity–Genericity. The third essential property of consciousness is that every experience is the particular way it is, from which it follows that it necessarily differs from any other experience. The category of *specificity* conveys the idea that experience cannot just exist generically, but rather must be in some particular way or another, i.e., be specific. *Genericity* is the negation of specificity, which means that something is not specified. For instance, a thing must always have a given color, not a generic one. All experiences are specific and cannot be generic, even though they can be general[1].

Fourth Axiom: Every experience is unified.

Unity–Reducibility. Unity is the fourth essential property of consciousness: an experience cannot be reduced to any proper subset of phenomenal contents. The category of *unity* expresses the insight that an experience is one, i.e., it is irreducible to its components. Unity is not opposed to plurality but to *reducibility*, which from the intrinsic perspective amounts to non-existence: a reducible experience is no experience to begin with.

[1] For instance, concepts as phenomenal invariances are general, yet they are still specific contents of consciousness.

Fifth Axiom: Every experience is definite.

Definiteness–Indefiniteness. The fifth essential property is that consciousness is definite, meaning that it has precise borders. The category of *definiteness* expresses the fact that an experience contains only what it contains, neither less nor more, at the level of content and spatiotemporal grain. Its negation is *indefiniteness*.

Axioms as Principles of Phenomenal Existence. Briefly, any given experience is a phenomenal structure that is subjective, specific, unified, and definite (Ellia et al. 2021). Also, consciousness is existence (being), and being is first discovered from within consciousness. According to their Aristotelian origins and development by Kant, categories are "ontological predicates" (Kant [1790]2000: 5: 181, p. 68). Logically, one must always use concepts to articulate judgments or propositions. In IIT, the categories predicate the essential properties that are true of every conceivable experience. Therefore, axioms are *principles of phenomenal existence*, obtained by predicating the categories of all experience as existence. The axioms are taken to be evident, essential (true of all experiences), complete, consistent, and mutually irreducible (Tononi 2015). One could think of these essential properties as "phenomenal forms", insofar they do not dictate any specific content, but they are rather the most general and invariant forms that bound any content that might fill them in a given experience. The categories are thereby the basic conceptual expression of such forms, through which the axioms are made possible. The logical form of the axioms, except the zeroth one—which merely predicates existence of experience—, is 'All *E* are *C*', rendered propositionally as <Every experience is [category]>.

3. Categorial Analysis of IIT's Foundations II: Operational Categories

Operationalization. As a scientific theory, IIT attempts to explain these five essential properties, discovered through introspection and reasoning, in physical terms. However, all analysis and categorial expression have been carried out so far only within consciousness, without considering an experience-independent world—so to speak, 'solipsistically'. The axiomatic categories articulate the axioms as principles of phenomenal, not physical, existence. Therefore, an *operationalization* is required: the essential properties need to be translated into a language that can be intersubjectively shared, one that allows for reproducible operations to be done by different agents, so as to extrinsically probe and validate findings about the subjective experience of a substrate. In order to provide this intersubjective rationality, the categorial lexicon must be enriched. I mention that, despite presenting the categories one-by-one by the necessity of exposition, they can be at play even if not yet introduced. As such, the categorial lexicon functions holistically as an interconnected system of concepts that cannot be analytically separated without loss of meaning.

Reality. According to the category of *reality*, there can be existence that is consciousness-independent, meaning that something can exist without being an intrinsic content of some experience. Against Berkeley, IIT assumes the existence of an external world. Note that neither the reality category nor those of substance and causation are divorced from phenomenology. Our experience exhibits countless regularities: e.g., spatial constancy relative to displacements of visual objects, coherent composition of objects that move smoothly in space, intermodality continuity and congruence when apprehending the same item, etc. (Tononi 2017a). Solipsism cannot explain these regularities, whereas positing mind-independent real entities via the reality category opens a way to understanding.

Realism. By combining categories, IIT can formulate operational principles called *assumptions* (see also Ellia 2021: ch. 3). For instance, by combining existence, extrinsicality, and reality, IIT articulates the assumption of *realism*, according to which entities can exist extrinsically in a way that does not depend on an observer's experience. The reality category is important to make sure that an existent can be 'public', i.e., does not depend for its existence on being experienced and can be an intrinsic content for multiple subjects.

Substance (substrate). Another category essential for operationalization is that of *substance*. The word '*substantia*' means 'something that stands under and grounds things' (Robinson 2018) and traditionally was

understood as a substrate in which accidents (attributes) inhere as predicable properties. As Kant ([1781/1786]1998: B274-9) emphasized, a substance is that which endures across the change of accidents (permanence). In IIT, the category of substance makes intelligible the notion of a *substrate* of consciousness as that on which operations can be carried out to extrinsically probe its intrinsic existence (i.e., consciousness). Adding reality, the substrate can be taken as real in a consciousness-independent sense. Operationally, an observer can consider multiple candidate substrates for analysis, some overlapping or completely contained in others.

Causality (cause-effect power). The category of *causality* is central to IIT's foundations: in the limit, consciousness should be explainable in causal terms, via one-to-one explanatory correspondences between its essential phenomenal properties and the operationally assessed causal properties of its substrate (Oizumi et al. 2014; Haun et al. 2019). Both substance and causation are indispensable tools to explain the regularities of experience, thus they have a basis in phenomenology, not merely in reason. Causality expresses the idea of a dependency relation by which something can change something else as a cause, but can also be changed as an effect. In other words, a thing must be able to "make and take a difference" (be a cause and effect). Combined with existence and intrinsicality/extrinsicality, causality becomes a criterion of existence from an operational point of view: an entity exists only if it can cause something and be caused by something (Tononi 2015). Indeed, operationalization cannot start without a causal interaction between observer and entity. For the moment, having both cause and effect power should be understood as a minimal criterion of all existence, intrinsic and extrinsic. I get back to the intrinsic/extrinsic existence distinction at the end of the section.

The Physical and Physicalism. Proving the conceptual power of the categories, by combining existence, intrinsicality/extrinsicality, reality, substance, and causality one makes intelligible a conception of *the physical* as a substrate which must have cause-effect power to really exist (operationally)—the concept of the *physical substrate of consciousness* (PSC) is essential for the postulates. The physical is a non-categorial concept based on the categories. Furthermore, the categories and the concepts derived from their combination allow IIT to formulate the operational assumption of *physicalism*, according to which the criterion of physical, not phenomenal, existence is having cause and effect power which can be assessed on a given substrate. The physical is thus a real substrate that can "make/take a difference". In agreement with its derivation from multiple categories, physicalism says something more than causation, which is just a minimal criterion of existence; note that IIT allows for both intrinsic and extrinsic physical existence. In the end, our scientific understanding of the existence discovered intrinsically depends on its operationalization in a causal language that allows us to understand the nature of physicality from within consciousness, but still explain the latter in terms of the former.

Atoms and Atomism. Through the categories of existence, structuredness, and substance one can understand a substrate that exists as a whole with constituents, which can themselves be constituted by smaller ones. Since a constituent must also exist (albeit qua constituent), by adding the categories of extrinsicality, causality, and reality one can formulate the concept of an *atom*[2] as the elementary unity that really exists operationally. Through it, the principle of *atomism* states that the explanatory account obtained through operationalization should probe existence down to the smallest units that have cause-effect power[3]. Neither the physical nor the atom are categories, but rather relevant concepts based on the latter. Thus, by means of its categorial lexicon, IIT can form three assumptive principles to bootstrap the operationalization of the axioms.

---

[2] Nothing is introduced through the back door, "atom" is just a name for the smallest unit that can be assessed operationally.

[3] This principle amounts to a methodological reductionism which should *not* be conflated with ontological reductionism (see Tononi 2017b and Grasso et al. 2021).

