

RESEARCH ARTICLE

Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms

Larissa Albantakis¹, Leonardo Barbosa^{1,2}, Graham Findlay^{1,3}, Matteo Grasso¹, Andrew M. Haun¹, William Marshall^{1,4}, William G. P. Mayner^{1,3}, Alireza Zaeemzadeh¹, Melanie Boly^{1,5}, Bjørn E. Juel^{1,6}, Shuntaro Sasai^{1,7}, Keiko Fujii¹, Isaac David¹, Jeremiah Hendren^{1,8}, Jonathan P. Lang¹, Giulio Tononi^{1*}

1 Department of Psychiatry, University of Wisconsin, Madison, Wisconsin, United States of America, **2** Fralin Biomedical Research Institute at VTC, Virginia Tech, Roanoke, Virginia, United States of America, **3** Neuroscience Training Program, University of Wisconsin, Madison, Wisconsin, United States of America, **4** Department of Mathematics and Statistics, Brock University, St. Catharines, Ontario, Canada, **5** Department of Neurology, University of Wisconsin, Madison, Wisconsin, United States of America, **6** Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway, **7** Araya Inc., Tokyo, Japan, **8** Graduate School Language & Literature, Ludwig Maximilian University of Munich, Munich, Germany

☞ These authors contributed equally to this work.

* gtononi@wisc.edu



OPEN ACCESS

Citation: Albantakis L, Barbosa L, Findlay G, Grasso M, Haun AM, Marshall W, et al. (2023) Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS Comput Biol* 19(10): e1011465. <https://doi.org/10.1371/journal.pcbi.1011465>

Editor: Lyle J. Graham, Université Paris Descartes, Centre National de la Recherche Scientifique, FRANCE

Received: January 11, 2023

Accepted: August 26, 2023

Published: October 17, 2023

Copyright: © 2023 Albantakis et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: There are no primary data in the paper; the code used to produce the results and analyses presented in this manuscript is available at <https://github.com/wmayner/pyphi/tree/feature/iit-4.0/pyphi>.

Funding: This project was made possible through the support of a grant from Templeton World Charity Foundation (TWCF0216, G.T.). In addition, this research was supported by the David P White Chair in Sleep Medicine at the University of

Abstract

This paper presents Integrated Information Theory (IIT) 4.0. IIT aims to account for the properties of experience in physical (operational) terms. It identifies the essential properties of experience (axioms), infers the necessary and sufficient properties that its substrate must satisfy (postulates), and expresses them in mathematical terms. In principle, the postulates can be applied to any system of units in a state to determine whether it is conscious, to what degree, and in what way. IIT offers a parsimonious explanation of empirical evidence, makes testable predictions concerning both the presence and the quality of experience, and permits inferences and extrapolations. IIT 4.0 incorporates several developments of the past ten years, including a more accurate formulation of the axioms as postulates and mathematical expressions, the introduction of a unique measure of intrinsic information that is consistent with the postulates, and an explicit assessment of causal relations. By fully unfolding a system's irreducible cause–effect power, the distinctions and relations specified by a substrate can account for the quality of experience.

Author summary

As a theory of consciousness, IIT aims to answer two questions: 1) Why is experience present vs. absent? and 2) Why do specific experiences feel the way they do? The theory's starting point is the existence of experience. IIT then aims to account for phenomenal existence and its essential properties in physical terms. It concludes that a substrate—a set of interacting units—can support consciousness if it can take and make a difference for itself (intrinsicity), select a specific cause and effect as an irreducible whole with a

Wisconsin-Madison, by the Tiny Blue Dot Foundation (UW 133AAG3451; G.T.), and by the Natural Science and Engineering Research Council of Canada (NSERC; RGPIN-2019-05418; W.M.). L. A. also acknowledges the support of a grant from the Templeton World Charity Foundation (TWCF-2020-20526, L.A.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: G.T. holds an executive position and has a financial interest in Intrinsic Powers, Inc., a company whose purpose is to develop a device that can be used in the clinic to assess the presence and absence of consciousness in patients. This does not pose any conflict of interest with regard to the work undertaken for this publication.

definite border and grain, and specify a structure of causes and effects through subsets of its units. To that end, IIT provides a mathematical formalism that can be employed to “unfold” the substrate’s cause–effect structure. This allows IIT to answer the two questions above: 1) Experience is present for any substrate that fulfills the essential properties of existence, and 2) specific experiences feel the way they do because of the specific cause–effect structure specified by their substrates. The theory is consistent with neurological data, and some of its core principles have been successfully tested empirically.

Introduction

A scientific theory of consciousness should account for experience, which is subjective, in objective terms [1]. Being conscious—having an experience—is understood to mean that “there is something it is like to be” [2]: something it is like to see a blue sky, hear the ocean roar, dream of a friend’s face, imagine a melody flow, contemplate a choice, or reflect on the experience one is having.

