Erkenntnis
https://doi.org/10.1007/s10670-022-00612-z

ORIGINAL RESEARCH

Emergentist Integrated Information Theory

Niccolò Negro[1]

Received: 17 September 2021 / Accepted: 21 August 2022
© The Author(s) 2022

Abstract
The integrated information theory (IIT) is an ambitious theory of consciousness that aims to provide both a neuroscientific and a metaphysical account of consciousness by identifying consciousness with integrated information. In the philosophical literature, IIT is often associated with a panpsychist worldview. In this paper, I show that IIT can be considered, instead, as a form of emergentism that is incompatible with panpsychism. First, I show that the panpsychist interpretation of IIT is based on two properties of integrated information: intrinsicality and fundamentality. I show that the way IIT deals with these two properties, however, aligns better with emergentism than panpsychism. Then, after plugging some anti-panpsychist assumptions into IIT's structure, I analyse different philosophical options for interpreting the formal result of causal emergence of integrated information in terms of dependence on and autonomy from its physical substrate. The upshot is that integrated information can be seen as dependent upon the fusion of the cause-effect powers of a physical substrate, and as autonomous in virtue of global-to-local determination. According to this interpretation, consciousness is the constraining power of the system as a whole upon itself, when this power emerges from the fusion on the cause-effect powers of the system's components.

1 Introduction

The integrated information theory (IIT) is one of the most influential and debated neuroscientific theories of consciousness (for reviews of the state of the field, see Yaron et al., 2022; Seth & Bayne, 2022). IIT is however an ambitious research programme that does not just aim to explain how consciousness relates to human brains,

✉ Niccolò Negro
niccolo.negro@monash.edu

[1] Cognition & Philosophy Lab, Department of Philosophy, Monash University, 20 Chancellors walk, 3800 Clayton, VIC, Australia

Published online: 14 October 2022

but also seeks to define what consciousness *is*. Thus, IIT comes with important ontological and metaphysical implications. Such implications usually align IIT with panpsychism: since the main claim of IIT is that consciousness is integrated information and integrated information is an intrinsic and fundamental property of reality, then consciousness is an intrinsic and fundamental property of reality (Tononi & Koch, 2015).

In this paper, I question this panpsychist interpretation of IIT (which I call 'Panpsychist IIT'), and I suggest that there is a better way to understand the intrinsic and fundamental nature of integrated information. The metaphysical interpretation of IIT that I develop here is a form of emergentism, and I will call it 'Emergentist IIT'. The upshot of Emergentist IIT is that consciousness, as integrated information, emerges from the fusion of the causal powers of a physical system, and it is causally autonomous by exerting global-to-local constraining on the system's components.

The main goal of this paper is to show that panpsychism is not necessarily implicated by IIT. This means that consciousness researchers that are reluctant to accept IIT because of its panpsychist implications can rethink their stance towards the theory and perhaps assess it under a different light.

In the first section, I briefly present IIT. In the second section, I explain why it has been considered as a panpsychist theory, but I show that the two properties upon which Panpsychist IIT is built, namely the intrinsicality and the fundamentality of integrated information, are expressed in a way that is at odds with what panpsychists typically claim. In the third section, then, I plug some anti-panpsychist assumptions into the IIT's structure, and argue that if these assumptions lead us to a coherent picture of IIT, then we have a compelling view that provides an alternative to Panpsychist IIT. In the fourth section, I consider the formal result of causal emergence, which can be used to support Emergentist IIT. In section five, I claim that such a result is metaphysically relevant, and not only epistemologically relevant. In section six, then, I try to make sense of causal emergence from a metaphysical viewpoint, under the assumption that an emergent phenomenon must be both dependent upon, and autonomous from, its basis. Given specific IIT-driven guidelines, I conclude that integrated information can be seen as an emergent phenomenon, and, in section seven, I draw the overall picture of Emergentist IIT. In the conclusion, I point out that Emergentist IIT can be a viable and attractive alternative to Panpsychist IIT, as it seems to satisfy the properties of intrinsicality and fundamentality of integrated information better than Panpsychist IIT does.

2 Integrated Information Theory: A Brief Outline

Here, I will provide a brief summary of IIT, to highlight just the aspects that are necessary for present purposes. Detailed introductions of the theory can be found in (Oizumi et al., 2014) and (Tononi et al., 2016). IIT is a theory of phenomenal consciousness – that is, it aims to explain the "what it is like" to be a conscious subject (Nagel, 1974; Block, 1995). The starting point of the theory is given by IIT's *axioms*, which are supposed to pick out the essential features of subjective experience from the first-person perspective. These essential features are then mapped onto the physi-

cal world courtesy of IIT's *postulates*. Each postulate is supposed to explain how physical reality should be in order for us to have the phenomenal features picked out by the corresponding axiom. Given that the postulates are formulated in information-theoretic terms, IIT arrives at an information-theoretic measure that is supposed to capture exactly the phenomenon from which the theory started, namely consciousness itself: this is integrated information, symbolized by Φ^{Max}.

A crucial aspect of IIT is that the IIT notion of information differs from the traditional notion of information (Shannon, 1948). IIT informational language is based instead on an intrinsic notion of information (Barbosa et al., 2020). While Shannon information is a measure of a *correlation* between variables, IIT information is supposed to measure *causal* relations. Thus, according to IIT, information is equivalent to causation, (Oizumi et al., 2014, p. 24), and causation, interpreted as a form of difference-making (Lewis, 1973; Woodward, 2003), can be measured in information-theoretic terms by using IIT formal apparatus (Albantakis et al., 2019). The causation that matters for consciousness, however, is *intrinsic causation*, where the *relata* of the causal/informational link are two states, at two different time-steps, of the same system. IIT information is not information for an extrinsic observer, but rather information for the system that is generating it: intrinsic information is information from the point of view of the system itself. This intrinsic notion of information is mandated by the very first axiom of IIT, which states that consciousness "exists from its own *intrinsic perspective*, independent of external observers" (Tononi & Koch, 2015, p. 7), and therefore it cannot be captured by an observer-relative measure. Describing consciousness in terms of integrated information, under the equivalence of information with intrinsic causation, thus means identifying consciousness with the intrinsic causal structure of a system in a state (Oizumi et al., 2014, p. 14).

The result is a theory of consciousness that conceptually defines what consciousness *is* (i.e. integrated information), and provides a formal tool to measure it (Φ^{Max}). On the one hand, the formalization of consciousness comes with explanatory and predictive power. On the other, the conceptualization of consciousness as integrated information comes with important metaphysical implications. Here, my focus is not on IIT as a scientific theory, but rather on its metaphysical implications: my aim is to provide a philosophical analysis of how IIT fits in the debate about the nature of consciousness.

3 Why IIT is not a form of Panpsychism

Two of the main proponents of IIT, Giulio Tononi and Christof Koch, frame IIT as a form of panpsychism: "in line with the central intuitions of panpsychism, IIT treats consciousness as an intrinsic, fundamental property of reality" (Tononi & Koch, 2015, p. 11). However, it is not clear that this metaphysical interpretation of IIT fits with the explanation IIT itself gives of the fundamentality and intrinsicality of integrated information, the two central properties upon which the panpsychist interpretation of IIT is built.

With respect to intrinsicality, consciousness as integrated information is said to be an intrinsic property because it "can be accounted for by the intrinsic cause–effect

power of certain mechanisms in a state—how they give form to the space of possibilities in their past and their future." (Tononi & Koch, 2015, p. 11). As mentioned above, this depends on the fact that integrated information is a measure of intrinsic causation – a measure of the cause-effect powers that a system can exert upon itself. Integrated information is thus instantiated any time a physical system satisfies IIT's postulates for physical existence, since it is an *intrinsic* property of that system. And given that consciousness just is integrated information, and some degree of Φ^{Max} can potentially be found in simple systems like atoms and molecules, consciousness is potentially ubiquitous and distributed anywhere physical reality satisfies IIT's postulates. This is why the intrinsicality of integrated information suggests a panpsychist reading of IIT. However, Tononi and Koch's rendition of the intrinsicality of integrated information does not correspond to how panpsychists traditionally conceive of the intrinsicality of consciousness[1].