Postulates as Principles of Physical Existence. The operational translation gives a sufficient reason, in physical terms, for why consciousness exists as subjective, structured, specific, unified, and definite. The theory must show how the causal properties of a PSC account one-to-one for the essential properties of experience. The operational translation makes possible a shift from the immediate phenomenological existence to the physical one, which is inferential or postulational, as shown by the introduction of three other categories besides the axiomatic ones. The task is to explicitly state the postulates as *principles of physical existence* only by recourse to the categories and concepts derived from them. Their logical form is <*P* □is *C*>, where the necessity operator (□) must be understood in an explanatory sense (as opposed to reductive), and which can be formulated propositionally as <The cause-effect power/structure of the PSC must be [category]>. Like its corresponding axiom, the zeroth postulate predicates physical existence of the PSC.

Zeroth Postulate: The PSC must exist (in a physical sense).

Physical Existence. As per above, a PSC exists physically only if it can be causally affected by and causally affect something else. In IIT, the PSC is considered a system of interconnected units, each of which can be assessed operationally according to atomism (Barbosa et al. 2021).

First Postulate: The cause-effect power of the PSC must be intrinsic.

Intrinsic Cause-effect Power. If the physical is defined in cause-effect power terms and intrinsicality refers to existence from one's own perspective, then the PSC must satisfy this intrinsicality criterion through its causal properties. The first postulate as a principle of physical existence states that, in order to be intrinsic, a PSC must have both causes and effects within itself. This principle can be formulated by means of the categories of existence, intrinsicality, and causality, as well as those required for defining the physical, with causation being central to the explanatory work. Via this principle, one can begin to probe whether a real, extrinsic substrate also exists intrinsically as a subject of experience; if the substrate has only extrinsic cause-effect power, i.e., making and taking a difference from something else, it cannot be conscious.

Second Postulate: The cause-effect power of the PSC must be structured.

Structured Cause-effect Power. To explain the structured nature of experience (phenomenal distinctions bound by relations), a correspondence to the cause-effect power of the PSC must be established. The second postulate as a principle of physical existence states that the cause-effect power of a PSC must be structured, meaning that it must be a cause-effect structure composed of causal distinctions bound by causal relations[4]. Haun & Tononi (2019: 9) defined causal relations as "joint cause-effects that *bind* cause-effects within a cause-effect structure". Within a PSC as a system of connected units, different combinations of subsets of units can have causes and effects within the system, specifying causal distinctions, and the cause-effects of different units can overlap, specifying causal relations. Causality, structuredness, and the categories involved in the conception of the physical explain the compositional nature of phenomenal experience. A real, extrinsic PSC must have a cause-effect structure to be a candidate for consciousness.

Third Postulate: The cause-effect structure of the PSC must be specific.

[4] Nothing is surreptitiously added by talking of "causal distinctions" and "causal relations". These are already contained in the category of structuredness, as the notion of a structure implies the notion of a part of the structure and of a relation between the parts. "Distinction" is just a name for a part of the cause-effect structure which is in "relation" to other parts.

Specific Cause-effect Power. There must be a causal property of the PSC that can account for the specificity of every experience. The third postulate as a principle of physical existence mandates that the cause-effect structure of a PSC must be specific, meaning that it must have a particular form. The cause-effect power of both the substrate (system) and each of its constituents (mechanisms) must be specific to yield a particularly shaped structure. This is the causal equivalent of the principle that an existent cannot be generic (non-specific) qua existence, and neither can its cause-effect power as an operational criterion of existence. For this postulate, causality, specificity, and the categories involved in the physical are germane to the explanation. Like the categories, the system of principles built by means of them is an interrelated whole: the cause-effect structure posited by the second postulate is put to use in the third, even though they are properly distinct. Nevertheless, the third postulate could be expressed using "the cause-effect power of the PSC" locution without any conceptual loss.

Fourth Postulate: The cause-effect structure of the PSC must be unified.

Unified Cause-effect Power. To account for the unity of consciousness, the fourth postulate as a principle of physical existence requires that the cause-effect structure of a PSC is unified, meaning that it is irreducible to the cause-effect structure specified by causally independent subsystems. Irreducibility is a further criterion of all genuine existence, as there is no point in considering a system, its constituents, and their relations as physically real if their cause-effect power is reducible to that specified by other entities; reducibility amounts to physical non-existence. Operationally, the method of partitioning, which involves cutting connections among constituents, is the means to extrinsically assess the unity of a PSC's cause-effect structure. Causality is explanatorily central, but only in relation to unity and physicality (derived through existence, intrinsicality/extrinsicality, reality, and substance). Despite being distinct, the fourth postulate also makes use of the cause-effect structure made intelligible through the second.

Fifth Postulate: The cause-effect structure of the PSC must be definite.

Definite Cause-effect Power. The fifth postulate as a principle of physical existence accounts for the definiteness of experience by positing that the cause-effect structure of a PSC must be definite, in the sense of having a border and grain. If consciousness exists and is definite—contains what it contains, neither less nor more, at the level of content and grain—the corresponding cause-effect structure must also be definite, yet the choice of border and grain must be justified. The sufficient reason for the identity of the border and grain, which also excludes alternative ones, is provided by IIT's maximum existence principle. According to it, among competing existents, the one that actually exists is the one that exists the most. Having Φ^{Max} (integrated information) value, which means being maximally irreducible, is IIT's operational criterion for probing maximal existence among different candidates. Note that an observer can operationally deploy the substance category recursively and consider multiple substrates for analysis, at different spatiotemporal and state grains. As per the reality category, a real existent cannot be dependent on an observer for its existence. Therefore, a PSC that exists intrinsically must trump other cause-effect structures specified by partially overlapping, sub-/super-ordinate substrates. It does so by specifying a maximally irreducible (Φ^{Max}) cause-effect structure, which is the operational credential for the strongest claim to existence. Because the PSC that specifies a Φ^{Max} cause-effect structure exists the most, the border and grain at which it exists are exclusive of other possible ones. There is no consciousness present if a cause-effect structure is irreducible, yet submaximally so. For this postulate, the categories of causality, definiteness, unity, and structuredness are central, in concert with the categorial requirements of the physical to kickstart operationalization. The fifth postulate is likewise formulated using the cause-effect structure posited by the second one.

Central Identity: A particular experience is identical with a particular maximally irreducible and intrinsic cause-effect structure.

Identity as Principle of Explanation. The last piece of IIT's theoretical foundation is the central, explanatory identity it posits. Here I am concerned with the origin of the identity, rather than its epistemic function within the theory, which is discussed at large in Chis-Ciure (2022c). The identity should be understood as a *principle of explanation* that is built and grounded by means of the same categorial lexicon as the axioms and postulates, plus an instance of the logical relation of identity (used extra-logically). The relevant categories to *formulate* the identity principle are existence[5], intrinsicality, structuredness, specificity, unity, definiteness, and causality, even if its *deployment* mandates also the other operational categories.

With this principle there are two logical aspects worth noting: (i) the use of the particular form ("an experience, a structure") and not the universal as in the axioms ("every experience"); (ii) the use of the identity relation. The first aspect does not imply that only one or some experiences are identical with such cause-effect structures; the identity applies to all experiences, so universally. The use of the particular stems from the concern with the shape of the cause-effect structure that can account operationally for a particular experience. A cause-effect structure must have a specific shape to explain a given experience. Regarding the second aspect, the identity relation should be understood primarily in an explanatory context. The operational language of cause-effect power allows us to *understand* the phenomenal structure of experience. Yet when it comes to ontology, consciousness has primacy (is fundamental; see also Ellia & Chis-Ciure 2022: Section 4.1), whereas cause-effect power has an explanatory nature (is instrumental). One can describe by means of two different languages, the phenomenological and the causal, the same existence as discovered immediately from the first-person perspective and assessed mediately (operationalized) from the third-person perspective (more on this in Section 4).