IIT aims to account for phenomenal properties—the properties of experience—in physical terms. IIT’s starting point is experience itself rather than its behavioral, functional, or neural correlates [1]. Furthermore, in IIT “physical” is meant in a strictly operational sense—in terms of what can be observed and manipulated.

The starting point of IIT is the existence of an experience, which is immediate and irrefutable [3]. From this “zeroth” axiom, IIT sets out to identify the essential properties of consciousness—those that are immediate and irrefutably true of every conceivable experience. These are IIT’s five axioms of phenomenal existence: every experience is for the experiencer (intrinsicity), specific (information), unitary (integration), definite (exclusion), and structured (composition).

Unlike phenomenal existence, which is immediate and irrefutable (an axiom), physical existence is an explanatory construct (a postulate), and it is assessed operationally (from within consciousness): in physical terms, to be is to have cause–effect power. In other words, something can be said to exist physically if it can “take and make a difference”—bear a cause and produce an effect—as judged by a conscious observer/manipulator.

The next step of IIT is to formulate the essential phenomenal properties (the axioms) in terms of corresponding physical properties (the postulates). This formulation is an “inference to a good explanation” and rests on basic assumptions such as realism, physicalism, and atomism (see [Box 1](#): Methodological guidelines of IIT). If IIT is correct, the substrate of consciousness (see (1) in [S1 Notes](#)), beyond having cause–effect power (existence), must satisfy all five essential phenomenal properties in physical terms: its cause–effect power must be for itself (intrinsicity), specific (information), unitary (integration), definite (exclusion), and structured (composition).

On this basis, IIT proposes a fundamental explanatory identity: an experience is identical to the cause–effect structure unfolded from a maximal substrate (defined below). Accordingly, all the specific phenomenal properties of any experience must have a good explanation in terms of the specific physical properties of the corresponding cause–effect structure, with no additional ingredients.

Based again on “inferences to a good explanation” (see [Box 1](#)), IIT formulates the postulates in a mathematical framework that is in principle applicable to general models of interacting units (but see (2) in [S1 Notes](#)). A mathematical framework is needed (*a*) to evaluate whether the theory is self-consistent and compatible with our overall knowledge about the world, (*b*) to

make specific predictions regarding the quality and quantity of our experiences and their substrate within the brain, and (c) to extrapolate from our own consciousness to infer the presence (or absence) and nature of consciousness in beings different from ourselves.

Ultimately, the theory should account for why our consciousness depends on certain portions of the world and their state, such as certain regions of the brain and not others, and for why it fades during dreamless sleep, even though the brain remains active. It should also account for why an experience feels the way it does—why the sky feels extended, why a melody feels flowing in time, and so on. Moreover, the theory makes several predictions concerning both the presence and the quality of experience, some of which have been and are being tested empirically [4].

While the main tenets of the theory have remained the same, its formal framework has been progressively refined and extended [5–8]. Compared to IIT 1.0 [5, 6], 2.0 [7, 9], and 3.0 [8], IIT 4.0 presents a more complete, self-consistent formulation and incorporates several recent advances [10–13]. Chief among them are a more accurate formulation of the axioms as postulates and mathematical expressions, the introduction of an Intrinsic Difference (ID) measure [12, 14] that is uniquely consistent with IIT's postulates, and the explicit assessment of causal relations [11].

In what follows, after introducing IIT's axioms and postulates, we provide its updated mathematical formalism. In the “Results and discussion” section, we apply the mathematical framework of IIT to representative examples and discuss some of their implications. The article is meant as a reference for the theory's mathematical formalism, a concise demonstration of its internal consistency, and an illustration of how a substrate's cause–effect structure is unfolded computationally. A discussion of the theory's motivation, its axioms and postulates, and its assumptions and implications can be found in a forthcoming book (see (3) in [S1 Notes](#)) and wiki [15] as well as in several publications [1, 16–21]. A survey of the explanatory power and experimental predictions of IIT can be found in [4]. The way IIT's analysis of cause–effect power can be applied to actual causation, or “what caused what,” is presented in [10].

From phenomenal axioms to physical postulates

Axioms of phenomenal existence

That experience exists—that “there is something it is like to be”—is immediate and irrefutable, as everybody can confirm, say, upon awakening from dreamless sleep. Phenomenal existence is immediate in the sense that my experience is simply there, directly rather than indirectly: I do not need to infer its existence from something else. It is irrefutable because the very doubting that my experience exists is itself an experience that exists—the experience of doubting [1, 3]. Thus, to claim that my experience does not exist is self-contradictory or absurd. The existence of experience is IIT's zeroth axiom.

Existence Experience *exists*: there is *something*.

Traditionally, an axiom is a statement that is assumed to be true, cannot be inferred from any other statement, and can serve as a starting point for inferences. The existence of experience is the ultimate axiom—the starting point for everything, including logic and physics.

On this basis, IIT proceeds by considering whether experience—phenomenal existence—has some axiomatic or essential properties, properties that are immediate and irrefutably true of every conceivable experience. Drawing on introspection and reason, IIT identifies the following five:

Intrinsicity Experience is *intrinsic*: it exists *for itself*.