A common way for (contemporary) panpsychists to interpret consciousness as an intrinsic property of reality is to hold that consciousness is a *categorical* property of matter: a non-relational property that displays an essential aspect of its bearer (Goff, 2017). For example, the shape and the mass of an object are categorical properties of that object, as they constitute how the object *is*. According to some ('Russellian' – see (Russell, 1927)) versions of panpsychism, physical properties of reality can account for what reality does, but not for what it is: the essential nature of reality is ultimately given by phenomenal properties. In this context, consciousness is intrinsic in virtue of constituting the intrinsic essence of physical reality. Thus, physical reality would be grounded on phenomenality, and not vice versa (Mørch, 2019b).

As pointed out by Grasso (2019), this categoricalist ontology is at odds with IIT, since IIT explicitly endorses (i) dispositional essentialism, the view that the essence of an entity is given by how that entity is disposed to affect and be affected by other entities (its cause-effect powers) (for a discussion, see (Bird, 2012), cited in (Tononi, 2017)); and (ii) pandispositionalism, according to which all properties are dispositional properties. The notion of 'intrinsic property', in IIT, must then be seen under this lens: consciousness as integrated information would be an intrinsic property of reality because it is identical to the intrinsic cause-effect powers of a physical system.

This ontology is not, strictly speaking, incompatible with panpsychism. Mørch (2020; Mørch, 2019a) has laid the basis to defend a panpsychist-flavoured version of IIT compatible with pandispositionalism, by maintaining that causation itself is grounded on phenomenality: "the only fundamentally dispositional properties we know or can positively conceive of are phenomenal properties—in particular, phenomenal properties associated with agency, intention and/or motivation" (Mørch, 2020, p. 1074). Although this version of IIT might be plausible and internally coherent, it requires endorsing the view that causation itself is ontologically grounded on phenomenal properties, a claim not explicitly endorsed by IIT. Rather, causal powers seem to be the criteria defining the physical world that is supposed to exist inde-

[1] Tononi & Koch (2015, p. 11) admit that IIT's panpsychism differs from traditional panpsychism insofar as it does not maintain that everything is conscious, but only that Φ^{Max}-generating systems are. However, given that some degree of Φ^{Max} can be found in simple systems like atoms and molecules, IIT implies that such systems are conscious, which is a claim that coheres better with panpsychism than other metaphysical views.

pendently of any subject or observer. Hence, reconciling IIT's pandispositionalist ontology with panpsychism seem to require a substantial modification of IIT. When panpsychists and IIT proponents claim that consciousness is an intrinsic property, they seem to refer to ultimately different ontologies: panpsychists refer to causal powers as grounded on (or being forms of) phenomenality, while IIT proponents refer to phenomenality as grounded on observer-independent causal powers[2].

This relates the intrinsicality of consciousness to its fundamentality. Perhaps, IIT and panpsychism could share the view that physical reality is imbued with phenomenal properties from the very bottom. According to this interpretation, consciousness would be an intrinsic property of reality because integrated information is an intrinsic property of microphysical elementary particles like muons, electrons and neutrinos. To be intrinsic in this sense, consciousness must also be *fundamental* (i.e., a property of the fundamental level of reality).

According to Tononi & Koch (2015), consciousness as integrated information is fundamental, but "in the case of experience the entities having the property are not elementary particles but complexes of elements" (Tononi & Koch, 2015, p. 11).

This sense of fundamentality clashes with the panpsychist view that consciousness or is a property of fundamental reality along with properties like mass and charge – for a discussion, see (Chalmers, 2016). IIT's sense of fundamentality is instead a dependence-based notion of fundamentality (Bennett, 2017; Leuenberger, 2020), since the instantiation of the property of integrated information essentially depends on the relations between the components of the physical substrate of consciousness – the 'complex', in IIT jargon (Tononi et al., 2016).

In metaphysics, philosophers (and scientists alike, see (Ellis et al., 2012)) have appealed to the notion of *strong emergence* to resolve the apparent tension between a property being both *fundamental* and *dependent* upon the instantiation of other properties. This is the idea that an entity can depend on other entities, and yet display fundamentally novel properties or causal powers, that is, properties or powers that cannot be reduced, not even in principle, to the properties or powers of the reality upon which they depend (O'Connor & Wong, 2005; Wilson, 2015; Chalmers, 2006).

Given IIT's pandispositionalism, according to which existence is defined in terms of cause-effect powers, the novelty of the emergent phenomenon must be accounted for in terms of cause-effect powers: integrated information, in being a property of a physical complex, rather than a property of elementary physical elements, comes into existence when a complex of elements generates causal powers that are irreducible to those of the component elements. Then, in the context of IIT, consciousness seems to be fundamental in the strong emergentist sense, not in the panpsychist sense (Cea, 2020).

Reconciling the panpsychist's claims about the intrinsicality and fundamentality of consciousness with how these two concepts are elaborated in IIT seems to be a difficult, if not impossible, task. Here, I argue that the intrinsicality and fundamentality of consciousness, in IIT, are better understood under an emergentist lens.

[2] I am using "grounding" as an ontological notion, here. There is no doubt that IIT assigns an epistemological priority to consciousness, and therefore even causation is *known* from within consciousness. But this does not imply that consciousness is ontologically prior to causation.

Although in the philosophical literature emergentism and panpsychism are often opposed (Van Cleve, 1990), several attempts to cash out panpsychism in an emergentist fashion have been made (Seager, 2012; Mørch, 2019a; Brüntrup, 2016). However, emergentism and panpsychism disagree on how consciousness is distributed. In what follows, I am going to sketch a picture of IIT that highlights precisely this distinction: the version of IIT I will present ('Emergentist IIT') is fully compatible with emergentism, but it will not be compatible with panpsychist emergentism since there is no phenomenality at the elementary level of reality. The core claim of Emergentist IIT is that consciousness is an emergent phenomenon that depends on, but is irreducible to, the cause-effect powers of a non-phenomenal system, when its cause-effect powers are interconnected in a way that satisfies IIT's postulates.

To show that Emergentist IIT is compatible with IIT, but incompatible with Panpsychist IIT, I will plug some 'anti-panpsychist' assumptions into IIT's theoretical structure and ontology, and assess whether the result is a coherent and plausible narrative about the place of consciousness in nature.

4 Anti-panpsychist Assumptions of Emergentist IIT

The claim that Emergentist IIT can be an alternative to Panpsychist IIT depends on two main *desiderata*. The first desideratum is to take IIT seriously as a scientific theory of consciousness, and this comes with two IIT-driven assumptions. The second desideratum is to paint an anti-panpsychist picture.

First desideratum: our conceptual work on the metaphysical implications of IIT should take the theoretical standpoints and formal results of IIT as a starting point, and find them a place in the metaphysical landscape while keeping the core of IIT intact. This assures that Emergentist IIT is a viable and faithful interpretation of IIT proper. We need to consider two central IIT-driven assumptions: the first concerns the above-mentioned pandispositionalist and dispositionalist essentialist ontology of IIT. To restate, this implies the view that to exist is to have causal powers (Grasso, 2019; Tononi, 2017). Emergentist IIT, in being a metaphysical interpretation of IIT, must be compatible with this ontology: the existence of consciousness as an emergent phenomenon must be identified with causal powers that are fundamentally novel and irreducible to those of its physical basis (i.e. the complex).

