Philosophia
https://doi.org/10.1007/s11406-023-00684-3

Integrated Information is not Causation: Why Integrated Information Theory's Causal Structures do not Beat Causal Reductionism

Javier Sánchez-Cañizares[1]

Received: 27 October 2022 / Revised: 24 August 2023 / Accepted: 4 September 2023

Abstract
In a recent work (Grasso et al., 2021), practitioners of the Integrated Information Theory (IIT) claim to have overcome the weaknesses of causal reductionism in producing a coherent account of causation, as causal reductionism would blatantly conflate causation with prediction and could not answer the question of 'what caused what.' In this paper, I reject such a dismissal of causal reductionism since IIT anti-reductionists misunderstand the reductionist stance. The reductionists can still invoke a causal account stemming from the causal power of the universe's basic units and interactions that, eventually, may lead to structures supporting integrated information. Additionally, I claim that the IIT-inspired misunderstanding of causal reductionism originates from the former's metaphysical deficit, conflating information with causation. However, as a possible way out, if IIT is complemented with a deeper metaphysical ground, such as nested hylomorphism, an improved argument against causal reductionism can be made to work by invoking formal causality as the ultimate cause of integration in natural systems.

Keywords Integrated Information Theory · Causal Structures · Causal Reductionism · Formal Causation · Nested Hylomorphism · Principle of Individuation

✉ Javier Sánchez-Cañizares
js.canizares@unav.es

[1] University of Navarra, "Mind-Brain Group" at the Institute for Culture and Society (ICS) and "Science, Reason and Faith" Group (CRYF), Campus Universidad de Navarra, Pamplona (Navarra) 31009, Spain

Published online: 16 September 2023

1 Introduction

Integrated Information Theory (IIT) (Balduzzi & Tononi, 2008; Koch & Tononi, 2011; Oizumi et al., 2014; Tononi, 2008, 2017; Tononi et al., 2016) stands out as one of the most promising scientific theories to fill in the epistemic gap regarding what Chalmers (1995) called the hard problem of consciousness. Among its many virtues, IIT provides an operational procedure that should reveal the presence of consciousness in any system, namely a maximum of integrated information (Φ). Even if calculation of Φ can be prohibitively lengthy in real life, surrogate measures thereof seem to enable mathematical discrimination among different conscious states in human beings such as minimally conscious states, locked-in syndrome, REM and non-REM dreams, anesthetized patients, and normal wakefulness (Casali et al., 2013). However, full operationalization of IIT, and how it relates to other complexity measures of consciousness remain an open problem (Arsiwalla & Verschure, 2018; Colombo et al., 2019; Golkowski et al., 2019; Li & Mashour, 2019; Pal et al., 2020; Ruiz de Miras et al., 2019; Sarasso et al., 2019). Moreover, given the relevance of IIT as a scientific theory of consciousness, a collaboration between IIT and rival theories like the Global Neuronal Workspace (GWS)—see for example (Baars, 2005; Dehaene & Changeux, 2011)—has recently started in search of agreement on how to discriminate between them empirically (Melloni et al., 2021).

From the slant of the philosophy of mind, IIT's main message underscores the identity between Φ and consciousness (Tononi, 2008, p. 232). Admittedly, new versions of IIT have developed the needed structure of integrated information—not to be understood as a scalar number any more—to match it to the rich composition of conscious experience (see, for example, (Albantakis et al., 2022; Tononi et al., 2022): each experience is identical to the cause–effect structure, a maximally irreducible conceptual structure where each concept-quale has its own integrated information φ, unfolded from a maximal substrate, i.e., with maximal Φ. Such identification between structured information and consciousness stems from the construction of this theory. In short, one translates phenomenological features of consciousness into IIT axioms from which the mathematical properties of consciousness naturally follow. Notice that the physical properties of the organic substrate of the conscious experience need an operationalization; they are given a mathematical form, but they are not themselves mathematical (Chis-Ciure, 2022). Whether such a methodological move makes sense remains controversial for several reasons: lack of differentiation between types of consciousness (Pautz, 2019), circularity as a result of inter-defining information and causation (Baxendale & Mindt, 2018), or the intrinsicality problem (Mørch, 2019; Sánchez-Cañizares, 2022c). However, IIT does not shy away from entering into deeper metaphysical waters by squaring the specific mechanisms providing Φ—a conceptual structure—with the maximal cause-effect power present in the system (Oizumi et al., 2014; Tononi et al., 2016, p. 453). Said structure will emerge in the system at some specific spatiotemporal coarse-graining. The identification of maximal integrated information with strongest or maximal causation automatically renders itself amenable to metaphysical analysis, ranging from panpsychism (Tononi, 2015) to functionalism (Cea, 2020), as well as different forms of hylomorphism (Owen, 2019, 2021b; Sánchez-Cañizares, 2022a).

Recently, based on the formalism of causal structure analysis inspired by IIT (Albantakis et al., 2019), some authors have claimed that the IIT-inspired measure of causal strength implies an ultimate blow to reductionism (Grasso et al., 2021). Reductionism would allegedly fail to account for the existence of composite mechanisms that are irreducible to their elementary constituents. Grasso et al. (2021) illustrate, through a toy model, the irreconcilable differences between the IIT-inspired understanding of causality and the reductionist one. Briefly stated, all reductionist accounts would lack a principled, explicit approach to analyzing causal structures, one of the benchmarks of IIT that supposedly requires rejecting reductionism as a metaphysical bedrock compatible with the theory. Nevertheless, as Sect. 5 will show, it is far from clear that a reductionist will accept such arguments since IIT relies on an allegedly non-reductive mechanistic model that might as well become reductive or, at least, be accounted for by a reductive interpretation. Consequently, something beyond mechanisms could be essential should one wish to rule out reductionism.