Intrinsic and Extrinsic Physical Existence. The physical is defined in terms of existence, intrinsicality/extrinsicality, reality, substance, and causality as a substrate that really exists in virtue of its causal powers. This led to the assumptive principle of physicalism under which the criterion of real physical existence is having cause-effect power assessable on a given substrate. Importantly, an intrinsic physical entity is not distinguished from an extrinsic one by the mere application of either intrinsicality or extrinsicality as a category. Intrinsic as well as extrinsic physical entities satisfy the postulate of intrinsicality. Rather, an intrinsic entity, in contrast to an extrinsic one, satisfies all the causal requirements enforced by the postulates. Only intrinsic physical entities satisfy exclusion, being absolute maxima of irreducible, specific, structured, and intrinsic cause-effect power. In IIT, to be an intrinsic entity is to be conscious, and this marks the "great divide of being" (Tononi 2017b). A purely extrinsic physical entity, being excluded from the domain of consciousness, is merely a relative maximum of cause-effect power, still irreducible, specific, structured, and causally affecting itself, which can help "carve nature at the joints" (Marshall et al. 2017). All this demonstrates the expressive power of the categories and principles when used in combination, allowing for theoretically interesting distinctions to be drawn[6].

4. The Transcendental Deduction of IIT's Categories

Transcendental arguments[7] rose to prominence with the advent of Kantian philosophy, especially the theoretical one (Kant [1781/1786]1998, [1790]2000; see Pereboom 2018 for review). In general, a transcendental argument says that, necessarily, without some *X* being the case, *Y* would be impossible. If *X* is the condition and *Y* the conditioned, then the conclusion of a transcendental argument is the

[5] To emphasize the existence category, the identity could be expressed as "a particular experience exists as identical with…" without any impact on the theory.

[6] See Tononi (2017b) for a more extended discussion of IIT's ontological taxonomy.

[7] See Bieri et al. (1979), Stern (1999, 2000, 2019), and Bitbol et al. (2009) for general discussion and debate on the scope, merits, weaknesses, and prospects of transcendental arguments, in both science and philosophy. There is also much recent work on transcendental arguments in epistemology (e.g., Schafer 2021) and in science (Hoffman 2019; Colombo & Wright 2021).

conditional <*Y*, only if *X*>, which should be read as "*X* is a necessary condition for the possibility of *Y*". In any such argument, there is a *transcendental claim* that posits a *necessary relation* between something being the case and the possibility of something else being the case, either as a premise or as a conclusion. Usually, *Y*, the conditioned, corresponds to an obvious, easily conceded fact about our consciousness, either relating to its existence, its structure as in the invariant architecture of all its particular instances, or its function as in the deployment of cognitive operations (e.g., thinking, judging, believing). On the other hand, *X* as a condition is a non-obvious, philosophically contentious, and thus interesting fact or principle about the world or ourselves. Therefore, the knowledge gained by either establishing *X* or by evidencing the necessary connection between *X* and *Y* constitutes a genuine conceptual advancement, thereby making transcendental arguments worthwhile.

Here I take the necessity to be *metaphysical*, so neither logical, conceptual, or nomological—that is, the necessary relation between the condition and the conditioned obtains in virtue of the nature of things or how the world is, even though I acknowledge other possible construals and uses. A transcendental claim is not to be taken as "valid for all times", but rather necessary in a *conditional* way, provided a set of principles are accepted or a certain research practice is implemented (Bitbol et al. 2009: 17). Transcendental reasoning usually begins with how I or we experience, think, or believe, i.e., it is *first-personal*, even though this is not a necessary feature. I cast IIT's transcendental deduction in a first-personal way, consistent with its methodology.

As discussed in Section 2.1, the axiomatic categories that capture the essential properties of consciousness do not need a transcendental grounding. The reason is that, for them, a phenomenological deduction suffices. This means that one can check the legitimacy of such concepts against one's experience. Likewise, the axioms find their ground in phenomenology. The axiomatic categories necessarily apply to all our possible experiences.

Without its operationalization through the corresponding categories and the principles they articulate, IIT would remain an exercise in pure phenomenology and not a scientific theory. Even if one can give a phenomenological grounding to the axioms, one must still find a ground for the use of operational categories, postulates, and identity. Moreover, one must make intelligible the move from the axioms to the postulates as a "translation" (Oizumi et al. 2014; Barbosa et al. 2021), "inference" (Tononi et al. 2016), or "positing" (Tononi 2017a). I argue that the move is an inference of a special kind: a *transcendental deduction*. Crucially, it is *not* the postulates themselves or the identity that are transcendentally deduced "from" the axioms. It is rather the operational categories that are the buildings blocks of these principles that are transcendentally proved, legitimating thus also the principles of physical existence (postulates) and the principle of explanation (identity). Therefore, I put forward a transcendental deduction that stands at the core of IIT's theoretical foundation and I proceed to unpack it:

> (P1) My consciousness is intrinsic, structured, specific, unified, and definite.
> (P2) The intrinsic, structured, specific, unified, and definite cause-effect power of a PSC is a necessary condition for the possibility of my consciousness.

Therefore,

> (C) There is intrinsic, structured, specific, unified, and definite cause-effect power of a PSC.

> The operational categories of reality, substance, and causation, on which the conceptual postulates, the central identity, and the assumptions are based, are necessary conditions for the possibility of experience, completing their transcendental deduction.

The logical form of the argument is a *modus ponens*, with (P1) being the conditioned, (P2) the conditional as a transcendental claim, and (C) an existential proposition asserting the condition. The necessity attached to the transcendental claim resides in the identity between my consciousness and the maximally irreducible, specific, and intrinsic cause-effect structure specified by a PSC. Albeit, as a principle

of explanation, the central identity is primarily epistemological, it has metaphysical consequences (discussed below), such that there is no need for further grounding: the identity metaphysically necessitates the connection between experience and cause-effect power[8]. IIT is a strictly *monistic* theory of consciousness. As conditioned, (P1) says that consciousness exists such that it has those five essential properties, summarizing the axioms. One might object that the pairing of existence–non-existence does not figure in the premises. The answer is that the "is" in (P1) is both of existence and of predication. It might have been expressed as: "My consciousness exists as intrinsic, specific, structured, unified and definite"; the same holds for (P2). Next, recall that a concept is logically antecedent to a principle, such that the focus in this premise is on the axiomatic categories, not the axioms. There is no further transcendental inquiry for the axiomatic categories, inasmuch as they are grounded in phenomenology. The transcendental conditional claim in (P2) can be seen as an abbreviation of the postulates, even if the accent is on both the axiomatic and operational categories. As premises, (P1) and (P2) can only be expressed propositionally, but the attention should fall on their component concepts. The conclusion asserts the existence of a substrate that specifies a maximally irreducible, specific, intrinsic cause-effect structure, proving that the categories of causation, substance, and reality are not merely figments of reason, but that they are objectively valid as a priori, necessary conditions of all objectivity understood operationally (observation, manipulation, and measurement).

The upshot of this transcendental deduction is a legitimation of IIT's categorial lexicon, which in its turn is the transcendental ground of the axioms, postulates, identity, and assumptions. The so-far problematic relation between the axioms and postulates has been elucidated: both sets of principles are based on some a priori categories that, as concepts of reason, are first discovered within consciousness by means of phenomenological reflection. The postulates are "translated" from the axioms through the operational categories which, in combination with the axiomatic ones, make their formulation possible. The axioms of phenomenology are the guide of this translation, with reason being instrumentally essential. The operational categories themselves have a basis in phenomenology, as the constant inquiry on behalf of reason for an explanation of phenomenal regularities. In agreement with its phenomenology-first methodology and with a common feature of transcendental arguments, the reasoning is cast in a first-person form. The intuition behind this is that everyone should be able to check on their own experience the soundness of the argument.