Information Experience is *specific*: it is *this one*.

Integration Experience is *unitary*: it is *a whole*, irreducible to separate experiences.

Exclusion Experience is *definite*: it is *this whole*.

Composition Experience is *structured*: it is composed of *distinctions* and the *relations* that bind them together, yielding a *phenomenal structure* that feels *the way it feels*.

To exemplify, if I awaken from dreamless sleep and experience the white wall of my room, my bed, and my body, the experience not only exists, immediately and irrefutably, but 1) it exists for me, not for something else, 2) it is specific (this one experience, not a generic one), 3) it is unitary (the left side is not experienced separately from the right side, and vice versa), 4) it is definite (it includes the visual scene in front of me—neither less, say, its left side only, nor more, say, the wall behind my head), 5) it is structured by distinctions (the wall, the bed, the body) and relations (the body is on the bed, the bed in the room), which make it feel the way it does and not some other way.

The axioms are not only immediately given, but they are irrefutably true of every conceivable experience. For example, once properly understood, the unity of experience cannot be refuted. Trying to conceive of an experience that were not unitary leads to conceiving of two separate experiences, each of which is unitary, which reaffirms the validity of the axiom. Even though each of the axioms spells out an essential property in its own right, the axioms must be considered together to properly characterize phenomenal existence.

IIT takes the above set of axioms to be complete: there are no further properties of experience that are essential. Other properties that might be considered as candidates for axiomatic status include space (experience typically takes place in some spatial frame), time (an experience usually feels like it flows from a past to a future), change (an experience usually transitions or flows into another), subject–object distinction (an experience seems to involve both a subject and an object), intentionality (experiences usually refer to something in the world, or at least to something other than the subject), a sense of self (many experiences include a reference to one’s body or even to one’s narrative self), figure–ground segregation (an experience usually includes some object and some background), situatedness (an experience is often bound to a time and a place), will (experience offers the opportunity for action), and affect (experience is often colored by some mood), among others. However, experiences lacking each of these candidate properties are conceivable—that is, conceiving of them does not lead to self-contradiction or absurdity. They are also achievable, as revealed by altered states of consciousness reached through dreaming, meditative practices, or drugs.

Postulates of physical existence

To account for the many regularities of experience ([Box 1](#)), it is a good inference to assume the existence of a world that persists independently of one’s experience (*realism*). From within consciousness, we can probe the physical existence of things outside of our experience operationally—through observations and manipulations. To be granted physical existence, something should have the power to “take a difference” (be affected) and “make a difference” (produce effects) in a reliable way (*physicalism*). IIT also assumes “operational reductionism,” which means that, ideally, to establish what exists in physical terms, one would start from the smallest units that can take and make a difference, so that nothing is left out (*atomism*).

By characterizing physical existence operationally as cause–effect power, IIT can proceed to formulate the axioms of phenomenal existence as postulates of physical existence. This

establishes the requirements for the *substrate of consciousness*, where “substrate” is meant operationally as a set of units that can be observed and manipulated.

Existence The substrate of consciousness can be characterized operationally by *cause-effect power*: its units must *take and make a difference*.

Building from this “zeroth” postulate, IIT formulates the five axioms in terms of postulates of physical existence that must be satisfied by the substrate of consciousness:

Intrinsicity Its cause-effect power must be *intrinsic*: it must take and make a difference *within itself*.

Information Its cause-effect power must be *specific*: it must be in *this state* and select *this cause-effect state*.

This state is the one with maximal *intrinsic information* (ii), a measure of the difference a system takes or makes over itself for a given cause state and effect state.

Integration Its cause-effect power must be *unitary*: it must specify its cause-effect state as a *whole set* of units, irreducible to separate subsets of units.

Irreducibility is measured by *integrated information* (φ) over the substrate’s minimum partition.

Exclusion Its cause-effect power must be *definite*: it must specify its cause-effect state as *this whole set* of units.

This is the set of units that is maximally irreducible, as measured by maximum φ (φ^*). This set is called a *maximal substrate*, also known as a *complex* [8, 13].

Composition Its cause-effect power must be *structured*: subsets of its units must specify cause-effect states over subsets of units (*distinctions*) that can overlap with one another (*relations*), yielding a *cause-effect structure* or Φ -structure (“Phi-structure”) that is *the way it is*.

Distinctions and relations, in turn, must also satisfy the postulates of physical existence: they must have cause-effect power, within the substrate of consciousness, in a specific, unitary, and definite way (they do not have components, being components themselves). They thus have an associated φ value. The Φ -structure unfolded from a complex corresponds to the quality of consciousness. The sum total of the φ values of the distinctions and relations that compose the Φ -structure measures its *structure integrated information* Φ (“big Phi,” “structure Phi”) and corresponds to the quantity of consciousness.

According to IIT, the physical properties characterized by the postulates are necessary and sufficient for an entity to be conscious. They are necessary because they are needed to account for the properties of experience that are essential, in the sense that it is inconceivable for an experience to lack any one of them. They are also sufficient because no additional property of experience is essential, in the sense that it is conceivable for an experience to lack that property. Thus, no additional physical property is a necessary requirement for being a substrate of consciousness.