This paper aims to show why (1) the reductionist account may still feel at ease with the IIT-inspired view of causality presented in (Grasso et al., 2021) and why (2) an improved metaphysical framework for IIT, particularly nested hylomorphism (Sánchez-Cañizares, 2022a), can do the trick, i.e., provide more cogent reasons that make reductionism untenable if Φ marks off the presence of consciousness. Nested hylomorphism accepts that maximal Φ is a necessary but otherwise insufficient condition for the emergence of consciousness, and thus rejects the pretense of exclusive causal power to integrated information. The crux of the issue relies on IIT's too-narrow view of causality as causal structures that, neglecting the 'nestedness' and specificity of different levels of causality, easily fall prey to the reductionist stance. In the following paragraphs, I will thus endeavor to show the gist of reductionism, IIT's attack on the former, potential reductionist replies, and how nested hylomorphism may complement IIT's framework. More specifically, I will present what reductionism is all about (Sect. 2), IIT's main argument against reductionism in general (Sect. 3) and in particular (Sect. 4) in (Grasso et al., 2021), and how the reductionist can answer IIT's objections, which, ultimately, boils down to rejecting causal power to the integration of information as presented by IIT (Sect. 5). Before reaching my conclusions (Sect. 7), I will show in Sect. 6 why and how nested hylomorphism provides a better route to disprove reductionism.

2 What is Reductionism all About?

To be sure, scientific reduction is an essential ingredient of the scientific enterprise.[1] According to the words of Frank Wilczek, summarizing the spirit of modern science since Newton, a complex object or subject has been reduced to something simpler when it has been shown, or made plausible, that the more complex thing can be analyzed into simpler parts, and its behavior understood from the behavior of those parts (Wilczek, 2015). Scientific reduction is far-reaching because it constantly searches

[1] However, how to interpret scientific reduction can be highly controversial, as Gillett (2016) has shown with several examples.

for deriving dynamics of physical systems from a minimal set of entities and interactions. Contemporary analytical philosophy of science distinguishes among different types of reductive frameworks (see for example (Stoljar, 2021; van Riel & Van Gulick, 2019) for a map of possibilities). Yet, for our interests, one can focus on the presentation and alleged refutation of reductionism made in (Grasso et al., 2021).

Nevertheless, before introducing the IIT arguments, one should stress as a crucial tenet of most reductionisms that they rely on an underlying or fundamental level—even if still unknown or unknowable—bearing the most relevant ontological weight. Depending on the ontological gradation attributed to the other levels of description, ranging from the diminished reality of mere appearances to sheer epiphenomenalism, one may classify reductionisms in keeping with their epistemic assumptions. The idea of "fundamentality" pervades all of them, though, as the "appearance from reality" criterion (Sánchez-Cañizares, 2019; van Fraassen, 2008, p. 292) only too well illustrates. The more reductionism confers a single ontological value to its favorite level of description, the more it becomes a type of ontological reductionism that embraces a germane kind of monism.

A telling example of reductionism in the slant of the philosophy of mind, extensible to biology and physics, is one of the possible interpretations of supervenience (Kim, 2010). A property Y is supervenient on a set of properties or facts X if and only if some difference in X is necessary for any difference in Y to be possible. Yet even if supervenience is different from reduction in general, a supervenient property might be reducible to the set of properties of its constituents (its supervenience base). Here, the essential point for our further discussion is that the supervenience base—taking the place of the most fundamental level for this kind of reductionism—contains the basic entities and their interactions. In that regard, supervenience does not necessarily imply functional irreducibility but rather, on the contrary, constitution. Said constitution at the most basic level of specific entities and their interactions, namely, *this* concrete dynamic, provides all the causality needed to explain the emergent properties. Consequently, one needs not speak of further causalities as inter-level causation since it is such concrete constitution or configuration of the fundamental level—the supervenience base—that causally explains what emerges at higher levels (Moreno & Mossio, 2015, Chap. 2). Ontological monism lives off its single ontological supervenience base.

Said reflections possess a sufficiently general value for the characterization of reductionism. Moreover, the reductionist foe that IIT sets out to defeat, appealing to the causality implicit in its formalism, proves to be more powerful than the characterization of (Grasso et al., 2021) suggests. Let us now address their main arguments against reductionism and the broader reductionist defense, in order to assess their respective validity.

3 IIT-inspired Attack on Reductionism

(Grasso et al., 2021) describe the reductionist's presuppositions as follows: "if we know what causes each element of a system to do what it does individually, we know all we need to know to predict the system's behavior as a whole, so there is no room

for additional causes to do anything". One may thus safely assume that IIT-inspired anti-reductionists deem the set of elements of a system as the most basic level of description—as explained in the previous section—and the system as a whole, with its specific dynamic, as a higher level. Consequently, if the initial premise holds—and this is a big if, but logically tenable—the reductionist will be able to predict the behavior of the higher level. Yet, as there are no additional causes, such procedure would imply conflating causation with prediction, two different notions that one can easily dissociate. Moreover, for the above-mentioned authors, the "widely shared reductionist intuition about causation seems to be based, ultimately, on two related notions: prediction and supervenience" (Grasso et al., 2021, sec. Box 1: Causation, prediction and supervenience), neither of which is synonymous with causation.

At this point, one may readily wonder whether such dissociation necessarily affects the reductionist's stance. As a telling example, a monist and supervenient account of biological causation denounces, on the one hand, the habitual merging in the literature of both concepts—reducibility and derivability in their jargon (Moreno & Mossio, 2015, Chap. 2.2). But, on the other hand, it defends "a constitutive interpretation of relational supervenience, according to which supervenient properties can in principle be reduced to the configurational properties of the supervenience base" (Moreno & Mossio, 2015, p. 45). Said work, however, provides a mixed response to the existence of irreducible properties which may generate distinctive causal powers: in the affirmative, if one compares them with their emergent base (an unstructured set of lower-level elements); in the negative, if one compares those causal powers with their supervenient base (a set of lower-level entities along with their relational properties, i.e., their specific interactions) (Moreno & Mossio, 2015, Chap. 2.3). The crucial question here is the point of comparison: there are irreducible properties if one compares the latter with the properties of an unstructured set of lower-level elements, i.e., if one only considers their generic, potential interactions; there are no such irreducible properties in the comparison of the alleged irreducible properties with the properties of a structured set of lower-level elements, i.e., with their specific, actual interactions. In Sect. 5 this argument will be further developed.