There is some maneuver space for how to interpret IIT's transcendental deduction. Thus, one can take an epistemic reading (more modest), an ontological reading (more ambitious), or both, as they are not mutually exclusive. In Stern's (2000, 2019) terms, the former yields a knowledge-directed argument, while the latter a world-directed one. In Chis-Ciure (2022a: ch. 2, 2022b), I emphasized the epistemic reading of categories and transcendental arguments; here I am open to both interpretations, acknowledging that the ambitious one might require more work. On one hand, the epistemic reading takes categories as necessary conditions for the *intelligibility* of experience, that is, as ineluctable presuppositions for scientific knowledge of consciousness. This interpretation coheres well with the instrumental role of reason and introspection in explaining the regularities of experience. It also goes hand-in-hand with the fundamental nature of experience, which, according to IIT, can be understood in some terms, but does not depend on anything else for its existence, i.e., it is basic. On the other hand, the ontological reading views categories as necessary conditions for my experience *being how it is*. Here, experience is still ontologically fundamental and the concept of cause-effect power is still operational; yet, due to their identity (as per the principle of

[8] To be clear, concerning its origin, the identity principle itself can only be formulated and grounded by means of the categories (and the law of identity used extra-logically), even though it is employed afterward to explain the connection between the phenomenal and the causal domains. Not only an identity, rather than, e.g., emergence or co-variance, ensures parsimony of categories and justifications, but it also gives sufficiency of explanation in the long run, by opening a space of empirical possibilities for confirming the theory (see also Chis-Ciure 2022c). I discuss below the issue of whether a categorial lexicon can be proved necessary and sufficient apriorily.

explanation) and the requirement of an external, real-world of cause-effect power (as per the operational assumptions), the categories can be seen as necessary prerequisites capturing how the world must be for consciousness to be how it is, i.e., a world of real substrates with cause-effect power. This interpretation hinges on the idea that the phenomenological and causal languages capture the same basic existence, i.e., consciousness, but the first has primacy when we express or characterize our experience, while the second comes to the fore when we aim to explain or understand it in a scientific sense.

A very important feature of all categories and principles built through them is that they are *a priori*. Kant's a priori has to do with the independence of particular sensory experiences or sequences thereof for the justification of some truth-evaluable propositions. For Kant, the a priori is invariant across sensorium transformations and trans-individual sensory variability (Waxman 2013: ch. 3A; see also Chis-Ciure 2022d). In this sense, the categories and the principles that articulate IIT's foundation are definitely a priori, which means that they are immune to empirical revision or confirmation. There is no particular sensory experience that can justify or nullify either the categories, axioms, postulates, or the identity in virtue of its occurrence. IIT's categorial lexicon and system of principles are the *philosophical core* of the entire framework, which enable the mathematization of the postulates and their empirical use. Following Whitehead ([1929]1978: 6), once cleared of logical contradictions, categorial systems of thought are "never refuted", "only abandoned" because of their inadequacy in including "some obvious elements of experience in the scope of the system" or of their incoherence understood as "arbitrary disconnection of first principles". In short, this must be also the fate of IIT's philosophical core if it is to be relinquished: there will be either some experiential dimension left unaccounted for and/or some discontinuity among its principles not obvious at this moment.

One could ask whether a given lexicon as a set of fundamental concepts is unique, and demand proof of it. Heuristically, I understand uniqueness here as necessity combined with sufficiency, given a specific theoretical goal. I distinguish between uniqueness *ab initio* and uniqueness *in fine*. The first one presupposes that a categorial lexicon can be proved a priori to be necessary and sufficient for an explanatory purpose. The second one implies that the adoption of a finite set of categories, which makes possible empirical evaluation, is reinforced through a feedback loop by the very same experimental confirmation it opens—a self-enforcing, virtuous circularity. I claim that is misguided to seek uniqueness *ab initio*: there can be no logical guarantee that a given set of concepts is sufficient for a task, in agreement with Bitbol et al. (2009: 17). Yet it is correct to expect uniqueness *in fine* as a "proof is in the pudding" scenario: if the explanation is completed only by recourse to a given lexicon, the latter is thereby proved unique. One must give credit, so to speak, and see how far one gets. In consciousness science, that scenario is an empirically hardened, global theory that accounts for each and every aspect of phenomenal experience, which lies far in the future. Meanwhile, any categorial lexicon proves its worth by making consciousness intelligible as much as possible[9]. Thus, categorial lexica and systems of experience should not be expected to be unique, as if there could not be in principle other concepts through which to start an account of experience, but rather logically consistent, phenomenologically adequate, and explanatorily exhaustive.

Mathematical Principles of Physical Existence. Insofar as IIT is a scientific theory, its conceptual core must be supplemented by mathematical expression through which empirical applicability (quantitative explanation, prediction, and measurement) is made possible. In this sense, there is an important distinction between *conceptual postulates* as a priori principles of physical existence and *formalized postulates* as mathematical

[9] In personal communication, Giulio Tononi claimed that, in the case of IIT's categories, uniqueness stems from the (conditional) completeness of the axioms and their translation into postulates which, coupled with the identity, can be in-principle experimentally probed. However, one still needs the test of experience to determine the explanatory completeness of IIT's principles beyond conditional status. Thus, an a priori argument could be made that IIT's categorial lexicon is unique if the axioms and the postulates are complete for the task, but this would be a *weak, conditional uniqueness*, the merits of which are not currently evident. Yet I believe *strong, unconditional uniqueness* of categories in a scientific context is an a posteriori affair.

principles of physical existence. The postulates are "assumptions about the physical world and specifically about the physical substrates of consciousness […], which can be formalized and form the basis of the mathematical framework of IIT" (Oizumi et al. 2014: 2). Also: "the postulates are conceptualized in terms of cause-effect power and given a mathematical formulation in order to make testable predictions and allow for inferences and explanations" (Barbosa et al. 2021: 1). However, once the categorial lexicon and system of principles are fixed as a core, they remain invariant through mathematical development (e.g., as in the transition from IIT 3.0 of Oizumi et al. 2014 to 4.0, which is still underway, but see Haun & Tononi 2019 and Barbosa et al. 2021). The conceptual postulates are principles of physical existence that, even though they can be mathematized via probability theory and IIT's intrinsic information formalism (Barbosa et al. 2020), are not in and of themselves mathematical. Neither should they be—otherwise, for example, the move from earth-mover distance to intrinsic distance in Barbosa et al. (2021) would be a change of foundations, which is hardly the case. Furthermore, given the identity, the mathematical principles of physical existence (i.e., formalized postulates) can be supplemented by auxiliary assumptions and used in an experimental context to actually probe the theory empirically. Yet to repeat, there are no explanations, predictions, or extrapolations without some basic categories and principles which are not themselves objects of empirical scrutiny, since they are its possibility conditions.

5. Objections and Replies

Tim Bayne has criticized IIT's axiomatic method:

> I argue that none of the five alleged axioms is able to play the role that is required of it, either because it fails to qualify as axiomatic or because it fails to impose a substantive constraint on a theory of consciousness. (Bayne 2018: 1)

He identified three relevant elements for his rebuke: (i) the axioms themselves, (ii) the relationship between the axioms and postulates, and (iii) the contribution of the axiomatic foundation to the overall epistemic status of IIT. The bulk of his paper deals with (i), devoting only a short paragraph to (iii) as a segue to the axioms' rebuttal. The five paragraphs discussing (ii) consider whether the relation is a logical deduction or an abduction, but ultimately "leave it to the advocates of IIT to clarify the relationship between the axioms and the postulates" (Bayne 2018: 2).