The postulates of IIT have been and are being applied to account for the location of the substrate of consciousness in the brain [4] and for its loss and recovery in physiological and pathological conditions [22, 23].

The explanatory identity between experiences and Φ -structures

Having determined the necessary and sufficient conditions for a substrate to support consciousness, IIT proposes an explanatory identity: every property of an experience is accounted

for in full by the physical properties of the Φ -structure unfolded from a maximal substrate (a complex) in its current state, with no further or “ad hoc” ingredients. That is, there must be a one-to-one correspondence between the way the experience feels and the way distinctions and relations are structured. Importantly, the identity is not meant as a correspondence between the properties of two separate things. Instead, the identity should be understood in an explanatory sense: the intrinsic (subjective) feeling of the experience can be explained extrinsically (objectively, *i.e.*, operationally or physically) in terms of cause–effect power (see (4) in [S1 Notes](#)).

The explanatory identity has been applied to account for how space feels (spatial extendedness) and which neural substrates may account for it [11]. Ongoing work is applying the identity to provide a basic account of the feeling of temporal flow [24] and that of objects [25].

Box 1. Methodological guidelines of IIT

Inference to a good explanation

We should generally assume that an explanation is good if it can account for a broad set of facts (*scope*), does so in a unified manner (*synthesis*), can explain facts precisely (*specificity*), is internally coherent (*self-consistency*), is coherent with our overall understanding of things (*system consistency*), is simpler than alternatives (*simplicity*), and can make testable predictions (*scientific validation*). For example, IIT 4.0 aims at expressing the postulates of intrinsicity, information, integration, and exclusion in a self-consistent manner when applied to systems, causal distinctions, and relations (see formulas).

Realism

We should assume that something exists (and persists) independently of our own experience. This is a much better hypothesis than solipsism, which explains nothing and predicts nothing. Although IIT starts from our own phenomenology, it aims to account for the many regularities of experience in a way that is fully consistent with realism.

Operational physicalism

To assess what exists independently of our own experience, we should employ an operational criterion: we should systematically observe and manipulate a substrate’s units and determine that they can indeed take and make a difference in a way that is reliable. Doing so demonstrates a substrate’s cause–effect power—the signature of physical existence. Ideally, cause–effect power is fully captured by a substrate’s transition probability matrix (TPM) (1). This assumption is embedded in IIT’s zeroth postulate.

Operational reductionism (“atomism”)

Ideally, we should account for what exists physically in terms of the smallest units we can observe and manipulate, as captured by unit TPMs. Doing so would leave nothing unaccounted for. IIT assumes that, in principle, it should be possible to account for everything purely in terms of cause–effect power—cause–effect power “all the way down” to conditional probabilities between atomic units (see (5) in [S1 Notes](#)). Eventually, this would leave neither room nor need to assume intrinsic properties or laws.

Intrinsic perspective

When accounting for experience itself in physical terms, existence should be evaluated from the intrinsic perspective of an entity—what exists for the entity itself—not from the perspective of an external observer. This assumption is embedded in IIT’s postulate of intrinsicity and has several consequences. One is that, from the intrinsic perspective, the quality and quantity of existence must be observer-independent and cannot be arbitrary. For instance, information in IIT must be relative to the specific state the entity is in, rather than an average of states as assessed by an external observer. Similarly, it should be evaluated based on the uniform distribution of possible states, as captured by the entity’s TPM (1), rather than on an observed probability distribution. By the same token, units outside the entity should be treated as background conditions that do not contribute directly to what the system is. The intrinsic perspective also imposes a tension between expansion and dilution (see below and [12, 14]): from the intrinsic perspective of a system (or a mechanism within the system), having more units may increase its informativeness (cause–effect power measured as deviation from chance), while at the same time diluting its selectivity (ability to concentrate cause–effect power over a specific state).

Overview of IIT’s framework

IIT 4.0 aims at providing a formal framework to characterize the cause–effect structure of a substrate in a given state by expressing IIT’s postulates in mathematical terms. In line with operational physicalism (Box 1), we characterize a substrate by the transition probability function of its constituting units.

On this basis, the IIT formalism first identifies sets of units that fulfill all required properties of a substrate of consciousness according to the postulates of physical existence. First, for a candidate system, we determine a maximal cause–effect state based on the intrinsic information (ii) that the system in its current state specifies over its possible cause states and effect states. We then determine the maximal substrate based on the integrated information (φ_s , “system phi”) of the maximal cause–effect state. To qualify as a substrate of consciousness, a candidate system must specify a maximum of integrated information (φ_s^*) compared to all competing candidate systems with overlapping units.