The second IIT-driven assumption implied by the first desideratum is the central thesis that consciousness is *explanatorily* identical (Haun & Tononi, 2019, p. 5) to integrated information. Emergentist IIT must assume that integrated information is somehow informative with respect to consciousness: we can explain consciousness in terms of integrated information, and use integrated information to infer and predict states of consciousness. This does not mean positing a numerical identity between consciousness and integrated information. It means instead that integrated information can be used to operationalize consciousness, and make it amenable to scientific investigation (for a discussion on implications of the notion of 'identity' in IIT, see (Mediano et al., 2019), (Mediano et al., 2022) and (Michel & Lau, 2020)). Importantly, according to IIT's Exclusion postulate, only a maximum of integrated information across nested systems contributes to consciousness (Oizumi et al., 2014,

p. 3). This postulate guarantees that one *definite* integrated information structure corresponds to consciousness, since our own phenomenology is not only unitary (as per Integration axiom), but also definite in space and time. Emergentist IIT will thus assume that consciousness can be achieved only at an optimal spatiotemporal scale: it is not just a matter of whether integrated information occurs, but of whether a *maximum* of integrated information is achieved from the intrinsic perspective of the system (Hoel et al., 2016; Moon, 2019)[3].

Second desideratum: Emergentist IIT must be incompatible with panpsychist emergentism. This means that the microphysical level cannot be itself phenomenal but, rather, mentality and consciousness should depend on reality that is in itself non-mental (Montero & Papineau, 2005). Notice that this requirement is restrictive enough to exclude several forms of panpsychism, but is compatible both with minimal physicalism (Lewis, 1983), property dualism (Chalmers, 1996), and versions of neutral monism, according to which fundamental reality is neither physical nor mental (Russell, 1921; Coleman, 2014; Banks, 2010). In what follows, I will thus use 'microphysical' meaning 'non-mental'.

We have some ground to build Emergentist IIT: the idea is to take IIT seriously as a theory of consciousness, and see whether these anti-panpsychist assumptions can make sense of the intrinsicality and the fundamentality of integrated information in a coherent way. If we can do so, it means that we have an alternative way to understand the two properties that, according to Tononi & Koch (2015), should support Panpsychist IIT. I will now turn to a formal result that can support Emergentist IIT.

5 Macro and Micro: Emergence as a Formal Result

Work led by Erik Hoel (Hoel et al., 2013, 2016 – see also (Grasso et al., 2021)) demonstrates *how* a system can achieve maximal integrated information at the macro-level. The rationale for this work is to be found in the exclusion postulate, according to which only the systems that specify a maximum of integrated information, among nested systems, contribute to consciousness. To identify this optimal grain for maximal integrated information, Hoel and colleagues have applied IIT analysis to physical systems in a state, and calculated Φ^{Max} at the micro-level and at the macro-level. For example, a micro-level system can be seen as composed of six interconnected binary elements ABCDEF. The macro-level could be given by taking each couple of elements as a single unit. The macro-level is thus defined via the mapping of different micro components into a single macro compound-element (say, α=AB; β=CD; γ=EF; see (Hoel et al., 2016, p. 8)). From a formal point of view, such a mapping can

[3] It could be asked whether the exclusion postulate is not in itself sufficient to prevent panpsychism. As noted above, there is a difference between traditional panpsychism, which assigns consciousness to *everything*, and the version of panpsychism implicated by IIT, which attributes consciousness to only Φ^{Max}-generating systems. Although it is true that the exclusion postulate could be used to prevent traditional panpsychism, it is not sufficient to prevent the conclusion that simple systems like atoms and molecules might be conscious.

take the form of coarse-graining or black-boxing (Marshall et al., 2018), but for space reasons I will focus only on the coarse-graining approach here.

The IIT analysis is based on a conception of causation that is interventionist and manipulationist in nature (Woodward, 2003; Pearl, 2000; Menzies & Price, 1993). That is, causation is a difference-making relation that can be analysed in terms of how a change of a variable's value changes another variable's value, when all the other variables are fixed (Woodward, 2003, p. 59). Importantly, this approach to causal analysis is not in itself incompatible with the above-mentioned alignment of IIT with a power-based view of causation. In fact, a power-based approach to causation tackles the metaphysical question of what causation is, whereas the interventionist approach tackles the epistemological question of how to track causation. And addressing this latter question is precisely the scope of IIT's causal analysis.

In particular, IIT formal analysis requires partitioning the system of interest in all the possible ways and measure how the current state of a particular part of the system makes a difference to the other parts and to the system as a whole. As a result, we obtain a transition probability matrix (TPM) that tracks the causal structure of the system under analysis – that is, it represents how the states of the system's components constrain other components' states. Hoel and colleagues apply this type of analysis both at the micro-level (e.g. the ABCDEF system) and at the macro-level (e.g. the αβγ system).

The result of this formal analysis is that Φ^{Max} (i.e., maximal integrated information) is achieved when the system is analysed at the macro-level. Given IIT's interventionist approach to causation, the resulting idea is that the macro-level is the true difference-maker: the causal power of the system occurs at the macro-level, not at the micro-level. Hoel and colleagues explain that "macro-level mechanisms can achieve greater irreducible selectivity. This means that the macro can win if macro mechanisms constrain past and future macro states irreducibly – above and beyond their parts – to a far greater extent than micro mechanisms do" (Hoel et al., 2016, p. 11). Claiming that the macro-level is more irreducibly selective than the micro-level amounts to claiming that the state transition of the system is more dependent on its macro-level states than its micro-level states, because the macro-level makes a difference to the system that goes beyond the difference made by the micro-level. To explain how this is possible, Hoel and colleagues argue that the coarse-grained elements are less noisy and more robust than the micro elements, and the gain in causal information of the macro-level would be constituted precisely by the fact that irrelevant micro-level details can be screened-off.

This point can be seen by resorting to Yablo's pigeon example (Yablo, 1992). In this case, a pigeon is trained to peck at red objects. The casual description of the pigeon's behaviour based on fine-grained properties (i.e. saying that the pigeon pecks because the object is magenta or crimson) does not seem to carry enough information, as it is too specific: saying that the pigeon pecked because the object in front of it is magenta omits the information that the pigeon would have pecked even if the object had been crimson (or carmine, or any other shade of red). Instead, the description of the pigeon's behaviour based on a coarse-grained property (i.e., the pigeon pecked because the object in front of it is red) seems to be more appropriate

(or, following Yablo, more *proportionate*). In Yablo's case, the relevant distinction is between determinable (e.g., red) and determinate (e.g., magenta) properties.

This discussion can be applied to Hoel et al.'s work by noticing how the determinable property "red" is doing a similar job that macro-level states do in Hoel et al.'s framework, while the determinate properties "magenta" and "crimson" are comparable to fine-grained states. In Hoel et al.'s terminology, the robustness of the determinable 'red' would be comparable to a coarse-grained state that summarizes micro-level information mainly by screening-off irrelevant details. According to IIT, the macro-level elements of a system can be the true cause; that is, the true difference-makers, of the system's state transition in the same way as the coarse-grained property 'red' is the true difference-maker of the pigeon's behaviour. IIT proponents have called this phenomenon "causal emergence" (Hoel et al., 2016, p. 1).

IIT thus provides us with a definition of consciousness in terms of integrated information, and with a formal result that demonstrates that the optimal spatiotemporal grain at which consciousness emerges is not at the micro-level, but rather at the macro – emergent – level. My scope, here, is to take the formal result of causal emergence and to pair it with philosophical analysis, so as to provide a conceptually coherent view of how consciousness can be seen as an emergent phenomenon, within the IIT context.

The first problem is whether a metaphysically robust reading of causal emergence is justified: perhaps we can successfully *describe* the causal profile of a system by analysing it at the macro-level, rather than the micro-level, but this does not mean that the macro-level is more causally efficacious. Causal emergence could be a matter of description, not of being. I will now argue that Emergentist IIT will have to push back on this "epistemic" interpretation of causal emergence.