The present paper provides ample material regarding (ii) and (iii), discussing at length the relation between axioms and postulates (Section 4), as well as the epistemic relevance for the theory of the axiomatic and operational categories, principles, and overall categorial system and its legitimation (Sections 2,3, and 4). In this section, I also reply to (i), namely to Bayne's charge that IIT's axioms are either non-axiomatic or non-substantive. While I cannot devote enough space here for an axiom-by-axiom defense—Bayne's entire piece covers that—, I focus on the unity of consciousness captured by integration, which, as a principle, falls short of axiomaticity. Using evidence from IIT sources and Bayne's own work, I show how the 2018 critical paper misconstrues the meaning of the fourth principle of phenomenal existence. Still, in defending the axiom of integration, I make general points that apply to skepticism vis-à-vis the other axioms also.

In *The Unity of Consciousness* book, Bayne (2014: 9-19) discussed three types of unity, what he called "subject", "representational" and "phenomenal unity", expressing interest only in the latter, which he described as:

> The multiplicity of objects and relations that we experience at any one point in time are not experienced in isolation from each other; instead, our experiences of them occur as components, aspects, or elements of more inclusive states of consciousness. It is this fact […] that the notion of phenomenal unity attempts to capture. (Bayne 2014: 11)

Contrast this with how the axiom of integration is discussed as early as Tononi (2008: 219):

> Phenomenologically, every experience is an integrated whole, one that means what it means by virtue of being one, and that is experienced from a single point of view. For example, the experience of a red square cannot be decomposed into the separate experience of red and the separate experience of a square. […] Indeed, the only way to split an experience into independent experiences seems to be to split the brain in two […], but then the surgery has created two separate consciousnesses instead of one.

The same idea is found in Tononi et al. (2016: 452), but also in the paper on spatial experience:

> Integration means that every experience is unified, being irreducible to independent components. Thus, the canvas of space cannot be reduced to a left side and a right side that are experienced independently—if it were so, there would be two independent consciousnesses rather than one. (Haun & Tononi 2019: 4)

Consider thus three facts. First: Bayne (2014) wrote a book in which he treated the topic of phenomenal unity at large. Second: by 2018, there was plenty of IIT literature available to him that made explicit that the integration axiom is rather about phenomenal unity, which the author himself quoted. Third: in the critical paper, Bayne (2018) failed to associate IIT's integration with his notion of phenomenal unity, which would have been a better fit than representational unity, gestalt unity, or phenomenal holism. To be clear, based on the IIT literature presented in Section 2 and quoted here, all of Bayne's readings (representational and gestalt unity, and holism) misrepresent this axiom. Yet, out of these three, phenomenal unity is linked with his discussion of phenomenal atomism and phenomenal holism, where he assigns the latter as a possible interpretation of integration. Thus, I focus on his criticism of integration qua holism because it is conceptually connected to phenomenal unity, which seems to be closer to IIT's fourth axiom, and because the interpretations of integration qua representational and gestalt unity are non-starters.

As a "thesis about the essential nature of consciousness", integration is supposedly not axiomatic. Before answering this objection, I must make explicit a methodological assumption: quoting IIT literature, the author takes "axiomatic" to mean "self-evident within the community of consciousness researchers" (Bayne 2018: 2). Indeed, Oizumi et al. (2014: 2) describe IIT's axioms as "self-evident truths about consciousness—the only truths that, with Descartes, cannot be doubted and do not need proof".

Specifying what this self-evidence means is required to see why, beyond him omitting phenomenal unity, Bayne's criticism of integration misses its target. On one hand, in Section 4, I distinguished the uniqueness *ab initio* from the *in fine* type when it comes to categorial lexica for consciousness. As part of IIT's categorial system, the axioms are built by logically predicating the axiomatic categories of every experience. It is mistaken to interpret self-evidence as a sort of a priori uniqueness: there is no *purely logical* guarantee that a set of categories is unconditionally necessary and sufficient to express the essential nature of experience. IIT's lexicon is no exception. Instead, besides logical consistency and phenomenological adequacy in the short-term, we can expect an *in fine* uniqueness proof of a lexicon, which involves a complete account of all the phenomenal properties of experience that is confirmed by experimental evidence. Moreover, self-evidence should not be conflated with phenomenological adequacy, as the former is way stronger, and Bayne (2018) focuses only on the axiomatic character of the principles of phenomenal existence, not on their fit with phenomenology.

On the other hand, the fact that IIT advocates say that the axioms are "self-evident truths" that "cannot be doubted and do not need proof" should be read charitably as implying that one should look no further than one's own experience to verify such principles, such that no derivation or justification from other sources is needed. Here, I call this a phenomenological deduction. Nevertheless, that does not mean the axioms are revealed to any suitable intellect as if by *lumen supranaturale* (that of God or some extra-ordinary Cartesian innatism). If that were the case, perhaps Greek philosophy would have bequeathed us the five geometric axioms of Euclid and, say, the five noetic axioms of Chyrisippus. On the contrary, despite their obviousness, it has been the experience of IIT advocates that the axioms are hard to convey even to specialist audiences, and are prone to misinterpretation (Giulio Tononi, personal communication).

Ellia et al. (2021) argue that introspection and reason are indispensable tools in making the structure of consciousness intelligible. Without the abstraction power of reason, no knowledge of our experience

would be possible. They also emphasize how limited their deliverances are: "attention, working memory, and reasoning (broadly understood as mental manipulation) are largely selective and sequential", and one "cannot penetrate the structure of experienced space down to its smallest components by introspection alone", while introspecting the structural properties of sound or color experiences is much harder, or even impossible (Ellia et al. 2021: 10). The proponents of IIT argue for a continuous to-and-fro between introspection and reasoning (which I grouped as phenomenological reflection), and neuroscientific evidence. I stress repeatedly this explanatory dialectic throughout the paper. Given the inherent limitations of any source of evidence—reason and introspection not exempted—, the expectation of universal agreement regarding the essential structure of experience, even just among specialists, is too high of a standard to judge a phenomenology-first theory. The variation in introspective and abstracting powers keeps the possibility of creativity and novelty in "axiomatizing" consciousness always open. This does not mean however that IIT's categorial lexicon has to expand. The task is rather to explain all aspects of experience with a finite—and ideally minimal—set of concepts. Hence, self-evidence is neither a priori uniqueness nor some 'inner light' that engenders unanimity. The variance in and the imperfect nature of phenomenological reflection make it rational to debate the axioms, even if, ultimately, their source of evidence should be transparent to all adequately-equipped thinkers[10].

According to Bayne (2018: 6), interpreted as phenomenal holism—the thesis that "[n]o experience can be built up out of simpler experiences"—, integration fails to be axiomatic for two reasons. The first reason: since we do not have first-person access to the cause-effect structure of our consciousness, we cannot adjudicate the debate between the atomist and the holist. The former claims phenomenal distinctions involve "distinct experiences that are bound together by a genuine relation", while the latter that distinctions are "merely differences in content and the binding relation [is] merely nominal". Because there is debate, integration cannot be axiomatic.

First reply to the first reason: none of phenomenal atomism or holism are correct interpretations of the integration axiom. Contents of an experience are neither distinct experiences "glued" together (as per atomism), nor are the relations between phenomenal distinctions nominal (as per holism). Integration says every experience is unified, meaning that it is irreducible to components that are experienced separately (Haun & Tononi 2019; Barbosa et al. 2021). As contents of experience, phenomenal distinctions and their relations are real (not nominal), but proper parts (not distinct experiences). Bayne's (2014) phenomenal unity quoted above seems closer in spirit to IIT's integration than both phenomenal atomism and holism, a connection missed in the 2018 critical paper.