The second part of the IIT formalism *unfolds* the cause–effect structure specified by a maximal substrate in its current state, its Φ -structure. To that end, we determine the distinctions and relations specified by the substrate’s subsets according to the postulates of physical existence. Distinctions are cause–effect states specified over subsets of substrate units (*purviews*) by subsets of substrate units (*mechanisms*). Relations are congruent overlaps among distinctions’ cause and/or effect states. Distinctions and relations are also characterized by their integrated information (φ_d , φ_r). The Φ -structure they compose corresponds to the quality of the experience specified by the substrate; the sum of their φ_{dir} values corresponds to its quantity (Φ).

While IIT must still be considered as work in progress, having undergone successive refinements, IIT 4.0 is the first formulation of IIT that strives to characterize Φ -structures completely and to do so based on measures that satisfy the postulates uniquely. For a comparison of the updated framework with IIT 1.0, 2.0, and 3.0, see [S2 Text](#).

Substrates, transition probabilities, and cause–effect power

IIT takes physical existence as synonymous with having cause–effect power, the ability to take and make a difference. Consequently, a substrate U with state space Ω_U is operationally defined by its potential interactions, assessed in terms of conditional probabilities (physicalism, [Box 1](#)). We denote the complete transition probability function of a substrate U over a system update $u \rightarrow \bar{u}$ as

$$\mathcal{T}_U \equiv p(\bar{u} | u), \quad u, \bar{u} \in \Omega_U. \quad (1)$$

A substrate in IIT can be described as a stochastic system $U = \{U_1, U_2, \dots, U_n\}$ of n interacting units with state space $\Omega_U = \prod_i \Omega_{U_i}$ and current state $u \in \Omega_U$. We define units in state u as a set of tuples, where each tuple contains the unit and the state of the unit, *i.e.*, $u = \{(U_i, \text{state}(U_i)) : U_i \in U\}$. This allows us to define set operations over u that consider both the units and their states. Ω_U is the set of all possible such tuple sets, corresponding to all the possible states of U . We assume that the system updates in discrete steps, that the state space Ω_U is finite, and that the individual random variables $U_i \in U$ are conditionally independent from each other given the preceding state of U :

$$p(\bar{u} | u) = \prod_{i=1}^n p(\bar{u}_i | u). \quad (2)$$

Finally, we assume a complete description of the substrate, which means that we can determine the conditional probabilities in (2) for every system state, with $p(\bar{u} | u) = p(\bar{u} | \text{do}(u))$ [10, 26–28], where the “do-operator” $\text{do}(u)$ indicates that u is imposed by intervention. This implies that U must correspond to a causal network [10], and \mathcal{T}_U is a transition probability matrix (TPM) of size $|\Omega_U|$ (see (6) in [S1 Notes](#)).

The TPM \mathcal{T}_U , which forms the starting point of IIT’s analysis, serves as an overall description of a system’s causal evolution under all possible interventions: what is the probability that the system will transition into each of its possible states upon being initialized into every possible state ([Fig 1](#))? (Notably, there is no additional role for intrinsic physical properties or laws of nature.) In practice, a causal model will be neither complete nor atomic (capturing the smallest units that can be observed and manipulated), but will capture the relevant features of what we are trying to explain and predict (see (7) in [S1 Notes](#)).

In the “Results and discussion” section, the IIT formalism will be applied to extremely simple, simulated networks, rather than causal models of actual substrates. The cause–effect structures derived from these simple networks only serve as convenient illustrations of how a hypothetical substrate’s cause–effect power can be unfolded.

Implementing the postulates

In what follows, our goal is to evaluate whether a hypothetical substrate (also called “system”) satisfies all the postulates of IIT. To that end, we must verify whether the system has cause–effect power that is intrinsic, specific, integrated, definite, and structured.

Existence. According to IIT, existence understood as cause–effect power requires the capacity to both take *and* make a difference (see [Box 2](#), Principle of being). On the basis of a complete description of the system in terms of interventional conditional probabilities (\mathcal{T}_U) (1), cause–effect power can be quantified as causal *informativeness*. Cause informativeness measures how much a potential cause increases the probability of the current state, and effect informativeness how much the current state increases the probability of a potential effect (as compared to chance).