IIT-inspired anti-reductionists affirm that causal reductionism cannot provide a complete and coherent account of 'what caused what' because "an account based on micro-units, such as individual neurons, taken one by one (first-order), might in principle predict what happens, but will not explain why or allow for meaningful inferences" (Grasso et al., 2021, sec. A simple example: causal reductionism). Noticeably, these authors call causal reductionism 'first-order prediction', drawing on first-order mechanisms that provide all that is needed to predict everything about the dynamics of a system. Yet, this is hardly surprising from the reductionist's viewpoint since it keeps tight bonds with a deterministic universe in which the state of the world at some specific time together with the laws of nature determine both the past and the future. On the contrary, from the IIT's perspective, "only the analysis of causal structures can provide a coherent account of 'what caused what'" (Grasso et al., 2021, sec. Conclusion). In short, the main point of contention here seems to depend on the causal power of 'higher orders', existent for IIT, but unnecessary and, consequently, with a compromised existence for the reductionist.

IIT anti-reductionists rely on an interventionist notion of causation, aligned with the Eleatic principle that makes existence and causation coextensive, and heavily drawing on counterfactuals. "[C]ausation should be understood as the ability of a mechanism to 'take' or 'make' a difference, as demonstrated through observation and manipulation" (Grasso et al., 2021, sec. Box 1: Causation, prediction and supervenience). That definition of causation perfectly aligns with the IIT protocol that invokes the comparison of informational content among different partitions and perturbations of the relevant system—amenable at least in principle to physical interventions for causal discrimination. "Whether a high-order mechanism can make (or take) a difference in a way that is irreducible to the difference made (or taken) by its parts should be assessed through causal structure analysis rather than ruled out based on unexamined intuitions" (Grasso et al., 2021, sec. Box 1: Causation, prediction and supervenience). Yet, whereas IIT anti-reductionists rightly point out that prediction and supervenience do not directly speak about causation, as supervenience only refers to general relations between *explanans* and *explanandum*, this last fact does reveal main differences in both approaches because of their different lens or explanatory logic. Hence, let us see how the IIT anti-reductionists try to downplay the reductionist logic via a concrete example.

4 Of Frogs, Bugs, and Super-bugs

In a toy model, (Grasso et al., 2021) introduce three kinds of frogs (F1, F2, and F3) that differ in their internal wiring and that may encounter left- or right-bugs (their prey) or super-bugs (their predator). It makes sense to assume that when frogs encounter bugs, the former tend to jump on the latter so that they catch bugs, whereas if frogs encounter super-bugs, the former tend to jump over the latter as otherwise super-bugs capture them (Grasso et al., 2021, Fig. 1). The gist of the author's argument focuses on the frogs in the F2 group, which have a more efficient brain than the F3 frogs: F2s possess just two central neurons, C_L and C_R, that trigger when a left- or right-bug shows up, respectively. Additionally, said neurons also trigger when a super-bug appears, and cause the F2 frog to jump over its predator. F2s' wiring thus preserves similar super-bug detection and avoidance functions, similarly to what happens with F3 frogs, but makes the latter's additional super-bug neuron C_C redundant.

Whereas the successful upshot for F2s and F3s is the same, there is a crucial difference in causal terms between F2 and F3 frogs: the super-bug as such never shows up as a cause in F2 frogs (Grasso et al., 2021), lacking the super-bug neuron C_C. In the reductionist causal account, so the argument goes, the firing of C_L and C_R corresponds to detection of the left and right side of the super-bug, respectively, but there is no proper representation of the whole super-bug. The reductionist will only admit first-order causation of C_L and C_R—as independent causes—but not high-order mechanisms of causation like the composition $C_L C_R$. In other words, "a first-order, reductionist causal account sees the super-bug as a cause (…) in F3 frogs, but excludes it as a cause in F2 frogs" (Grasso et al., 2021, sec. A simple example: causal reductionism).

F2 and F3 frogs were assumed to have three sensors, S_L, S_C, and S_R, conveniently reporting to their two or three neurons, respectively. The authors elaborate their point by imagining another kind of frog, F1, with only two sensors, S_L and S_R, and one central neuron, in two varieties: left-F1 frogs detect only left-bugs (the central neuron only triggers when S_L signals); and right-F1 frogs detect only right-bugs (the central neuron only triggers when S_R signals). Quite obviously, both varieties remain incapable of spotting super-bugs. However, if a left-F1 and a right-F1 are, side by side, in front of a super-bug, they would still avoid it by jumping on the left and the right, respectively. "In the case of two F1 frogs, then, the reductionist account happens to capture everything: there are two separate causes leading two separate F1 frogs to a fortuitous escape" (Grasso et al., 2021). Nevertheless, according to the authors, here lies the problem, as "causal reductionism cannot distinguish this case from that of an F2 frog: unlike the two F1 frogs, F2 frogs have evolved an efficient, second-order mechanism, C_LC_R, whose activation has a clear cause, the detection of a super-bug, and a clear effect, the escape response" (Grasso et al., 2021, sec. A simple example: causal reductionism). One could still simplify matters further by saying that, ultimately, for a reductionist, a super-bug is simply two bugs, left and right, showing up partially overlapped and simultaneously aligned, and an F2 frog is just a left- and a right-F1, side by side.

On the other hand, if one introduces the analysis of IIT, the presence of integrated information for the specific arrangement—structure and dynamic—of units reveals causation at different levels: "causal structure analysis establishes that both F3 and F2 frogs have irreducible mechanisms for the detection and avoidance of super-bugs, with corresponding causes and effects, whereas pairs of F1 frogs do not" (Grasso et al., 2021, sec. Causal structures). Benefitting from the equivalence between causality and integrated information, the same authors claim that "causal reductionism fails to recognize the super-bug detector C_LC_R as an irreducible high-order mechanism with a cause in its own right. When it comes to high-order mechanisms constituted of two or more units, then, causal reductionism becomes incoherent with respect to irreducibility and ends up missing obvious causes" (Grasso et al., 2021).

Beyond the additional technical points that IIT provides for its analysis of causal structures, the toy model presented here illustrates the crux of the argument in this controversy. According to IIT anti-reductionists, C_LC_R is a cause—a second-order mechanism—different from the pair C_L, C_R, as the reductionist would ultimately have it. Moreover, there seem to be cogent reasons for it since the conjoint activation of C_LC_R, together represents quite a different reality in the case of F2 and F1 frogs. One might still reach the gist of the argument by saying that the information provided—and consequently, in the IIT language, the causation implied—by C_L (or C_R) is different in the structure C_LC_R than in the pair C_L, C_R. The context for C_L (or C_R) is different in both cases and such distinction, supposedly, remains in the dark for causal reductionism. "It assumes that first-order mechanisms are causally irreducible but fails to recognize that higher-order mechanisms can be just as irreducible, having their own irreducible cause and effect" (Grasso et al., 2021, sec. Conclusion).