Second reply to the first reason: in my terminology, if self-evidence is properly understood as phenomenological deduction, and not as uniqueness *ab initio* or 'inner light', the possibility of debate does not taint the axiomatic character of integration. If one grasps the fourth principle of phenomenal existence, and finds an experience that fails to be unified (i.e., a counterexample), then integration fails to be axiomatic because it is not essential anymore (i.e., does not apply to all experiences). That would amount to proving a mismatch with phenomenology—Bayne (2018: 5-6) did not provide the proof, nor questioned the match. However, the axiomatic character of integration does not collapse because one can debate how to understand it, for reasons invoked above. The difference is subtle, but crucial. Indeed, given his criticism, the author himself seems to have misunderstood the axiom.

Bayne's (2018: 6) second reason for why integration is not a proper axiom: even if phenomenal holism is true of our experience, "there is no reason to think that it captures an essential feature of consciousness". The point here seems to be that the axiom of integration fails to be true of all experiences. Below, I expound in-detail the third response to Doerig et al. (2019) and Kleiner & Hoel (2021), which

[10] A historical parallel: equating thinking with judging, Kant ([1781/1786]1998) believed he discovered the twelve immutable pure concepts of understanding. Ironically, perhaps with the exception of the thing-in-itself, no other part of the *Critique* was more virulently attacked afterward than his categories table and their deduction.

targets the relevant issue of whether IIT's categorial lexicon is true of all conceivable experience. The reasoning there also applies against Bayne's second point, so I keep it short here. We can only start from our own experience, articulate a categorial system that is the conceptual foundation of a theory, then empirically confirm the theory. We must first account for all properties of our experience in physical terms, and only after extrapolate to other possible forms of consciousness—not the other way around. To bootstrap the scientific study of consciousness, we must start with ourselves, as we have no access to other forms of experience. The allegation that we do not *prima facie* know the character of experience *generaliter* is self-stultifying.

Doerig et al. (2019) proposed an unfolding argument according to which causal structure theories like IIT are either false or non-falsifiable, thus non-scientific. Generalizing this reasoning to all theories of consciousness, Kleiner & Hoel (2021) proposed the substitution argument, which states that if the inferred contents of experience as assessed through report or behavior are independent from the predictions regarding those contents that a theory advances, based on observable internal variables of a system relevant to that theory, then that theory is already falsified. The alternative to the substitution argument is pathological unfalsifiability, which happens when the inferences and predictions are strictly dependent. Considering IIT's phenomenology-driven approach, the authors argue that:

> [I]n order to avoid our results, and indeed the need for any experimental testing at all, a theory constructed from phenomenology has to be *uniquely derivable* from conscious experience. However, to date, no such derivation exists, as phenomenology seems to generally underdetermine the postulates of IIT […], and because it is unknown what the scope and nature of nonhuman experience is. (Kleiner & Hoel 2021: 11)

I have three lines of reply to this criticism[11]. Firstly, what Kleiner & Hoel (2021) miss, besides the common root of axioms and postulates in the same categorial lexicon evidenced in this contribution, is the important distinction between the conceptual postulates as physical principles of existence and the formalized postulates as mathematical principles. This distinction was made explicit through the categorial exposition, which evinced how the general, not mathematical, principles can be formulated (Section 3). True, the underdetermination of the postulates by their mathematical expression cannot be solved purely a priori, but it presupposes a back-and-forth between a priori, mathematical construction, and a posteriori, experimental evidencing, a point to which Negro (2020) also draws attention. I agree with the critics that this cannot be solved merely by means of reason, and so do the IIT authors: "in the end, […] a good account of consciousness will require inferences and extrapolations based on a systematic, complementary back-and-forth between introspection, reasoning, and neuroscientific evidence" (Ellia et al. 2021: 10). In fact, the changes in IIT's formalism are a testament to the multiple ways in which the principles of existence can be mathematized, in a constant trial-and-error in which formal updates are checked for internal coherence and consistency with data, replacing older, antiquated versions. Nevertheless, IIT proponents made a notable attempt to *uniquely derive* the measure of intrinsic information from the postulates (Barbosa et al. 2020, 2021), which is a significant step forward out of this underdetermination quagmire.

However, the crux of the matter is that the axioms are *not* underdetermined relative to the conceptual postulates. Since consciousness is ontologically primary then, if the postulates of cause-effect power would be *only* necessary conditions for its possibility understood explanatorily (that is, as indispensable requirements of intelligibility and empirical objectivity), then the problem of the underdetermination of axioms by postulates would loom large. In this case, there could be no guarantee that this specific set of postulates is the only possible "translation". However, the identity principle is central here, since it claims that phenomenal existence *is* physical existence in operational terms. This preserves the ontological primacy of consciousness as fundamental existence but shows that this existence can be understood in some physical

[11] See Ellia (2021: ch. 4) for a different strategy in answering the unfolding argument, and Chis-Ciure & Ellia (2021) for a defense of IIT against Chalmers' conceivability argument that employs some overlapping resources.

terms precisely because the causal characterization is co-extensive and co-referring with the phenomenological description, insofar as they target the one and only consciousness in which existence in first discovered (recall the ontological interpretation in Section 4). Consciousness is ultimate or fundamental relative to any possible characterization and comprehension, be it phenomenological or physical—indeed, these are descriptions of the two apparently irreconcilable aspects of experience, its subjectivity captured phenomenologically and its objectivity expressed operationally in cause-effect terms. While we cannot help but express the essential structure of experience conceptually, the bridge is built by the identity principle, grounding both these intelligible aspects in consciousness as the fundamental, the ultimate.

Therefore, the uniqueness[12] of the derivation of postulates from axioms is proved: physical and phenomenal existence are one and the same, but made intelligible in different ways, meaning that each and every essential property of phenomenology must correspond one-to-one to an essential physical property defined in cause-effect terms. Intrinsic, structured, specific, unified, and definite cause-effect power is a necessary condition for the possibility of consciousness as intrinsic, structured, specific, unified, and definite, with the necessity grounded in their identity. The identity is first and foremost explanatory but, at the same time, it has inevitable *metaphysical consequences*, namely the metaphysical necessary relation between the phenomenal and the physical. As argued, the axiomatic categorial pairs are immediately grounded in phenomenology, indeed, in no need of a transcendental deduction because they define what it takes for something to be an experience. However, the operational categories of reality, substance, and causality have been thereby transcendentally deduced as necessary conditions for the possibility of consciousness. Therefore, if the axioms as principles of phenomenal existence have a transcendental ground in the categories of phenomenology, the operational principles and the postulates as principles of physical existence have found their transcendental ground in both the axiomatic and operational categories, with the latter being transcendentally proved.

Thus, the a priori foundations of IIT are made explicit, their coherence demonstrated and their completeness tentatively established by their self-containment understood as independence from other principles. This is already enough to diffuse some of McQueen's (2021) and Merker et al.'s (2021) criticism, which pivoted on the alleged *ad hoc* character of the postulates in relation to the axioms, as well as on a confused view of the nature and role of the identity. The principles of phenomenal and physical existence, as well as the identity, originate uniquely[13] in the axiomatic and operational categories, which have been phenomenologically and transcendentally proved.