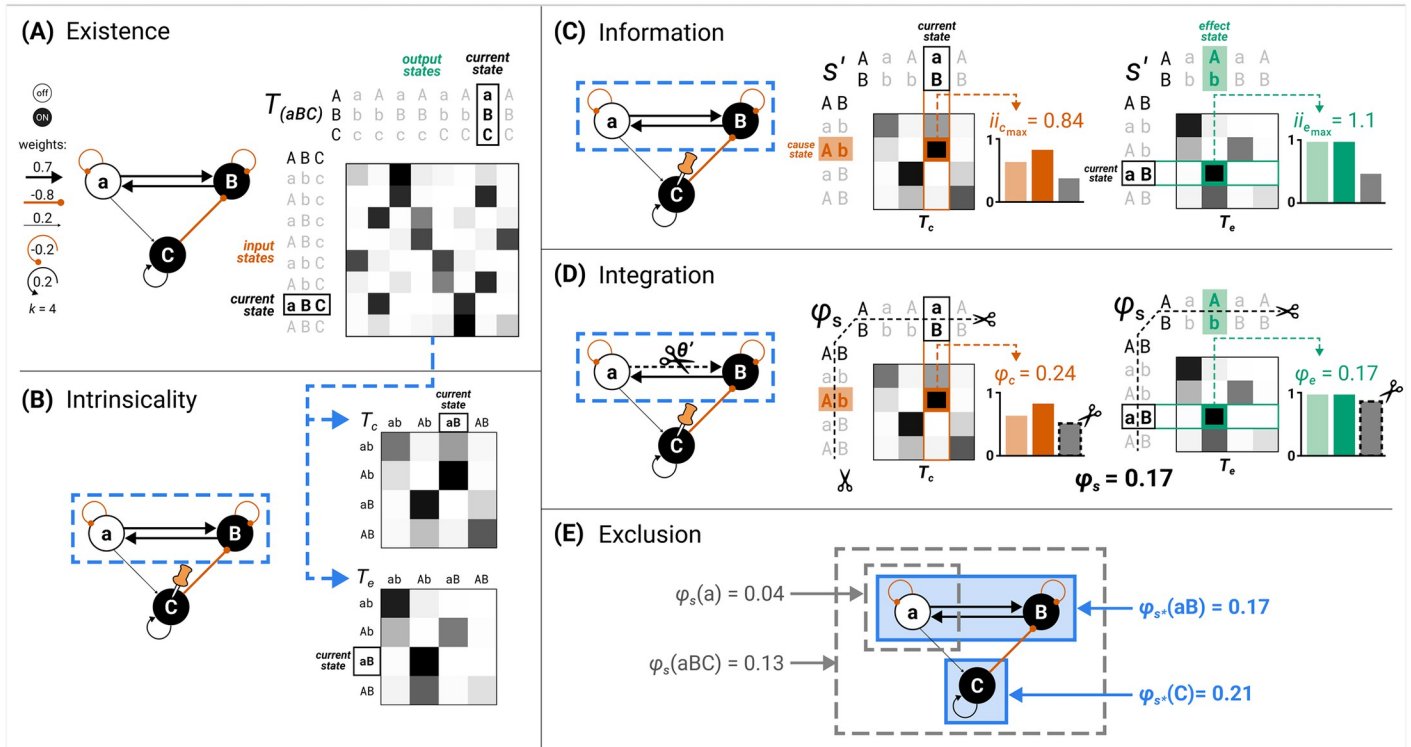


Fig 1. Identifying substrates of consciousness through the postulates of existence, intrinsicality, information, integration, and exclusion. (A) The substrate $S = aBC$ in state $(-1, 1, 1)$ (lowercase letters for units indicated state “-1,” uppercase letters state “+1”) is the starting point for applying the postulates. The substrate updates its state according to the depicted transition probability matrix (TPM) (gray shading indicates probability value from white ($p = 0$) to black ($p = 1$); each unit follows a logistic equation (see “Results” for definition) with $k = 4.0$ and connection weights as indicated in the causal model). Existence requires that the substrate must have cause–effect power, meaning that the TPM among substrate states must differ from chance. (B) Intrinsicality requires that a candidate substrate, for example, units aB , has cause–effect power over itself. Units outside the candidate substrate (in this case, unit C) are treated as background conditions. The corresponding cause and effect TPMs (T_c and T_e) of system aB are depicted on the right. (C) Information requires that the candidate substrate aB selects a specific cause–effect state (s'). This is the cause state (red) and effect state (green) for which intrinsic information (ii) is maximal. Bar plots on the right indicate the three probability terms relevant for computing ii_c (7) and ii_e (5): the selectivity (light colored bar), as well as the constrained (dark colored bar) and unconstrained (gray bar) effect probabilities in the informativeness term. (D) Integration requires that the substrate specifies its cause–effect state irreducibly (“as one”). This is established by identifying the minimum partition (MIP; θ') and measuring the integrated information of the system (φ_s)—the minimum between cause integrated information (φ_c) and effect integrated information (φ_e). Here, gray bars represent the partitioned probability required for computing φ_c (20) and φ_e (19). (E) Exclusion requires that the substrate of consciousness is definite, including some units and excluding others. This is established by identifying the candidate substrate with the maximum value of system integrated information (φ_s^*)—the maximal substrate, or complex. In this case, aB is a complex since its system integrated information ($\varphi_s = 0.17$) is higher than that of all other overlapping systems (for example, subset a with $\varphi_s = 0.04$ and superset aBC with $\varphi_s = 0.13$).

<https://doi.org/10.1371/journal.pcbi.1011465.g001>

Intrinsicality. Building upon the existence postulate, the intrinsicality postulate further requires that a system exerts cause–effect power *within itself*. In general, the systems we want to evaluate are open systems $S \subseteq U$ that are part of a larger “universe” U . From the intrinsic perspective of a system S (see Box 1), the set of the remaining units $W = U \setminus S$ merely act as background conditions that do not contribute directly to cause–effect power. To enforce this, we causally marginalize the background units, conditional on the current state of the universe, rendering them causally inert (see “Identifying substrates of consciousness” for details).