6 The Metaphysical Robustness of Causal Emergence

Emergentist IIT is a metaphysical interpretation of IIT. This means that it takes the formal results of IIT and it translates them into the debate about the place of consciousness in nature. If causal emergence were just an observer-dependent representation of a system, causal emergence would only be an epistemologically relevant result, not a metaphysically relevant one. So, we could not use it to justify a *metaphysical* reading of IIT that depends on it. For Emergentist IIT to justifiably use causal emergence to promote an emergentist metaphysical picture of IIT, we need to make sure that causal emergence is a metaphysically relevant result.

Dewhurst (2021) argues that there is no reason to think that a system of interest gains genuinely novel causal powers at the emergent macro-level, since causal emergence just refers to the best scale at which a system should be described. Dewhurst appeals to the traditional labels of 'strong' and 'weak' emergence to make his point: causal emergence could be a form of 'weak' emergence, but not of 'strong' emergence. Following Chalmers (2006), Dewhurst maintains that weak emergence occurs when a phenomenon is only epistemologically irreducible to its basis, but not ontologically: the irreducibility of the weakly emergent phenomenon would be due to our inability to explain the higher-level phenomenon in terms of the lower-level phe-

nomena upon which it depends. Although the strong/weak distinction can be formulated with more nuanced modal and metaphysical arguments – for a discussion, see (Wilson, 2015) [4] – what matters for present purposes is the question of whether causal emergence is irreducible only 'in the eye of the beholder'. We can thus frame this problem in terms of the observer-relative nature of causal emergence: if causal emergence is observer-dependent, then it is not a metaphysically relevant result and it is weakly emergent, whereas if it is observer-independent, then it is a metaphysically relevant result and it is strongly emergent.

The question, then, is how to justify the metaphysical robustness of causal emergence. The problem, as Dewhurst highlights, seems to be rooted in the *tool* IIT employs to measure causation, namely the interventionist approach itself: even if we concede that it is possible to track macro-level causation, it is not at all clear that this causation is observer-independent.

Indeed, it seems to be the observer who intervenes on and perturbs the system, and averages over the system's components to derive a macro-component: the notion of causal emergence, according to Dewhurst, is essentially observer-dependent, as the very distinction between a micro-level and macro-level system is a *representation* of the system imposed by the observer.

This objection, however, seems to misinterpret the notion of intrinsic information in IIT on the one hand, and the very scope of interventionism on the other hand. First, in IIT, perturbations, interventions, and other observer-relative notions are necessary to describe, from the outside, a system; Φ^{Max} can indeed be the result of the extrinsic analysis an observer performs over a system of interest, but, and this is the crucial point, this extrinsic search is supposed to pick out a phenomenon that exists in its own right, independently of that description (Tononi, 2008, p. 234). And we have good reasons to think that the extrinsic search corresponds to the intrinsic phenomenon of interest, namely consciousness, because it is axiomatically built upon how consciousness *intrinsically is*. This does not mean that IIT's extrinsic analysis of consciousness – based on observer-relative notions like interventions, perturbations – is accurate in describing consciousness: the axioms might be revisable (Bayne, 2018), as is their translation into postulates (Hanson & Walker, 2021). But what is important here is the Janus-face nature of IIT's causal analysis: the extrinsic aspect, given by the procedure an observer must perform over a system, is supposed to have an intrinsic counterpart – how the system *is* from its own intrinsic perspective.

In the context of causal emergence, this means that the emergence of macro-level causation individuated by IIT's causal analysis reveals a genuine phenomenon: IIT's formal analysis provides an *extrinsic description* of how that phenomenon *intrinsically is*.

This is in line with the scope of the interventionist framework upon which IIT causal analysis is built. As Woodward puts it, even if we concede that "what matters

[4] For example, Van Cleve (1990) draws the weak/strong distinction in terms of whether a property supervenes with purely nomological (for weak emergence) or logical (for strong emergence) necessity on its base properties. Wilson (1999; 2015) draws the distinction, instead, in terms of token powers: strong emergence occurs when a higher-order property has at least one token power that is not identical with any token power of the base property, whereas weak emergence occurs when the higher-order property has a proper subset of the token powers of the base property (see Wilson 2015, p. 362).

for whether X causes Y is the 'intrinsic' character of the X-Y relationship […] the attractiveness of an intervention is precisely that it provides an extrinsic way of picking out or specifying this intrinsic feature" (Woodward, 2000, p. 204).

Thus, the interventionist framework maintains that our extrinsic way of tracking causation faithfully represents an intrinsic causal phenomenon. When applied to IIT, this attitude is further strengthened by the fact that we know from IIT axioms that there is an intrinsic perspective, and that this intrinsic perspective *exists* independently of any extrinsic, observer-relative, analysis. Thus, if one accepts IIT and interventionism, the metaphysical robustness of causal emergence follows: consciousness corresponds to the optimal spatiotemporal scale at which integrated information reaches its maximum. This spatiotemporal scale is captured by causal emergence, and therefore, given that consciousness is clearly not an observer-relative phenomenon, causal emergence itself cannot be an observer-relative phenomenon. As IIT proponents put it:

> The search for a maximum of integrated information (Φ^{Max}) thus identifies a definite spatiotemporal scale at which a set of elements "self-defines" as a complex—the grain size at which it "comes into focus" causally from its own intrinsic perspective. Such a grain size is determined by the intrinsic cause–effect structure of the system itself, as opposed to being the most convenient or interesting scale for an external observer (Hoel et al., 2016, p. 11).

The only option available to push against the metaphysical robustness of causal emergence (from the perspective of someone who accepts IIT), seems to be to reject interventionism. But given the widespread use and popularity of interventionism in science and philosophy, it seems reasonable to use this framework as an epistemological account of causality: we have at least good reasons to think that the interventionist-flavoured IIT's causal analysis is able to faithfully pick out the intrinsic phenomenon of interest, namely consciousness, as the "coming into focus causally" (Hoel et al., 2016, p. 11) of a system, from its own intrinsic perspective.

The next step in building Emergentist IIT is to understand the causal emergence of consciousness in metaphysical terms. Rigorous philosophical analysis is needed in order to understand the relation between consciousness as integrated information and its basis. The scope of the next section is to provide such an analysis.

7 Emergence as Dependence and Autonomy

In order to give a metaphysically precise account of causal emergence, we need to frame integrated information as an emergent phenomenon. In the philosophical literature, emergent phenomena are usually characterized by two necessary and jointly sufficient conditions: they are (i) *dependent upon* their basis; and (ii) *autonomous from* (or *irreducible to*) their basis (O'Connor, 1994; Wilson, 2015). That is, integrated information, to be an emergent phenomenon as causal emergence shows, must be both dependent upon the causal properties of its basis, and autonomous from such causal properties. But there are several different ways of interpreting these two condi-

tions. Loosely following (O'Connor, 2020) and (Wilson, 2015), we can individuate four main ways philosophers have used to express the dependence of the macro on the micro[5]: (i) functional realization; (ii) causation; (iii) supervenience; (iv) fusion. Similarly, there are mainly four ways for interpreting the autonomy of the macro from the micro: (i) non-linearity; (ii) multiple realizability; (iii) fundamentality; (iv) downward causation. The goal, here, is to understand which of these possibilities fits best with IIT; that is, with how causal emergence expresses the relation between integrated information and its basis.

Before embarking in this project, it is important to restate that, according to IIT, integrated information is equivalent to causation, since integrated information is a form of irreducible difference-making. This means that integrated information is an emergent phenomenon not in virtue of *having* novel causal properties, but rather in virtue of *being* a novel causal property. The equivalence IIT draws between integrated information and causation must constrain the way we conceptualize integrated information as an emergent phenomenon: the basis itself is given by the *cause-effect powers* of the complex, and not by its physical architecture as such.