5 Potential Reductionist Replies

To avoid misunderstandings, one should begin this section by recalling the first and foremost tenet of reductionism, at least in the view of (Moreno & Mossio, 2015) and that was already introduced in Sect. 2: the constitution or configuration of the fundamental level (supervenience base) causally explains what emerges at higher levels. At this point, one should already notice that, strictly speaking, there is always an asymmetry for the reductionist between the levels of reality which, ultimately, reflects their different ontological weight. Even if the reductionists refer to 'higher levels,' they lack the tools to define them unambiguously. One presumably refers to a higher level as just what is manifested, observed, or experienced closely enough to a human-cognitive level. Hence, an implicit suspension of the ontological value of such (supposedly) higher levels is at play within the reductionist's stance. Does the IIT anti-reductionist fare much better here?

The reductionist holds that micro-causes make a difference. Moreover, they are the only stuff capable of doing so. As reductionism does not have a language to care about higher-level effects, its narrative of causality clings to fundamentality in entities and to their interactions as bringing about this specific constitution of the world (or part of it). The problem with the toy example of frogs is that, to be consistent, a reductionist will not deem frogs entities in themselves but just a specific coarse-graining or set of fundamental entities (sensors + neurons + muscles) that, together with the rest of the world (left-, right-bugs, and super-bugs as another convenient coarse-graining), could help to get a picture of what is going on. But no coarse-graining of fundamental entities and interactions, whatever these are, causes something different from what the above-mentioned specific constitution causes—'this' particular state and dynamic in the configuration space. Of course, reductionists admit that they might be wrong in the particulars, but does the IIT argument presented throughout Sects. 3 and 4 prove that reductionists are wrong in general? Hardly.

First and foremost, reductionism does care about analysis. Therefore, it must start assuming what the basic entities and their interactions are—not differently from IIT. But reductionism refrains from conferring ontological value to any of the possible subsets of states in the configuration space—subsets of states, especially those most-likely visited, would usually represent systems with a certain identity over time. In our referred-to toy model, neurons are as important at a fundamental level as sensors and muscles of the so-called frogs, and as the basic units that bugs and super-bugs consist of. For the reductionist, the basic units have no precedence over each other. Does one wish to talk about integration? Fair enough, but beware that: (1) such integration proceeds from the ultimate laws of nature—interactions among the fundamental entities—that IIT does not discuss, since it initially draws on the correlation of information in the working of logical gates—IIT's favorite modeling; (2) integration is secondary or derivative for the reductionist, as all its alleged causal power still depends on its constitution. Of course, reductionism, in its different varieties, could hardly avoid the accusation of Humeanism—that the world consists of configurations of fundamental particles in space-time—but that is a slightly different problem (or perhaps not, as I will show in the next Section).

As stated in previous sections, IIT anti-reductionists accuse reductionism of conflating prediction and causation. Grasso and coworkers (2021, sec. Box 2: Dissociation between causation and prediction: an example) state that "in general, it is not possible to predict the next state of a neuron or set of neurons based on the output of each individual input unit, for example, due to nonlinear interactions"[2]. They use an XNOR gate as an instance of how individual inputs (B and B') may not produce predictive information about the next state. The authors thus deem the reductionist account insufficient since it always refers to micro-causes as single, individual inputs, neglecting a holistic view of causes that exclusively can predict the final state. In brief, B and B'—together and not individually—are the sole cause of the XNOR gate's output. Analogously, turning back to the frogs, $C_L C_R$, together, are the cause for F2 to jump over the super-bug, very differently from lucky F1s jumping over the super-bug because of other reasons. In the case of F2s, $C_L C_R$, together, cause the activation of the left and right muscles of F2. On the contrary, for F1, C_L alone is the cause of the activation of its left muscle, and C_R alone is the cause of the activation of its right one. In other words, the F1s present a causal reduction, whereas there is a non-reductive causal structure in F2s.

However, reductionists could willingly retort that their reductionism does not depend on finding 'single causes' as they also care about the whole process entailing the complete picture. There is little surprise for reductionists if a non-linear combination of micro-causes brings about non-trivial effects. Similar to what happens with IIT, reductionism struggles to individuate the most fundamental units relevant to the description of a particular phenomenon, amongst the different possible coarse-graining of physical quantities. Reductionists would undoubtedly employ manipulations of the system and counterfactuals to sift among apparent fundamental causes; the problem being—not unlike in IIT—what such causes actually are. But the reductionists do not assume that only individual causes bring about individual effects, or that individual causes add up linearly to produce an effect.

Against what (Grasso et al., 2021) present in their toy model, reductionists will merely claim that they are also entitled to consider the system's fundamental units, not in isolation but in actual (generally non-linear) interaction, and that said units and their specific interactions bring about the final result or the system's state of interest. But claiming that $C_L C_R$ is different from C_L, C_R mischaracterizes the reductionist position. IIT-inspired anti-reductionists affirm that "to demonstrate causation, we need to show that something takes or makes a difference, as assessed through perturbations and partitions. Unlike prediction, high-order causation must be assessed in its own right and does not automatically follow from first-order causation" (Grasso et al., 2021, sec. Box 2: Dissociation between causation and prediction: an example). Nevertheless, these authors seem to identify reductionism with their so-called 'first-order causation' by equating the former's supervenient base with a set of single units bereft of their actual interactions. (Grasso et al., 2021) are the ones who reduce causal

[2] For reductionism, predictions are tools that facilitate the painstaking task of determining actual causes in a concrete process. But reductionism is aware of the many situations where one cannot make predictions with the required precision to discriminate between causes. Such inability to predict do not jeopardize reductionism's basic tenet, i.e., there only exist fundamental particles and their interactions.

reductionism to first-order causation, which makes no sense for reductionism, at least because of the ambiguous ontological weight attributed to higher levels in reality. However, it is not that kind of irreducibility that will defeat reductionism. (Grasso et al., 2021) are wrong in equating the reductionist's supervenient base with a set of single units and a first-order mechanism because the reductionists' supervenient base consists of a set of single units and their specific, actual interactions. Whereas these authors say that $C_L C_R$ is an irreducible cause in itself, the reductionist rejects such claim. What does exist for the reductionist is C_L, C_R, and the interactions among them and with other basic units, irrespective of the fact that IIT practitioners wish to consider $C_L C_R$ as a second-order mechanism because they act simultaneously.