Information. The information postulate requires that a system’s cause–effect power be specific: the system in its current state must select a specific cause–effect state for its units. Based on the *principle of maximal existence* (Box 2), this is the state for which intrinsic information is maximal—the *maximal cause–effect state*. *Intrinsic information* (ii) measures the difference a system takes or makes over itself for a given cause and effect state as the product of

informativeness and selectivity. As we have seen (existence), *informativeness* quantifies the causal power of a system in its current state as a reduction of uncertainty with respect to chance. *Selectivity* measures how much cause–effect power is concentrated over that specific cause or effect state. Selectivity is reduced by uncertainty in the cause or effect state with respect to other potential cause and effect states.

From the intrinsic perspective of the system, the product of informativeness and selectivity leads to a tension between *expansion* and *dilution*, whereby a system comprising more units may show increased deviation from chance but decreased concentration of cause–effect power over a specific state [12, 14].

Integration. By the integration postulate, it is not sufficient for a system to have cause–effect power within itself and select a specific cause–effect state: it must also specify its maximal cause–effect state in a way that is irreducible. This can be assessed by *partitioning* the set of units that constitute the system into separate parts. The system integrated information (φ_s) then quantifies how much the intrinsic information specified by the maximal state is reduced due to the partition (see (8) in S1 Notes). Integrated information is evaluated over the partition that makes the least difference, the *minimum partition* (MIP), in accordance with the *principle of minimal existence* (see Box 2).

Integrated information is highly sensitive to the presence of *fault lines*—partitions that separate parts of a system that interact weakly or directionally [13].

Exclusion. Many overlapping sets of units may have a positive value of integrated information (φ_s). However, the exclusion postulate requires that the substrate of consciousness must be constituted of a definite set of units, neither less nor more. Moreover, units, updates, and states must have a definite grain. Operationally, the exclusion postulate is enforced by selecting the set of units that maximizes integrated information over itself (φ_s^*), based again on the principle of maximal existence (see Box 2). That set of units is called a *maximal substrate*, or *complex*. Over a universal substrate, sets of units for which integrated information is maximal compared to all competing candidate systems with overlapping units can be assessed recursively (by identifying the first complex, then the second complex, and so on).

Composition. Once a complex has been identified, composition requires that we characterize its *cause–effect structure* by considering all its subsets and fully *unfolding* its cause–effect power.

Usually, causal models are conceived in holistic terms, as state transitions of the system as a whole (1), or in reductionist terms, as a description of the individual units of the system and their interactions (2) [29]. However, to account for the structure of experience, considering only the cause–effect power of the individual units or of the system as a whole would be insufficient [17, 29]. Instead, by the composition postulate, we have to evaluate the system’s cause–effect structure by considering the cause–effect power of its subsets as well as their causal relations.

To contribute to the cause–effect structure of a complex, a system subset must both take *and* make a difference (as required by existence) *within* the system (as required by intrinsicity). A subset $M \subseteq S$ in state $m \in \Omega_M$ is called a *mechanism* if it *links* a cause and effect state over subsets of units $Z_{ce} \subseteq S$, called *purviews*. A mechanism together with the cause and effect state it specifies is called a *causal distinction*. Distinctions are evaluated based on whether they satisfy all the postulates of IIT (except for composition). For every mechanism, the cause–effect state is the one having maximal intrinsic information (ii), and the cause and effect purviews are those yielding the maximum value of integrated information (φ_d) within the complex—that is, those that are maximally irreducible. By the information postulate, the cause–effect power of a complex must be specific, which means that it selects a specific cause–effect state at the system level. Consequently, the distinctions that exist for the complex are only those whose cause–effect state

is congruent with the cause–effect state of the complex as a whole (incongruent distinctions are not components of the complex and its specific cause–effect power because they would violate the specificity postulate, according to which the experience can only be “this one”).

Distinctions whose cause or effect states overlap congruently within the system (over the same subset of units in the same state) are *bound* together by *causal relations*. Relations also have an associated value of integrated information (φ_r), corresponding to their irreducibility.

Together, these distinctions and relations compose the *cause–effect structure* of the complex in its current state. The cause–effect structure specified by a complex is called a Φ -*structure*. The sum of its distinction and relation integrated information amounts to the structure integrated information (Φ) of the complex.

In the following, we will provide a formal account of the IIT analysis. The first part demonstrates how to identify complexes. This requires that we (a) determine the cause–effect state of a system in its current state, (b) evaluate the system integrated information (φ_s) over that cause–effect state, and (c) search iteratively for maxima of integrated information (φ_s^*) within a universe. The second part describes how the postulates of IIT are applied to unfold the cause–effect structure of a complex. This requires that we identify the causal distinctions specified by subsets of units within the complex and the causal relations determined by the way distinctions overlap, yielding the system’s Φ -structure and its structure integrated information (Φ).