Courtesy of this important clarification, we can now try to make sense of the nature of integrated information as an emergent phenomenon.

7.1 Dependence Relation

We need to identify some guidelines along which we can adjudicate whether a certain dependence relation is good enough to fit with IIT. First, the dependence relation must account for the *spatiotemporal* emergence of integrated information (Hoel et al., 2016, p. 11). Second, the dependence relation must account for the identity of integrated information with a physical, rather than functional or mathematical, property (Koch, 2019). There is an apparent tension between the idea that integrated information is a physical property and, at the same time, an emergent property, especially if the metaphysical robustness of causal emergence is interpreted as a form of strong emergence. The dependence relation must be able to resolve this tension. Third, the dependence relation must be illuminating; meaning, it must not create more problems than it solves: it must provide a way to understand reasonably well how the causal powers of the complex relate to integrated information. Let us see which metaphysical dependence relation complies best with these guidelines.

I. *Functional Realization*: the macro might depend on the micro in virtue of being *realized* by it. For example, a line of code can be considered as a macro-level phenomenon dependent on its micro-level implementation on a hardware. To be a line of code is to perform a certain role at the software level, and such a role is dependent on its physical realization at the hardware level (Baysan, 2019). In this sense, the macro should be functionally defined first. This does not seem to be the case in IIT, since integrated information is not defined in terms of functional role, but rather in terms of structural properties (Ellia et al., 2021): as the second

[5] In what follows, I will be using the labels 'macro' and 'micro' respectively for emergent phenomena and their bases. These labels have to be understood as ontological, rather than representational.

guideline points out, consciousness does not coincide with an abstract property, but it coincides instead with a physical property – with a system's actual physical architecture (Tononi, 2015). Thus, considering realization as the dependence relation of integrated information on its basis does not seem to capture the nature of causal emergence.

II. *Causation*: the macro could depend on the micro in virtue of being *caused* by it. (O'Connor & Wong, 2005). Whether causation between levels is a coherent concept is matter of debate (Craver & Bechtel, 2013). Here, the main problem with interpreting the relation between integrated information and its basis as *causal* is that it is not particularly illuminating. The two *relata* of this causal relation would themselves be causal properties, and therefore, we would end up with an account of macro-level causation (integrated information) as caused by micro-level causation (the causal powers of the complex), and it is not at all clear that causation can be caused. This interpretation does not help much in our conceptualization of integrated information as an emergent phenomenon, and therefore cannot be considered as illuminating. Although interpreting the dependence relation between integrated information and its basis as causal relation is not, strictly speaking, incompatible with IIT, such an interpretation seems to create more problems than it solves.

III. *Supervenience*: the macro-level can be said to depend on the micro-level in virtue of being *supervenient* on it (Davidson, 1970). For example, the aesthetic properties of a painting (e.g. "being beautiful") supervene on its non-aesthetic properties (e.g. "this spot being blue", "that spot being red", and so on). The idea is that every macro-level change necessarily implies a micro-level change, while micro-level change can occur without any change at the macro-level. So, the macro-level is not necessarily reducible to its micro basis. To change the painting's macro-level property from "being beautiful" to "being awful", we need to change at least some of its micro-level properties. But the macro-level property "being beautiful" is not reducible to a particular set of micro-level properties, as there are innumerable ways for a painting to "be beautiful" (Kim, 1990). Supervenience is explicitly mentioned in (Hoel et al., 2016, p. 4) as relation between the macro and the micro. There are two problems with this relation, though. First, supervenience is a synchronic relation between the macro and the micro, and this is at odds with the idea that a set of elements "self-defines" as a complex both in space *and time* (Hoel et al., 2016, p. 11). We can perhaps consider several time steps as one coarse-grained time step, and consider the relation between the micro time steps and the macro time step as synchronic supervenience, but only from the extrinsic perspective – i.e. when we represent the system through coarse-graining. From the intrinsic perspective of the system, however, causal emergence takes time. And this temporal aspect of emergence is not captured by supervenience itself, since supervenience is an instantaneous relation. Second, supervenience usually implies the co-existence, at the same time, of the micro level property and the macro level property. If this were the case in IIT, integrated information would co-exist with the causal powers of its basis, which is incompatible (given that to exist is to have causal powers) with the idea that causal emergence renders inert the causal powers of the micro level (Hoel et al., 2016,

p. 10). Thus, supervenience does not seem able to capture all the aspects of causal emergence as a metaphysically robust phenomenon.

IV. *Fusion*: the macro-level might depend on the micro-level in virtue of being the result of a physical fusion between two micro-level property instances in interaction. The fused macro-level property is a whole whose causal powers inherit and subsume the cause-effect powers of its micro-level constituents. Crucially, once fused into the macro-level whole, the micro-level property instances cease to exist as independent entities (Humphreys, 1996, 1997). For example, during fertilization, a male gamete and a female gamete merge together and fuse into a novel diploid organism. Fusion seems to be able to account for the spatiotemporal emergence of integrated information (i.e., the idea that the complex self-defines in space and time) because the two original entities become something new courtesy of their temporally extended interaction. Moreover, it is able to account for the idea that integrated information and the cause-effect powers of the micro level do not synchronically co-exist, as the micro-level causal powers are subsumed by the macro-level causal powers. Fusion is also illuminating enough, as it is a physical operator used to account for relations between property instances in general, and it can be extended to relations between causal powers.

Then, fusion seems to be the best way to cash out the metaphysical relationship between integrated information and its basis: the complex instantiating integrated information thus self-defines as a whole by fusing the cause-effect powers of its components in space (i.e., depending on which components contribute causally to the complex) and time (i.e., depending on the temporal scale at which the components contribute to the complex). The fused cause-effect powers of the complex constitute integrated information, namely a new form of causal power. This resolves the tension between thinking that integrated information is both an emergent property and a physical property: it is an emergent property because it depends on the relations between the cause-effect powers of a physical substrate, but also a physical property because it is the fusion of its physical causal powers and therefore, a new version of those causal powers[6].

7.2 Autonomy Relation

If fusion is the best way to understand how integrated information is dependent upon the cause-effect powers of its physical substrate, the next question is in which sense integrated information is autonomous from, or irreducible to, the cause-effect powers of the complex. As we did for dependence, we need some criteria to evaluate the goodness of each possible autonomy relation in fitting IIT. First, the autonomy relation should be compatible with the identified dependence relation, namely fusion. Second, the autonomy of integrated information should be observer-independent;

[6] Note that, in this context, "physical" does not mean "microphysical", and therefore it is not equivalent to "non-mental". Rather, the idea of consciousness as physical, but emergent, property has to be intended as the idea that consciousness is a *natural* property. That is, amenable to be investigated by the natural science courtesy of its cause-effect powers. Thanks to Ignacio Cea for pointing this out.

that is, integrated information must be *in principle* irreducible to its basis. Third, it needs to be specific enough to account for the feature that renders integrated information irreducible to its basis; namely, it must explain in which sense integrated information is a new form of causal power. Let us see which autonomy relation complies best with these criteria.