To sum up, the main confusion in the IIT characterization of reductionism stems from considering the causal influence of what IIT practitioners call micro-, first-order units as independently producing an effect. On the contrary, for the reductionist, the causal power of true fundamental units is not separated into mechanisms of different orders. It means that IIT's first-order units are not necessarily what the reductionist calls fundamental units or fundamental particles. The reductionist logic assumes an ontology where the primary entities are fundamental particles and interactions. Of course, one may describe reality at different levels of description—e.g., mechanisms of different order in IIT—because of the existence of non-linear and reciprocal interactions.[3] Said interactions, simplified by IIT in a transition probability matrix, allow for different levels of description. But, according to the reductionists, the composite causality defended by (Grasso et al., 2021) can always be reduced to the causality of fundamental particles and interactions.

6 Beyond IIT's Metaphysics of Causation: Nested Hylomorphism

Are we thus at a stalemate? Indeed, reductionism as a metaphysical position may be unassailable if one accepts unknown (maybe unknowable) initial, fundamental causes and a deterministic law of nature. However, there might be a better metaphysical route to overcome reductionism that, at the same time, benefits from IIT's insights and improves them. And it has to do with the integration not of information but of causation itself. As a matter of fact, IIT anti-reductionists underscore the existence of causal composition (Albantakis & Tononi, 2019). Causal structures make a difference that overcomes the sheer input-output modeling of a black box, i.e., regardless of the system's internal dynamics. However, IIT merely identifies integrated information with integrated causation without further insight about how the former entails the latter. It unavoidably runs into circularity by inter-defining information and causation (Baxendale & Mindt, 2018).[4] True, Φ may reveal the presence of an irreduc-

[3] Whereas reductionism assumes that specific non-linear interactions between fundamental particles, along with those fundamental particles, constitute a sufficient cause for what there is, IIT anti-reductionism places non-linear interactions and effects among higher-order mechanisms. However, IIT mechanisms are just an epistemic description for the reductionist.

[4] If information and causation are two sides of the same coin, as IIT claims, there is no room for differences between these two concepts. However, the concept of information allows for qualifications that may be different from the ones employed in the conceptualization of causation. Such a problem hints at one of

ible cause-effect structure, but it does not explain what causes what, as IIT-inspired measures of causal strength claim (Albantakis et al., 2019). Assuming the above-mentioned IIT identity, one may ask as a follow-up question whether Φ is causing the specific mechanisms, structure, or constitution of the system, or, conversely, the latter cause the former. While pacifically accepting the IIT protocol to determine Φ, the reductionist will undoubtedly reject the first option and assume the second one.

Remarkably, in the face of frequent criticisms about the ambiguous meaning of information used by IIT (Pautz, 2019; Sánchez-Cañizares, 2022c), Tononi has always defended the univocal definition of information as it "has to be evaluated from the perspective of the system itself, starting from its elementary, indivisible components (…), and not by arbitrarily imposing 'units' from the perspective of an observer" (Tononi, 2008, p. 234). Why is that possible? Because "there will often be a privileged spatiotemporal 'grain size' at which a given system forms a complex of highest Φ—the spatiotemporal scale at which it 'exists' the most in terms of integrated information" (Tononi, 2008, p. 236). Reductionists may pretty well raise strong objections in front of that claim since (1) it remains contingent upon the very definitions of both information and information bearers in the system, and (2) it begs the question of why such 'grain size' deserves the qualification of 'privileged'.[5]

The relevance of this privileged spatiotemporal grain size postulated by IIT defenders becomes additionally compromised by the fact that, even though coarse-grainings for the emergence of the classical world from the quantum world are inescapable, they are not objectively and univocally determined (Gell-Mann & Hartle, 2007; Hartle, 2011; Sánchez-Cañizares, 2022b; Wallace, 2008). However, such apparent drawback points towards improving the metaphysical bedrock of IIT, which need not embrace that "each mechanism must have just one cause and one effect" (Grasso et al., 2021, sec. Causal structures). Why should one endorse that "irreducible mechanisms do not exclude each other, but causes and effects do"? (Grasso et al., 2021, sec. Causal structures). Why not then allow for the existence of different levels of causality in and through which the IIT's beloved composition and integration eventually occur? In the ontology market, one available option for IIT is the so-called nested hylomorphism, a neo-Aristotelian metaphysical approach[6] that assumes a many-to-many correspondence between causes and effects featuring complexity, as well as a gradation of matter-and-form composition in different, increasingly emergent levels.

All hylomorphisms agree that causal forms determine the way of being of each kind of thing, overcoming the implicitly Humean assumption that it is merely the arrangement of particles what determines the form of a system. In nested hylomorphism, formal causation at a higher level selects the specific dynamics—among the whole

the main points of this paper, expressed in its title: integrated information is not causation; the former is a consequence of an improved version of the latter, as this section aims to show.

[5] Information is not causation for the reductionist; causation is part and parcel of physical interactions. Even if the notion of information is secondary in (Grasso et al., 2021), maximal integrated information in IIT determines the maximal cause-effect structure, i.e., what exists the most. In other words, the causal analysis of (Grasso et al., 2021) crucially depends on IIT's take on information, particularly maximal integrated information, to access the maximal cause-effect structure.

[6] For the essentials of the Aristotelian tradition in contemporary philosophy of science, see (Owen, 2021a; Simpson, 2022; Simpson et al., 2018).

bunch of possibilities (matter) in the phase space of the immediate lower level—responsible for the extant unity and identity of the upper level. The form in the upper level is causally responsible, yet not uniquely, for the emergence of a new composite at the upper level. Of course, the specific dynamics of the system is its own actualized efficient cause, which efficiently causes the system to be, but not independently or in isolation from formal causation at each level (Sánchez-Cañizares, 2022a). Our argument focuses on the ability of nested hylomorphism to overcome reductionism on better terms than one-sided, causally-irreducible, integrated information.