Secondly, the unfalsifiability thorn of the dilemma in both Doerig et al.'s (2019) and Kleiner & Hoel's (2021) accounts can be given a quick response in light of the proposed categorial analysis. The reply hinges on the *explicit definition* of the physical and formulation of the physicalism assumption in terms of the same categorial lexicon. What this definition amounts to is an *explicit* account of objectivity, albeit one only in its essentials at the moment. The criticism rests on an *implicit* definition of objectivity when it comes to consciousness study, according to which *only* report, behavior, or function can be used as an independent measure to probe the existence and quality of experience in a physical system. But IIT proceeds differently: it starts phenomenologically from the essential properties of experience and translates them into the language of cause-effect power via its categories, and doubly dissociates consciousness from function/behavior. The redefinition of the physical in terms of the cause-effect power of a real substrate, and the identity between the phenomenal and the physical thus defined, are the nucleus of IIT's account of objectivity. This account, like everything else so far, is traced back to the categories, which in Chis-Ciure (2022a: ch. 2, 2022b) I argue should be the case for any major theory of consciousness. In more detail, by

[12] To clarify, the question of whether a categorial framework is unique for a theoretical task is separate from whether the postulates can be uniquely translated from the axioms. My answer to the first is no, at least when it comes to an a priori proof; my answer to the second is yes, by means of an a priori proof.

[13] Logical relations and laws being taken for granted, e.g., the identity principle.

recourse to its categorial lexicon, IIT proposes a meta-language of physicality that defines the physical and thus objectivity (as intersubjectivity) conditions: an entity exists physically only as a real substrate with cause-effect power that can be assessed by an observer. With a conception of the physical in place, IIT builds a language of physicality that specifies what objective (i.e., intersubjectively assessable) properties can account for the structure of experience. In particular, the postulates specify how the cause-effect power of a PSC must explain one-to-one the existence, intrinsicality, structuredness, specificity, unity, and definiteness of consciousness. Then, by the explanatory identity, the gap between the purely subjective phenomenological domain and the intersubjective causal domain is closed—they are one and the same from different perspectives.

IIT cannot be refuted by the unfolding or substitution argument on "logical" grounds alone, because these operate with a different categorial system. Dialectically, the position of the critics is to either prove a mismatch between phenomenology and the axiomatic categories, and/or that IIT leaves some quality of our experience unaccounted for and in principle unaccountable by means of its resources, and/or to show some logical inconsistency in its conceptual foundations. Two competing categorial systems enter the "deadlock phase" when they criticize each other without regard to their specific categorial commitments. If there is no consensus on a minimal set of categories, the debate cannot be settled within any of the two systems. This stage cannot be scientific because it predates all empirical evaluability that is the hallmark of science; indeed, it makes it possible. Instead, what currently takes place is a meta-theoretical debate about the proper paradigm, i.e., objectivity conditions, for the science of consciousness. Given all of the above and the present state of the debate, the integrated information theorist is justified in dismissing the unfalsifiability thorn of the dilemma insofar as the cause-effect power criterion of physical existence and identification of the causal and phenomenal are contradicted—from the viewpoint of another categorial framework. The ball is the challengers' court: either correct IIT's logic, prove its phenomenological inadequacy, or unmask its explanatory incompleteness.

In addition, Kleiner & Hoel (2021: 12) claim that to avoid their results one strategy would be to find "novel forms of inference", which "would likely constitute a major change in the methodologies of experimental testing of theories of consciousness". I have shown here how the method of categorial analysis works in IIT's case and how central a transcendental form of reasoning is to this endeavor, and elsewhere to categorial systems in general (Chis-Ciure 2022a: ch. 2, 2022b). Time will tell whether the results of this investigation will bear fruit for other frameworks as well, and enable an enlarged conception of what objectivity amounts to for a theory of consciousness broadly understood.

Thirdly, is there any guarantee that IIT's categorial lexicon is true of all conceivable experiences, human and non-human? The principle that the same invariant structure applies to all instances of consciousness irrespective of their differentiable richness or non-essential particularities is the experience-equivalent of Hume's uniformity of nature principle, according to which the laws of nature apply homogeneously throughout the universe, in both space and time. This principle is unprovable by the scientific method because it is an a priori presupposition of science, one that grounds induction. While there is a debate on how indispensable complete uniformity is (or whether regularity suffices), the key idea is that one cannot prove this principle directly, especially when it comes to consciousness. We can at most *abstract from within our experience* but we can never *abstract the experience* and check whether alternative forms of it share its essential properties. This is at the core of IIT's phenomenology-first approach: beginning from the experience of a typical adult human being—the only starting point available—, the theory captures its essential properties in axioms and operationalizes them via the assumptions and postulates. If by the central identity the theory is empirically validated in ourselves, which implies a satisfactory physical account of all the properties of experience, then one can extrapolate *from one's own experience* and adjudicate questions about the presence and quality of consciousness in other organisms or non-biological systems. This would be where the *in fine* uniqueness of IIT's categories is proved. At this final stage, an appeal to the principle of the uniformity of experience in nature would add *extra* credibility to the thesis that a categorial system

based on the consciousness of neurotypical adult humans applies to non-human experience too. But the theory based on that system must first account exhaustively for the intricacies of our own.

We are currently far from that stage. However, something can be said about the relation between theory and the uniformity of experience in nature that is relevant at this point. Thus, there is a seesaw between this principle and theory: one could find an indirect justification of the principle if, by assuming it, one is able to construct a theory that ends up explaining and predicting many instances of consciousness in nature, further grounding the inductive claim that all instances of experience follow the same laws. Nevertheless, it is central to my argument that, if one cannot prove the principle, then one cannot use the indeterminacy of its truth-value to argue against a theory, as Kleiner & Hoel (2021) concluded the quote above. As demonstrated, the uniqueness of derivation of IIT's postulates from the axioms is grounded in the phenomenological and transcendental deduction of its categorial lexicon by means of which such principles are first articulated, as concepts are primary relative to propositions. The principle of uniformity of consciousness has no bearing on this, as IIT is silent about its truth-value at this moment. Yet the principle is compatible with the theory, even if the latter proceeds *without* assuming it and tries to fully account for experience. Be that as it may, there is a simple *reductio* against Kleiner & Hoel's (2021) use of the principle in the argument quoted above: if a theory of experience assumes outright that the principle is false, so other forms of experience are incommensurably dissimilar to ours, it pays the huge price of forever being just a provincial theory of human experience. We should not sell ourselves that cheap just yet.

Coming full circle, if a theory assumes from the start that experience is essentially *homogeneous* in nature—IIT does not—, it must still prove the explanatory completeness of its categorial lexicon by getting empirical confirmation. Or, if a theory assumes from the start that experience is essentially *heterogeneous* in nature—IIT does not—, it remains undesirably parochial *ex principio*. Finally, if a theory is agnostic about the truth-value of the uniformity of consciousness in nature principle—IIT is—, it must still prove the explanatory exhaustiveness of its categorial system via the 'tribunal of experience'. The outcome is straightforward: even if it is "unknown what the scope and nature of nonhuman experience is" (Kleiner & Hoel 2021: 11), this putative ignorance cannot be a ground to doubt the "unique derivability" of IIT's categories and ensuing principles from our experience. Indeed, I argued at length why phenomenology does not underdetermine the postulates of IIT—because of the unique origin of the axioms and postulates in the categorial lexicon, and because the phenomenological and causal languages describe from different perspectives the same fundamental existence that is consciousness—, and why IIT can withstand the unfalsifiability charge of the unfolding/substitution argument—because the latter rejects some of the explicit categorial commitments of the integrated information theorist from within another system. Applying also against Bayne's (2018) second point regarding the integration axiom, the defense is completed by noting that ignorance regarding the "scope and nature of nonhuman experience" is orthogonal to the adequacy of IIT's categorial system, which makes possible empirical evaluation in the human case, is (ideally) strengthened by experimental confirmation, and ultimately grounds extrapolating its categories beyond human experience.