I. *Non-linearity/Non-aggregativity*: macro-phenomena can be thought of as autonomous from micro-phenomena because the properties of macro-phenomena cannot be obtained through linear models (or linear summation) of the properties of micro-phenomena. For example, the emergent dynamic of clouds cannot be captured by a linear model of the clouds' component particles (Silberstein & McGeever, 1999). Although this relation might capture an important aspect of integrated information (e.g. we cannot obtain integrated information by summing up linearly the causal powers of its basis), it is hardly compatible with the criterion of observer-independency, as non-linearity and non-aggregativity seem to be properties of our models, rather than properties of the phenomenon itself. The relation of non-linearity/non-aggregativity, in itself, does not seem capture the hallmark of the ontological autonomy of integrated information.
II. *Multiple realizability*: the macro-level can be considered as autonomous from the micro-level in virtue of being multiple realizable. For example, the property of being a mouse trap is irreducible to its physical realiser, because there are many different ways of trapping mice; that is, there are multiple ways to physically implement the functional role played by a mouse trap. As seen above, this is not compatible with IIT, since integrated information is a physical property, and not a functional one. Thus, multiple realizability does not seem to fit nicely with IIT, and for this reason it does not seem to be the best autonomy relation upon which Emergentist IIT can be built.
III. *Fundamentality*: macro phenomena can be thought of as autonomous from micro phenomena in virtue of their ontological novelty; that is, in virtue of their novel causal powers (Barnes, 2012). For example, the dynamics of a country can be considered as autonomous from that of its citizens because the country can do things that its citizens cannot do (e.g., printing money or starting a war). If this is the case, the causal powers of the country are irreducible to that of its citizens because fundamentally novel. As mentioned above, Emergentist IIT must account for the fundamentality of integrated information in the strong emergentist sense. However, identifying the mark of the autonomy of integrated information with fundamentality does not illuminate what it means for integrated information to be such a novel form of causation. If we accept that integrated information *is* a form of causal power, then we need an account of the nature of these novel causal powers, and fundamentality, in itself, does not deliver such an account (Wilson, 2015, p. 374). Identifying fundamentality as the hallmark of the autonomy of integrated information is thus not specific enough, and therefore, although it is true that integrated information must be fundamental in the strong emergentist sense, it does not seem appropriate to cash out the autonomy of integrated information just in terms of fundamentality.

IV. *Downward Causation*: macro phenomena could achieve autonomy from their basis because they exhibit downward causation – the ability to influence micro-level phenomena (Kim, 2000; Flack, 2017). For example, the behaviour of a country can influence the behaviour of its citizens. Intended as a form of macro-to-micro determinative influence (Thompson & Varela, 2001), downward causation seems to be able to account for the irreducibility of integrated information: the fused cause-effect powers of a physical system would achieve autonomy courtesy of their ability to affect micro-level dynamics. The irreducibility of integrated information would then be given by the global constraint that integrated information, as a property of the whole complex, places on micro-level dynamics. Downward causation is thus *prima facie* compatible with fusion, is observer-independent (as it is a property of the emergent phenomenon itself and not of our models), and is specific enough to tell us in which sense integrated information can be considered as irreducible.

The autonomy relation between integrated information and its basis can thus be captured by downward causation, intended as macro-to-micro influence. Integrated information would be a form of causal powers that constrains micro-dynamics, and it is a novel causal property because it cannot be traced at the level of micro-causal powers only. Moreover, given the intrinsic aspect of integrated information, the micro-dynamics constrained by integrated information must be those encompassing the causal powers of the complex's components. This means that the constraint imposed by integrated information is a form of *intrinsic constraining*: a macro-to-micro influence where the global causal properties of the system as a whole influence the dynamics of the components whose interactions gave rise to the dynamic of the whole in the first place.

8 Emergentist IIT: The Overall Picture

The result of the previous analysis shows that the best way to understand integrated information as an emergent phenomenon is to think of it as dependent on the fusion of the cause-effect powers of the complex's components, and autonomous from such cause-effect powers because able to constrain the components' future states.

The claim is not that Emergentist IIT must necessarily be built upon fusion and downward causation. IIT may in fact be compatible with different characterizations of what it means for a property to be emergent. My claim, here, is that fusion and downward causation seem to constitute the pair of relations that fits best with IIT's claim because it raises fewer thorny questions than the other above-considered options. That is, resorting to these relations seems to be the least problematic strategy, in building Emergentist IIT.

A possible objection is that, independently of how they fit with IIT, fusion and downward causation are incompatible on their own. We can express the objection in the following way: if micro-level properties fuse together, and thus cease to have independent existence, it remains nothing, at the micro level, on which the macro-level property can exert its causal power. In the case of human fertilization, the male

and female gametes fuse in a diploid organism, but the diploid organism does not exert its novel causal powers upon the gametes.

However, in the context of Emergentist IIT, to fuse are not entities or properties, but the causal powers of the physical substrate (e.g., in the case of the human brain, it is not the neurons to fuse, but rather their causal powers). Thus, the fused whole is nothing but the doing of combined causal powers, and its existence is given by its ability to influence the state of the components of the physical substrate.

Addressing the objection of the incompatibility of fusion with downward causation would require formulating a more detailed account of the nature of causal powers (Hiddleston, 2005; Mumford, 2009). For example, the micro-level causal powers can be considered as non-operating after fusion, or, alternatively, it could be argued that only a proper subset of operating micro-level causal powers fuse, while the remaining operating micro-level powers constitute the existence of the micro-level properties affected by the fused whole[7]. Independently of how one sketches the details of the proposal, it seems possible to see that, at least in the context of Emergentist IIT, fusion and downward causation can be compatible.

To clarify, each physical component of the complex has dispositional properties that specify how it could constrain the other components of the complex. When the interrelation of these properties satisfies the five IIT postulates, the causal powers of the components fuse together and transform into a novel form of dispositional property, namely integrated information. The constraining power of each local physical component merge into a novel constraining power (i.e., integrated information) that belongs to the physical substrate as a whole.

Such global constraining power is an *intrinsic* constraining, because it does not affect physical components outside the complex, but rather it affects the complex itself. After the emergence of integrated information, the cause-effect powers of the complex's components are manifested in virtue of this global property of the complex. In doing so, integrated information determines the complex's state transition. Consciousness, in Emergentist IIT, is the emergent property of a system that determines its own unfolding, and in this sense, can be considered as global-to-local intrinsic constraining at an optimal spatiotemporal scale.

The picture portrayed by Emergentist IIT, according to which consciousness is a form of irreducible intrinsic constraining, accounts for the intrinsicality and the fundamentality of integrated information, but it does not suggest a panpsychist reading of IIT. The intrinsicality of integrated information is accounted for by the global-to-local constraining that corresponds to integrated information, since this is *intrinsic* constraining. The fundamentality of integrated information is instead accounted for by the ontological novelty of this intrinsic constraining: integrated information is a global property of a physical system that emerges through the transformation (via fusion)[8] of local cause-effect powers into a novel form of causal power. Nothing, in this picture, suggests that the physical components of a complex must be themselves rooted in phenomenality, as phenomenal properties emerge from the interrelation of

[7] A similar point is raised by Fallon & Blackmon (2021).

[8] Not surprisingly, Humphreys (2016) considers his own fusion model as a form of transformational emergence.

dispositional properties[9]. Thus, the anti-panpsychist assumptions that I have plugged into IIT's structure have provided a coherent picture that is able to account in anti-panpsychist terms for the intrinsicality and fundamentality of integrated information, namely the two properties upon which Panpsychist IIT is built. Given that IIT seem to account for these properties in a way that is different from how panpsychists intend them, and that the anti-panpsychist picture I have portrayed seem to be coherent and faithful to IIT, I argue that Emergentist IIT, as a metaphysical interpretation of IIT, is better than Panpsychist IIT.

It might be objected that Emergentist IIT does not importantly differ from Panpsychist IIT in terms of how consciousness is distributed throughout the universe. After all, if a simple system like an atom generates Φ^{Max}, and Φ^{Max} is taken to be an indicator of consciousness for Emergentist IIT too, then Emergentist IIT should be committed to the claim that the atom is conscious – and this is dangerously close to what panpsychism is committed to.