If one takes IIT at face value, without metaphysical commitments, irreducibility might merely mean irreducibility of Φ in terms of information present at its subsystems. Nevertheless, inasmuch as one can explain away the emergence of such Φ through its underlying most basic mechanisms—on-off neurons wired as logical gates, no matter how many recursive internal loops may give rise to effectively nonlinear interactions—reductionists remain sound and safe, at least in their constitutive version of emergence (of Φ).[7] But, at least in the field of philosophy of mind, the IIT procedure has been accused of latent inconsistency as it runs into the problem of intrinsicality (Mørch, 2019; Sánchez-Cañizares, 2022c), namely, one could shrewdly add more connections to the initial system to increase Φ.[8] In other words, the domain of definition, or initial system, where to look for Φ is ill-defined in IIT without further assumptions. Nested hylomorphism adds, as crucial causal power underlying the whole IIT procedure, a formal cause at each level picking out the necessary, immediately lower-level dynamic that constitutes the system as an ontological unity.

More importantly for the argumentation against reductionism: (1) Such a selection entails a formal dimension of causality distinct from any efficient or material causality, as the phenomenon of multiple realizability in nature also witnesses. Each formal cause is irreducible because it is different. (2) Such a selection is immediately in charge of the ontological emergence of a new, higher level dependent on its lower levels, but irreducible to them. Hence formal causation co-cause together with other nested causes. (3) Such a selection rejects the reductionist's claim that the more complex thing can be analyzed into simpler parts, and its behavior synthesized from the behavior of those parts (Wilczek, 2015). Simpler parts and their potential behaviors do not suffice to reduce or explain the unity of those things that present more complexity. In practical terms, what defeats the reductionist stance is its remnant of indetermination because it merely deals with fundamental entities and interactions. The reductionist that is not allured by superdeterminism remains powerless in front of the question of why 'this' specific configuration or constitution as a system gets updated,

[7] Logical gates and logical states do not exist as such in nature. They must be instantiated by physical processes. The reductionist merely claims that the physical causality behind the scenes—that of fundamental particles plus fundamental interactions—underlies the IIT logical description. However, useful IIT might be, it is insufficient to ascertain natural causality.

[8] The intrinsicality problem exemplifies one of the problems in identifying integrated information with causation. Even though, for IIT practitioners, maximal integrated information would mark the physical scale at which the cause-effect structure exists the most, IIT still has a problem when defining the conditions for a candidate set to be chosen. In other words, additional causality seems necessary in order to explain why one may restrict the IIT searching procedure to a particular candidate set. Consequently, integrated information cannot be equated with all causation bearing on the system.

since the selection of dynamics is the activity of formal causation, providing a new level of determination and unity in nature truly irreducible to micro-causes. In other words, reductionists simply have no cogent metaphysical reasons to speak about individual systems in nature and, consequently, nothing to explain away.[9]

IIT anti-reductionists can still argue that their high-order composite mechanisms cause an effect. Indeed, a "quantitative, operational approach to causal analysis that can be applied across spatiotemporal scales can thus identify those macro levels that are particularly relevant for our understanding of a system. By contrast, a reductionist account cannot explain why some spatiotemporal scales seem causally more relevant than others" (Grasso et al., 2021, sec. Box 3: Macro and micro: causation at different levels of organization). That partly explains why Φ is a superb mark of the reasons to reject reductionism. But reductionists will not accept that IIT's causal structures are ontological as long as they can also obtain Φ through basic units and their specific interactions. IIT's causal analysis reveals nothing of the sort, especially when compared with a reductionist analysis in which the same results can hold. IIT manifests integrated information that can be maximal in some cases. But IIT begs the question of the relationship between causation and information by merely equating maximal integrated information with a cause-effect structure. Such an equation does not hold.

However, a hylomorphic co-causal combination accounts for the unity of higher-order mechanisms—and ultimately complex systems. IIT trusts everything to integrated information in higher-order mechanisms and complexes; however, the reductionist can obtain the same results without invoking higher-order causation. The problem really lies in that, to speak about a causal structure, one needs formal causation, as Φ alone does not explain the specific constitution of its underlying mechanisms. In nested-hylomorphic terms, the irreducibility that defeats reductionism cannot be deduced from a material and efficient causal account that merely invokes micro-causes, even if it comes to producing maximal integrated information. Only the irreducibility that stems from the system's ultimate ontological determination, provided by formal causation, does.

In other words, IIT practitioners have no account explaining what causes the integration of information. Integration of information depends on the model they build up (nodes and interactions). Of course, the question is what, in real life, instantiates such a designed model. The reductionist will answer that the model is instantiated by 'this' particular set of fundamental particles and their specific interactions. However, the reductionist cannot answer why this particular instantiation of fundamental particles and their interactions arises. Nested hylomorphism, considering systems and levels in nature as ontological and not just epistemic descriptions, defends that, at each level, formal causation selects or determines the specific physical interaction featuring in the system. This hylomorphic co-causation is responsible for the existence of the new level or system. Integrated information can be a marker for it but cannot definitely

[9] The reductionist must introduce additional information (usually in the form of initial conditions, boundary conditions, privileged bases, wave function reduction, symmetry breakings, semiclassical assumptions, and the like) to select the specific dynamics that natural processes realize: this is the critical point to attack the reductionist. Unfortunately, IIT-inspired causal analysis cannot provide such a perspective. It ultimately relies on the non-linear combination of physical processes in many-body systems that can be approximated and used as logical gates and logical states.

be its cause. Such a co-causation—physical and formal—is metaphysically different from the IIT's constellation of nth-order physical mechanisms that reductionism rejects. Nevertheless, IIT can acquire sounder philosophical foundations insofar as it acknowledges that what integrates information and selects specific dynamics is formal causation, not integrated information.

7 Conclusions

IIT-inspired measures of causality suffer from a metaphysical deficit, as pointed out in the title of this paper, namely, integrated information is not causation. Whereas IIT deserves credit in underscoring the relevance of integration thanks to its procedure to individuate Φ, postulating a mere identity between maximal integrated information and causation without further ado becomes a red herring when fighting against reductionism. As information, even if integrated, is not causation, IIT-inspired arguments against reductionism turn flawed since the reductionist may always argue that fundamental physical causality—provided by elementary units and the basic laws of nature—is the metaphysical bedrock of integrated information—the latter being just one effect of the former. Even if Φ is irreducible as information, its underlying nth-order mechanisms are not, remaining an epistemic construct of IIT. The recalcitrant reductionist will consider the supposedly irreducible information the mere effect of some non-linear dynamics and, consequently, not fundamentally irreducible.