6. Conclusion

To sum up, the categorial analysis in Sections 2 and 3 uncovered IIT's categorial lexicon. Namely: (i) the *axiomatic categories* of existence–nonexistence, intrinsicality–extrinsicality, structuredness–unstructuredness, specificity–genericity, unity–reducibility and definiteness–indefiniteness; and (ii) the *operational categories* of reality, substance, and causation. By means of them, the theory puts forward a system of principles: (i) those of phenomenological existence (axioms); (ii) those of physical existence (postulates); (iii) the operational ones (assumptions); and (iv) that of explanation (identity). After a brief on transcendental arguments, Section 4 presented a transcendental deduction of IIT's operational categories, by showing how the maximally irreducible, specific, intrinsic cause-effect structure specified by a PSC is a necessary

presupposition of consciousness. Thereby, the operational categories have been transcendentally grounded and, together with the phenomenological basis of the axiomatic ones, the foundational principles of the theory and their use have found legitimation. In Section 5, the categorial analysis and some of its results, specifically the unique grounding of axioms and postulates in the categories, the distinction between conceptual and formalized postulates, and the redefinition of the physical as an account of objectivity, were put to use against some criticism targeting IIT's theoretical core.

Not only an elucidation of the epistemological foundations of the theory has been achieved, which I hope to engender further debate on a basis of reciprocal understanding between defenders and critics of IIT, but also a new research venue has been opened. In particular, the method of categorial analysis, which involves systematic exposition of the conceptual framework of other theories of consciousness, that is, an inquiry into their a priori, philosophical core as theories of an ineliminably subjective phenomenon, might help solve some conceptual morasses that have plagued this field of study since its infancy, setting its future development on a firmer footing that balances the speculative impetus with empirical checking.

Acknowledgments

I am very grateful to Giulio Tononi for his patience in giving me feedback on an earlier draft, as well as his generosity in filling me in on IIT aspects that were never published before. He was a necessary condition for the possibility of this paper in many ways. Moreover, I thank Larissa Albantakis for greatly improving the paper and my thinking on key matters. Garrett Mindt was instrumental in getting me to write about but not like German Idealists—I'm in his debt for the willingness to read and comment on multiple drafts of this paper. Precious feedback on a first draft is one of the many things I have to thank Matteo Grasso for. I also appreciate Andrew Haun's openness to discussing obscure philosophical issues, quite removed from his interests. In addition, I benefitted greatly from Jon Mallatt's suggestions, which are always thorough and to the point. Francis Fallon was so kind as to help me not put reviewers in a position to have to meet me halfway. I am also grateful to Mircea Dumitru for his generous support in all academic matters. Finally, I want to thank two anonymous reviewers who corrected missteps and made the argument punchier. All remaining faults of the paper are solely due to the author.

References

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

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

Bayne, T. (2012). *The unity of consciousness*. Oxford University Press.

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, 2018(1). https://doi.org/10.1093/nc/niy007

Bieri, P., Horstmann, R.-P., Krüger, L. (Eds.). (1979). *Transcendental arguments and science: Essays in epistemology*. Dordrecht Reidel Publishing Company.

Bitbol, M., Kerszberg, P., Petitot, J. (Eds.). (2009). *Constituting objectivity: Transcendental perspectives on modern physics*. Springer.

Chis-Ciure, R. (2022a). *The A Priori Foundations of Integrated Information Theory: Toward a Transcendental Science of Consciousness*, doctoral dissertation.

Chis-Ciure, R. (2022b). Categorial Systems and Transcendental Reasoning: Why and How Theories of Consciousness Must Redefine the Meaning of Objectivity, under review.

Chis-Ciure, R. (2022c). The Central Identity of Integrated Information Theory of Consciousness as Constitutive A Priori Principle, under review.

Chis-Ciure, R. (2022d). Kant's A Priori in the Context of Constitutive Principles, under review.

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

Colombo, M., Wright, C. (2021). First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese*, 198(S14), 3463–3488. https://doi.org/10.1007/s11229-018-01932-w

Doerig, A., Schurger, A., Hess, K., Herzog, M. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, *72*, 49–59. https://doi.org/10.1016/j.concog.2019.04.002

Ellia, F. (2021). *Integrated Information Theory: An Empirically Testable Solution to the Mind-Body Problem*, doctoral dissertation.

Ellia, F., Chis-Ciure, R. (2022). Consciousness and Complexity: Neurobiological Naturalism and Integrated Information Theory, *Consciousness and Cognition*, 100, 103281. https://doi.org/10.1016/j.concog.2022.103281

Ellia, F., Hendren, J., Grasso, M., et al. (2021). Consciousness and the fallacy of misplaced objectivity. *Neuroscience of Consciousness*, 2, niab032. https://doi.org/10.1093/nc/niab032

Grasso, M., Albantakis, L., Lang, J., Tononi, G. (2021). Causal reductionism and causal structures. *Nature Neuroscience*. http://dx.doi.org/10.1038/s41593-021-00911-8

Haun, A., Tononi, G. (2019). Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy*, 21(12), 1160. https://doi.org/10.3390/e21121160

Hoffmann, M. (2019). Transcendental Arguments in Scientific Reasoning. *Erkenntnis*, 84(6), 1387–1407. https://doi.org/10.1007/s10670-018-0013-9

Kant, I. ([1781/1786]1998). *Critique of Pure Reason* (P. Guyer & A. Wood, Eds). Cambridge University Press.

Kant, I. ([1790]2000). *Critique of the power of judgment* (P. Guyer & E. Matthews, Eds.). Cambridge University Press.

Kleiner, J., Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, 1, niab001. https://doi.org/10.1093/nc/niab001

Marshall, W., Kim, H., Walker, S., Tononi, G., Albantakis, L. (2017). How causal analysis can reveal autonomy in models of biological systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 375(2109), 20160358. https://doi.org/10.1098/rsta.2016.0358

McQueen, K. (2019). Interpretation-Neutral Integrated Information Theory. *Journal of Consciousness Studies*, 26(1–2), 76-106(31).

Merker, B., Williford, K., Rudrauf, D. (2021). The Integrated Information Theory of consciousness: A case of mistaken identity. *Behavioral and Brain Sciences*, 1–72. https://doi.org/10.1017/S0140525X21000881

Negro, N. (2020). Phenomenology-first versus third-person approaches in the science of consciousness: The case of the integrated information theory and the unfolding argument. *Phenomenology and the Cognitive Sciences*, 19(5), 979–996. https://doi.org/10.1007/s11097-020-09681-3

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. https://doi.org/10.1371/journal.pcbi.1003588

Pereboom, D. (2019). Kant's Transcendental Arguments. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2019/entries/kant-transcendental/

Robinson, H. (2018). Substance. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/fall2021/entries/substance/

Schafer, K. (2021). A Kantian virtue epistemology: rational capacities and transcendental arguments. *Synthese*, 198, 3113–3136 (2021). http://dx.doi.org/10.1007/s11229-018-02005-8

Stern, R. (Ed.) (1999). *Transcendental arguments: Problems and prospects*. Oxford University Press.

Stern, R. (2000). *Transcendental Arguments and Scepticism: Answering the Question of Justification*. Oxford University Press.

Stern, R. (2019) Transcendental Arguments. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/fall2021/entries/transcendental-arguments/>

Tononi, G. (2015). Integrated information theory. *Scholarpedia*, 10(1), 4164. https://doi.org/10.4249/scholarpedia.4164

Tononi, G. (2017a). Integrated Information Theory of Consciousness: An Outline. In S. Schneider & M. Velmans (Eds.), *The Blackwell Companion to Consciousness* (pp. 243–256). Wiley Blackwell.

Tononi, G. (2017b). Integrated Information Theory of Consciousness: Some Ontological Considerations. In S. Schneider & M. Velmans (Eds.), *The Blackwell Companion to Consciousness* (pp. 621–633). Wiley Blackwell.

Tononi, G., Boly, M., Massimini, M., Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Waxman, W. (2013). *Kant's anatomy of the intelligent mind*. Oxford University Press.