However, Emergentist IIT provides a metaphysical framework to justify the research into conditions that would limit the distribution of consciousness: approaching the problem of consciousness through an emergentist perspective provides a metaphysical platform for scientifically explaining consciousness as an emergent property. Since Emergentist IIT (as noted above) is not committed to a strict metaphysical identity between consciousness and integrated information, and IIT itself remarks that the identity between consciousness and integrated information is an *explanatory* one, an emergentist account might hold that the most informative level of explanation is not necessarily the physical one, but an emergent one: perhaps the most informative explanatory level is given by looking at adaptive systems that learn to exploit intrinsic constraining over learning and evolutionary time (Flack, 2017), rather than *any* physical system. This does not mean that Φ^{Max} can no longer be used as an indicator of consciousness, but rather that we are justified in applying it only to certain types of systems. In sum, an emergentist framework provides the metaphysical justification for the research program of finding those conditions that would exclude systems like atoms and molecules from the realm of conscious systems (despite generating Φ^{Max}). This would render Emergentist IIT less liberal, in the distribution of consciousness, than IIT proper.

Finally, there is an interesting point to be made about the ontological continuity between consciousness and the microphysical in the Emergentist IIT picture. On the one hand, the fundamentality of integrated information suggests that consciousness cannot be ontologically continuous with the microphysical, but rather that consciousness and physical reality are distinct in nature. On the other hand, however, the fact that integrated information is a form of causal power suggests that there is some sort of ontological continuity between consciousness and the microphysical world, as the microphysical is itself a tapestry of cause-effect powers. Emergentist IIT can resolve this tension by pointing at a form of causal monism according to which consciousness, as integrated information, is not a novel property of reality, but rather the

[9] As mentioned above, in Emergentist IIT dispositional properties are not phenomenal. One could hold on a panpsychist reading of Emergentist IIT by claiming that causal powers are themselves rooted in (proto) phenomenal properties. See (Mørch, 2019a) for a panprotopsychist view of IIT also based on fusion.

transformation of physical causal powers into a new form: the only entities existing in nature would be causal powers, and consciousness would be one of them. But consciousness would be nonetheless a peculiar form of causal power – a form manifested in intrinsic constraining.

9 Conclusion

The metaphysics of IIT is often seen as a panpsychist metaphysics, based on the intrinsicality and the fundamentality of integrated information. Here, I have argued that there is an emergentist way to interpret these two properties, and that this emergentist interpretation fits IIT better than the panpsychist interpretation. I have built Emergentist IIT on (i) IIT-driven assumptions like pandispositionalism and the explanatory identity between consciousness and integrated information; and (ii) anti-panpsychist assumptions that devoid the physical of phenomenality. Given these foundations and the formal result of causal emergence, I have argued that integrated information can be thought of as an emergent phenomenon. More specifically, it can be seen as dependent upon the fusion of the cause-effect powers of a physical substrate, and as autonomous in virtue of global-to-local determination. According to Emergentist IIT, consciousness is the constraining power of the system as a whole upon itself, when this power emerges from the fusion on the cause-effect powers of the system's components.

More work is needed to develop Emergentist IIT in details. For example, more research is needed to determine (i) whether more micro-to-micro transitions can occur during the fusion process, so that the temporal scale at which the macro-level unfolds can be slower than that of the micro-level; (ii) how exactly the constraining power of integrated information is exerted upon the local properties of the complex's components; (iii) whether it is possible that the macro-level can escape the Markovian dynamic (Großmann et al., 2020; Muñoz et al., 2020), thus influencing not only the manifestation of the micro-powers at the next time steps, but at successive time steps too.

Furthermore, more work is necessary to distinguish how Emergentist IIT accounts for the distribution of consciousness differently than IIT proper. Emergentism can be used to defend the claim that special sciences can be more informative than physics in certain contexts and conditions, and therefore consciousness, being an emergent property, must be explained via special sciences rather than physics. But more research is needed to establish and individuate the conditions in which we are justified in using integrated information as indicator of consciousness and those in which we are not.

Despite its embryonal form, Emergentist IIT can be seen as a viable metaphysical option for whoever takes IIT seriously as a theory of consciousness, and it seems to be the best way to place IIT within the debate about the place of consciousness in nature.

Acknowledgements The author wishes to thank Jakob Hohwy, Tim Bayne, Ignacio Cea, Francesco Ellia, Andrew McKilliam for many stimulating discussions around the topics of this paper. A previous version of

this paper has been presented at the *Realisation and Composition across the Sciences Workshop*, organized by the MetaScience team (Bristol, UK). The author wishes to thank the participants of the workshop for their valuable comments and feedbacks.

Author Contribution sole author.

Funding The author did not receive support from any organization for the submitted work.
Open Access funding enabled and organized by CAUL and its Member Institutions

Data Availability n/a.

Declarations

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

Ethics Approval n/a.

Consent n/a.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

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://www.mdpi.com/1099-4300/21/5/459

Banks, E. C. (2010). Neutral monism reconsidered. *Philosophical Psychology*, *23*(2), 173–187. doi:https://doi.org/10.1080/09515081003690418

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

Barnes, E. (2012). Emergence and Fundamentality. *Mind*, *121*(484), 873–901. http://www.jstor.org/stable/23407311

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

Baysan, U. (2019). Emergence, Function and Realization. In S. Gibb, R. Hendry, & T. Lancaster (Eds.), *The Routledge Handbook of Philosophy of Emergence*. Routledge

Bennett, K. (2017). *Making Things Up*. Oxford University Press

Bird, A. (2012). Monastic Dispositional Essentialism. In A. Bird, B. D. Ellis, & H. Sankey (Eds.), *Properties, Powers, and Structures: Issues in the Metaphysics of Realism* (pp. 35–41). Routledge

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247

Brüntrup, G. (2016). Emergent Panpsychism. In G. Brüntrup, & L. Jaskolla (Eds.), *Panpsychism: Contemporary Perspectives*. New York: Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199359943.003.0003

Cea, I. (2020). Integrated information theory of consciousness is a functionalist emergentism. *Synthese*. doi:https://doi.org/10.1007/s11229-020-02878-8

Chalmers, D. J. (1996). *The conscious mind: in search of a fundamental theory*. New York: Oxford University Press

Chalmers, D. J. (2006). Strong and weak emergence. In P. Davies, & P. Clayton (Eds.), *The Re-Emergence of Emergence: The Emergentist Hypothesis From Science to Religion*. Oxford University Press

Chalmers, D. J. (2016). Panpsychism and Panprotopsychism. *Panpsychism* New York: Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199359943.003.0002

Coleman, S. (2014). The Real Combination Problem: Panpsychism, Micro-Subjects, and Emergence. *Erkenntnis*, *79*(1), 19–44. doi:https://doi.org/10.1007/s10670-013-9431-x

Craver, C. F., & Bechtel, W. (2013). Interlevel Causation. In W. Dubitzky, O. Wolkenhauer, K. H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 1044–1047). New York, NY: Springer New York. doi:https://doi.org/10.1007/978-1-4419-9863-7_69

Davidson, D. (1970). Mental Events. In L. Foster, & J. W. Swanson (Eds.), *Essays on Actions and Events* (pp. 207–224). Clarendon Press

Dewhurst, J. (2021). Causal emergence from effective information: Neither causal nor emergent? *Thought: A Journal of Philosophy*. doi:https://doi.org/10.1002/tht3.489

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

Ellis, G. F. R., Noble, D., & O'Connor, T. (2012). Top-down causation: an integrating theme within and across the sciences? *Interface Focus*, *2*(1), 1–3. doi:doi:https://doi.org/10.1098/rsfs.2011.0110

Fallon, F., & Blackmon, J. C. (2021). IIT's Scientific Counter-Revolution: A Neuroscientific Theory's Physical and Metaphysical Implications. *Entropy*, *23*(8), 942. https://www.mdpi.com/1099-4300/23/8/942