Since IIT lacks the tools to describe the causal power that confers unity to the system, the reductionist will deem IIT-inspired derivation of causal structures hardly a proof of ontological irreducibility, begging the question. Admittedly, IIT offers a freshly-minted operational approach to integrated information but not to causation, nor to integrated causation. However, integration hints at the need for a deeper metaphysical grounding in the interpretation of the theory, like the one provided by nested hylomorphism. Here, formal causality becomes the keystone that selects a system's specific dynamic or constitution, conferring its ultimate determination. Whereas both IIT and reductionist causal accounts introduce by fiat the concrete dynamic that provides unity to a system, nested hylomorphism makes it contingent on the formal dimension that enables the ontological update of nature to a new higher level that nests its underlying lower levels.

Nested hylomorphism, as a metaphysical position on causation, does not oppose IIT. Whereas IIT stands as a superb mathematical protocol on the lookout for neural correlates of consciousness, it still suffers from shortcomings in its understanding of causation: if causation is equated with information, even if qualified as integrated information, IIT still falls prey to the reductionist argumentation. All IIT models could be reduced, as a matter of principle, to a perspective in which fundamental units and interactions determine the whole picture. What IIT fundamentally lacks is a metaphysical account of causation that, going beyond the causal structure of mechanisms, explains, especially, why some physical structures possess concrete dynamics of interactions that preserve their unity and identity. IIT may thus regard nested hylomorphism as a friend, not a foe.

Funding This paper received partial financial support from the project "New approaches to biological causality: ecological psychology, enactivism and teleodynamics" (NACB) of the Institute for Culture and Society (ICS).
Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Competing interests The author has no relevant financial or non-financial interests to disclose.

References

Albantakis, L., & Tononi, G. (2019). Causal Composition: Structural Differences Among Dynamically Equivalent Systems. *Entropy*, *21*(10), 989. https://doi.org/10.3390/e21100989.

Albantakis, L., Marshall, W., Hoel, E., & Tononi, G. (2019). What Caused What? A Quantitative Account of Actual Causation Using Dynamical Causal Networks. *Entropy*, *21*(5), 459. https://doi.org/10.3390/e21050459.

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2022). *Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms*. 1–53. http://arxiv.org/abs/2212.14787.

Arsiwalla, X. D., & Verschure, P. (2018). Measuring the Complexity of Consciousness. *Frontiers in Neuroscience*, *12*(JUN), 1–6. https://doi.org/10.3389/fnins.2018.00424.

Baars, B. J. (2005). Global Workspace Theory of Consciousness: Toward a Cognitive Neuroscience of Human Experience. *Progress in Brain Research*, *150*, 45–53. https://doi.org/10.1016/S0079-6123(05)50004-9.

Balduzzi, D., & Tononi, G. (2008). Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLoS Computational Biology*, *4*(6), e1000091. https://doi.org/10.1371/journal.pcbi.1000091.

Baxendale, M., & Mindt, G. (2018). Intervening on the Causal Exclusion Problem for Integrated Information Theory. *Minds and Machines*, *28*(2), 331–351. https://doi.org/10.1007/s11023-018-9456-7.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M. A., Laureys, S., Tononi, G., & Massimini, M. (2013). A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior. *Science Translational Medicine*, *5*(198), https://doi.org/10.1126/scitranslmed.3006294.

Cea, I. (2020). Integrated Information Theory of Consciousness is a Functionalist Emergentism. *Synthese*. https://doi.org/10.1007/s11229-020-02878-8.

Chalmers, D. J. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219. http://www.ingentaconnect.com/content/imp/jcs/1995/00000002/00000003/653.

Chis-Ciure, R. (2022). The Transcendental Deduction of Integrated Information Theory: Connecting The Axioms, Postulates, and Identity Through Categories. *Synthese*, *200*(3), 236. https://doi.org/10.1007/s11229-022-03704-z.

Colombo, M. A., Napolitani, M., Boly, M., Gosseries, O., Casarotto, S., Rosanova, M., Brichant, J. F., Boveroux, P., Rex, S., Laureys, S., Massimini, M., Chieregato, A., & Sarasso, S. (2019). The Spectral Exponent of the Resting EEG Indexes the Presence of Consciousness During Unresponsiveness Induced by Propofol, Xenon, and Ketamine. *Neuroimage*, *189*(September 2018), 631–644. https://doi.org/10.1016/j.neuroimage.2019.01.024.

de Ruiz, J., Soler, F., Iglesias-Parro, S., Ibáñez-Molina, A. J., Casali, A. G., Laureys, S., Massimini, M., Esteban, F. J., Navas, J., & Langa, J. A. (2019). Fractal Dimension Analysis of States of Consciousness and Unconsciousness Using Transcranial Magnetic Stimulation. *Computer Methods and Programs in Biomedicine*, *175*, 129–137. https://doi.org/10.1016/j.cmpb.2019.04.017.

Dehaene, S., & Changeux, J. P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018.

Gell-Mann, M., & Hartle, J. B. (2007). Quasiclassical Coarse Graining and Thermodynamic Entropy. *Physical Review A*, *76*(2), 022104. https://doi.org/10.1103/PhysRevA.76.022104.

Gillett, C. (2016). *Reduction and Emergence in Science and Philosophy*. Cambridge University Press. https://doi.org/10.1017/CBO9781139871716.

Golkowski, D., Larroque, S. K., Vanhaudenhuyse, A., Plenevaux, A., Boly, M., Di Perri, C., Ranft, A., Schneider, G., Laureys, S., Jordan, D., Bonhomme, V., & Ilg, R. (2019). Changes in Whole Brain Dynamics and Connectivity Patterns During Sevoflurane- and Propofol-induced Unconsciousness Identified by Functional Magnetic Resonance Imaging. *Anesthesiology*, *130*(6), 898–911. https://doi.org/10.1097/ALN.0000000000002704.

Grasso, M., Albantakis, L., Lang, J. P., & Tononi, G. (2021). Causal Reductionism and Causal Structures. *Nature Neuroscience*, *24*(10), 1348–1355. https://doi.org/10.1038/s41593-021-00911-8.