Flack, J. C. (2017). Coarse-graining as a downward causation mechanism. *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences*, *375*(2109), 20160338. doi:doi:https://doi.org/10.1098/rsta.2016.0338

Goff, P. (2017). Consciousness and Fundamental Reality. Oup Usa

Grasso, M. (2019). IIT vs. Russellian Monism: A Metaphysical Showdown on the Content of Experience. *Journal of Consciousness Studies*, *26*(1–2), 48–75. https://www.ingentaconnect.com/content/imp/jcs/2019/00000026/f0020001/art00004

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

Großmann, G., Bortolussi, L., & Wolf, V. (2020). Efficient simulation of non-Markovian dynamics on complex networks. *PLOS ONE*, *15*(10), e0241394. doi:https://doi.org/10.1371/journal.pone.0241394

Hanson, J. R., & Walker, S. I. (2021). On the Non-uniqueness Problem in Integrated Information Theory. *bioRxiv*, 2021. 04.07.438793. https://doi.org/10.1101/2021.04.07.438793

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://www.mdpi.com/1099-4300/21/12/1160

Hiddleston, E. (2005). Causal Powers. *The British Journal for the Philosophy of Science*, *56*(1), 27–59. doi:https://doi.org/10.1093/phisci/axi102

Hoel, E., Albantakis, L., Marshall, W., & Tononi, G. (2016). Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neurosci Conscious*, 2016(1), niw012. doi:https://doi.org/10.1093/nc/niw012

Hoel, E., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49), 19790–19795. doi:https://doi.org/10.1073/pnas.1314922110

Humphreys, P. (1996). Aspects of Emergence. *Philosophical Topics*, *24*(1), 53–70. www.jstor.org/stable/43154222

Humphreys, P. (1997). How Properties Emerge. *Philosophy of Science*, *64*(1), 1–17. doi:https://doi.org/10.1086/392533

Humphreys, P. (2016). *Emergence: A Philosophical Account*. Oxford University Press

Kim, J. (1990). Supervenience as a philosophical concept. *Metaphilosophy*, *21*(1/2), 1–27. http://www.jstor.org/stable/24436754

Kim, J. (2000). Making sense of downward causation. In P. B. Andersen, C. Emmeche, N. O. Finnemann, & P. V. Christiansen (Eds.), *Downward Causation* (pp. 305–321). University of Aarhus Press

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

Leuenberger, S. (2020). The fundamental: Ungrounded or all-grounding? *Philosophical Studies*, *177*(9), 2647–2669
Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*(17), 556–567
Lewis, D. K. (1983). *Philosophical Papers*. Oxford University Press
Marshall, W., Albantakis, L., & Tononi, G. (2018). Black-boxing and cause-effect power. *PLOS Computational Biology*, *14*(4), e1006114. doi:https://doi.org/10.1371/journal.pcbi.1006114
Mediano, P. A. M., Rosas, F. E., Bor, D., Seth, A. K., & Barrett, A. B. (2022). The strength of weak integrated information theory. *Trends in Cognitive Sciences*. doi:https://doi.org/10.1016/j.tics.2022.04.008
Mediano, P. A. M., Seth, A. K., & Barrett, A. B. (2019). Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy*, *21*(1), 17. https://www.mdpi.com/1099-4300/21/1/17
Menzies, P., & Price, H. (1993). Causation as a Secondary Quality. *The British Journal for the Philosophy of Science*, *44*(2), 187–203. http://www.jstor.org/stable/687643
Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), doi:https://doi.org/10.33735/phimisci.2020.II.54
Montero, B., & Papineau, D. (2005). A defence of the via negativa argument for physicalism. *Analysis*, *65*(3), 233–237
Moon, K. (2019). Exclusion and Underdetermined Qualia. *Entropy*, *21*(4), 405. https://www.mdpi.com/1099-4300/21/4/405
Mørch, H. H. (2019a). Is Consciousness Intrinsic?: A Problem for the Integrated Information Theory. Journal of Consciousness Studies, 26(1–2), 133–162(30).
Mørch, H. H. (2019b). Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism? *Erkenntnis*, *84*(5), 1065–1085. doi:https://doi.org/10.1007/s10670-018-9995-6
Mørch, H. H. (2020). Does Dispositionalism Entail Panpsychism? *Topoi*, *39*(5), 1073–1088. doi:https://doi.org/10.1007/s11245-018-9604-y
Mumford, S. (2009). Causal Powers and Capacities. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford Handbook of Causation*. Oxford University Press
Muñoz, R. N., Leung, A., Zecevik, A., Pollock, F. A., Cohen, D., van Swinderen, B., et al. (2020). General anesthesia reduces complexity and temporal asymmetry of the informational structures derived from neural recordings in Drosophila. *Physical Review Research*, *2*(2), 023219. doi:https://doi.org/10.1103/PhysRevResearch.2.023219
Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435–450. doi:https://doi.org/10.2307/2183914
O'Connor, T. (1994). Emergent properties. *American Philosophical Quarterly*, *31*(2), 91–104
O'Connor, T. (2020). 'Emergent Properties' Fall 2020 E. N. Zalta *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/fall2020/entries/properties-emergent/
O'Connor, T., & Wong, H. Y. (2005). The Metaphysics of Emergence. *Noûs*, *39*(4), 658–678. doi:https://doi.org/10.1111/j.0029-4624.2005.00543.x
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. doi:https://doi.org/10.1371/journal.pcbi.1003588
Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press
Russell, B. (1921). *The Analysis of Mind*. Duke University Press
Russell, B. (1927). The Analysis of Matter. Kegan Paul
Seager, W. (2012). Emergentist Panpsychism. *Journal of Consciousness Studies*, *19*(9–10), 19–39. https://www.ingentaconnect.com/content/imp/jcs/2012/00000019/f0020009/art00002
Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*. doi:https://doi.org/10.1038/s41583-022-00587-4
Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. doi:https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
Silberstein, M., & McGeever, J. (1999). The Search for Ontological Emergence. *The Philosophical Quarterly (1950-)*, *49*(195), 182–200. http://www.jstor.org/stable/2660261
Thompson, E., & Varela, F. J. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences*, *5*(10), 418–425. doi:https://doi.org/10.1016/S1364-6613(00)01750-2
Tononi, G. (2008). Consciousness as Integrated Information: A Provisional Manifesto. *Biological Bulletin*, *215*(3), 216–242. doi:https://doi.org/10.2307/25470707

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

Tononi, G. (2017). Integrated Information Theory of Consciousness. *The Blackwell Companion to Consciousness* (pp. 621–633). doi:https://doi.org/10.1002/9781119132363.ch44

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. doi:https://doi.org/10.1038/nrn.2016.44

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions Of The Royal Society Of London. Series B, Biological Sciences*, *370*(1668), doi:https://doi.org/10.1098/rstb.2014.0167

Van Cleve, J. (1990). Mind–Dust or Magic? Panpsychism Versus Emergence. *Philosophical Perspectives*, *4*, 215–226. doi:https://doi.org/10.2307/2214193

Wilson, J. (1999). How superduper does a physicalist supervenience need to be? *Philosophical Quarterly*, *49*(194), 33–52

Wilson, J. (2015). Metaphysical emergence: Weak and Strong. In T. Bigaj, & C. Wuthrich (Eds.), *Metaphysics in Contemporary Physics* (pp. 251–306). Poznan Studies in the Philosophy of the Sciences and the Humanities

Woodward, J. (2000). Explanation and Invariance in the Special Sciences. *The British Journal for the Philosophy of Science*, *51*(2), 197–254. doi:https://doi.org/10.1093/bjps/51.2.197

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press

Yablo, S. (1992). Mental Causation. *The Philosophical Review*, *101*(2), 245–280. doi:https://doi.org/10.2307/2185535

Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2022). The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nat Hum Behav*. doi:https://doi.org/10.1038/s41562-021-01284-5

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.