Hartle, J. B. (2011). The Quasiclassical Realms of this Quantum Universe. *Foundations of Physics*, *41*, 982–1006. https://doi.org/10.1007/s10701-010-9460-0.

Kim, J. (2010). *Philosophy of Mind* (3rd ed.). Westview Press.

Koch, C., & Tononi, G. (2011). A Test for Consciousness. *Scientific American*, *June*, 44–47.

Li, D., & Mashour, G. A. (2019). Cortical Dynamics During Psychedelic and Anesthetized States Induced by Ketamine. *Neuroimage*, *196*(April), 32–40. https://doi.org/10.1016/j.neuroimage.2019.03.076.

Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the Hard Problem of Consciousness Easier. *Science*, *372*(6545), 911–912. https://doi.org/10.1126/science.abj3259.

Mørch, H. H. (2019). Is Consciousness Intrinsic? A Problem for the Integrated Information Theory. *Journal of Consciousness Studies*, *26*(1–2), 133–162(30).

Moreno, Á., & Mossio, M. (2015). *Biological Autonomy. A Philosophical and Theoretical Enquiry*. Springer.

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

Owen, M. (2019). Exploring Common Ground between Integrated Information Theory and Aristotelian Metaphysics. *Journal of Consciousness Studies*, *26*(1–2), 163–187. https://www.ingentaconnect.com/contentone/imp/jcs/2019/00000026/f0020001/art00008.

Owen, M. (2021a). *Measuring the Immeasurable Mind: Where Contemporary Neuroscience Meets the Aristotelian Tradition*. Lexington Books.

Owen, M. (2021b). Circumnavigating the Causal Pairing Problem with Hylomorphism and the Integrated Information Theory of Consciousness. *Synthese*, *198*(S11), 2829–2851. https://doi.org/10.1007/s11229-019-02403-6.

Pal, D., Li, D., Dean, J. G., Brito, M. A., Liu, T., Fryzel, A. M., Hudetz, A. G., & Mashour, G. A. (2020). Level of Consciousness is Dissociable from Electroencephalographic Measures of Cortical Connectivity, Slow Oscillations, and Complexity. *Journal of Neuroscience*, *40*(3), 605–618. https://doi.org/10.1523/JNEUROSCI.1910-19.2019.

Pautz, A. (2019). What is the Integrated Information Theory of Consciousness? A Catalogue of Questions. *Journal of Consciousness Studies*, *26*(1–2), 188–215.

Sánchez-Cañizares, J. (2019). Classicality First: Why Zurek's Existential Interpretation of Quantum Mechanics Implies Copenhagen. *Foundations of Science*, *24*(2), 275–285. https://doi.org/10.1007/s10699-018-9574-y.

Sánchez-Cañizares, J. (2022a). Integrated Information Theory as Testing Ground for Causation: Why Nested Hylomorphism Overcomes Physicalism and Panpsychism. *Journal of Consciousness Studies*, *29*(1), 56–78. https://doi.org/10.53765/20512201.29.1.056.

Sánchez-Cañizares, J. (2022b). Teleology Writ Large: In Search of New Optimization Principles in Nature. In M. Fuller, D. Evers, A. Runehov, K. W. Saether, & B. Michollet (Eds.), *Studies in Science and Theology, volume 17 (2019–2020): Nature - and beyond: Immanence and Transcendence in Science and Religion* (pp. 327–343). Martin-Luther-University Halle-Wittenberg.

Sánchez-Cañizares, J. (2022c). Formal Causation in Integrated Information Theory: An Answer to the Intrinsicality Problem. *Foundations of Science*, *27*(1), 77–94. https://doi.org/10.1007/s10699-020-09775-w.

Sarasso, S., Brichant, J., Massimini, M., Gosseries, O., Laureys, S., Bodart, O., Comolatti, R., Casarotto, S., Tononi, G., Casali, A., Pigorini, A., Ledoux, D., Faria, G., Rosanova, M., Fecchio, M., Boly, M., & Nobili, L. (2019). A Fast and General Method to Empirically Estimate the Complexity of Brain Responses to Transcranial and Intracranial Stimulations. *Brain Stimulation*, *12*(5), https://doi.org/10.1016/j.brs.2019.05.013.

Simpson, W. M. R. (2022). From Quantum Physics to Classical Metaphysics. In W. M. R. Simpson, R. C. Koons, & J. Orr (Eds.), *Neo-Aristotelian Metaphysics and the Theology of Nature* (pp. 21–65). Routledge. https://doi.org/10.4324/9781003125860-3.

Simpson, W. M. R., Koons, R. C., & Teh, N. J. (Eds.). (2018). *Neo-Aristotelian Perspectives on Contemporary Science*. Routledge.

Stoljar, D. (2021). Physicalism. In *The Stanford Encyclopedia of Philosophy* (Summer). Edward N. Zalta. https://plato.stanford.edu/archives/sum2021/entries/physicalism/.

Tononi, G. (2008). Consciousness as Integrated Information: A Provisional Manifesto. *The Biological Bulletin*, *215*(3), 216–242. https://doi.org/10.2307/25470707.

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

Tononi, G. (2017). The Integrated Information Theory of Consciousness. In *The Blackwell Companion to Consciousness* (pp. 243–256). Wiley. https://doi.org/10.1002/9781119132363.ch17.

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.

Tononi, G., Albantakis, L., Boly, M., Cirelli, C., & Koch, C. (2022). Only What Exists Can Cause: An Intrinsic View of Free Will. 1–26. http://arxiv.org/abs/2206.02069.

van Fraassen, B. C. (2008). *Scientific Representation: Paradoxes of Perspective*. Clarendon Press.

van Riel, R., & Van Gulick, R. (2019). Scientific Reduction. In *The Stanford Encyclopedia of Philosophy* (Spring). Edward N. Zalta. https://plato.stanford.edu/archives/spr2019/entries/scientific-reduction/.

Wallace, D. (2008). Philosophy of Quantum Mechanics [Quantum Physics]. In D. Rickles (Ed.), *The Ashgate Companion to the New Philosophy of Physics* (Issue July, pp. 16–98). Ashgate. http://arxiv.org/abs/0712.0149.

Wilczek, F. (2015). *A Beautiful Question: Finding Nature's Deep Design*. Penguin Press.