Box 2. Ontological principles of IIT

Principle of being

The *principle of being* states that *to be is to have cause–effect power*. In other words, in physical, operational terms, to exist requires being able to take and make a difference. The principle is closely related to the so-called Eleatic principle, as found in Plato’s Sophist dialogue [30]: “I say that everything possessing any kind of power, either to do anything to something else, or to be affected to the smallest extent by the slightest cause, even on a single occasion, has real existence: for I claim that entities are nothing else but power.” A similar principle can be found in the work of the Buddhist philosopher Dharmakīrti: “Whatever has causal powers, that really exists.” [31] Note that the Eleatic principle is enunciated as a disjunction (either to do something. . . or to be affected. . .), whereas IIT’s principle of being is presented as a conjunction (take *and* make a difference).

Principle of maximal existence

The *principle of maximal existence* states that, when it comes to a requirement for existence, *what exists is what exists the most*. The principle is offered by IIT as a good explanation for why the system state specified by the complex and the cause–effect states specified by its mechanisms are what they are. It also provides a criterion for determining the set of units constituting a complex—the one with maximally irreducible cause–effect power—for determining the subsets of units constituting the distinctions and relations that compose its cause–effect structure, and for determining the units’ grain. To exemplify, consider a set of candidate complexes overlapping over the same substrate. By the postulates of integration and exclusion, a complex must be both unitary and definite. By the maximal existence principle, the complex should be the one that lays the greatest claim to existence as *one* entity, as measured by system integrated information

(φ_s). For the same reason, candidate complexes that overlap over the same substrate but have a lower value of φ_s are excluded from existence. In other words, if having maximal φ_s is the reason for assigning existence as a unitary complex to a set of units, it is also the reason to exclude from existence any overlapping set not having maximal φ_s .

Principle of minimal existence

Another key principle of IIT is the *principle of minimal existence*, which complements that of maximal existence. The principle states that, when it comes to a requirement for existence, *nothing exists more than the least it exists*. The principle is offered by IIT as a good explanation for why, given that a system can only exist as one system if it is irreducible, its degree of irreducibility should be assessed over the partition across which it is least irreducible (the minimum partition). Similarly, a distinction within a system can only exist as one distinction to the extent that it is irreducible, and its degree of irreducibility should be assessed over the partition across which it is least irreducible. Moreover, a set of units can only exist as a system, or as a distinction within the system, if it specifies both an irreducible cause and an irreducible effect, so its degree of irreducibility should be the minimum between the irreducibility on the cause side and on the effect side (see (9) in [S1 Notes](#)).

Identifying substrates of consciousness

Our starting point is a substrate U in current state u with TPM \mathcal{T}_U (1). We consider any subset $s \subseteq u$ as a possible complex and refer to a set of units $S \subseteq U$ as a candidate system. (Note that s and u are sets of tuples containing both the units and their states.)

By the intrinsicality postulate, the units $W = U \setminus S$ are background conditions, and do not contribute directly to the cause–effect power of the system. To discount the contribution of background units, they are *causally marginalized*, conditional on the current state of the universe. This means that the background units are marginalized based on a uniform marginal distribution, updated by conditioning on u . The process is repeated separately for each unit in the system, and they are then combined using a product (in line with conditional independence), which eliminates any residual correlations due to the background units. Accordingly, we obtain two TPMs \mathcal{T}_e and \mathcal{T}_c (for evaluating effects and causes, respectively) for the candidate system S . For evaluating effects, the state of the background units is fully determined by the current state of the universe. The corresponding TPM, \mathcal{T}_e , is used to identify the effect of the current state:

$$\mathcal{T}_e = \mathcal{T}_e(\mathcal{T}_U, u, w) \equiv p_e(\bar{s} | s) = p(\bar{s} | s, w), \quad s, \bar{s} \in \Omega_S, \tag{3}$$

where $w = u \setminus s$. For evaluating causes, knowledge of the current state is used to compute the probability distribution over potential prior states of the background units, which is not necessarily uniform or deterministic. The corresponding TPM, \mathcal{T}_c , is used to evaluate the cause of the current state:

$$\mathcal{T}_c = \mathcal{T}_c(\mathcal{T}_U, u, w) \equiv p_c(s | \bar{s}) = \prod_{i=1}^{|\bar{s}|} \sum_{\bar{w}} p(s_i | \bar{s}, \bar{w}) \left(\frac{\sum_{\hat{s}} p(u | \hat{s}, \bar{w})}{\sum_{\hat{u}} p(u | \hat{u})} \right), \quad s, \bar{s} \in \Omega_S. \tag{4}$$

In both TPMs, the background units W are rendered causally inert, so that causes and effects are evaluated from the intrinsic perspective of the system.

The intrinsic information $ii_{c/e}$ is a measure of the intrinsic cause or effect power exerted by a system S in its current state s over itself by selecting a specific cause or effect state \bar{s} . The cause–effect state for which intrinsic information (ii_c and ii_e) is maximal is called the maximal cause–effect state $s' = \{s'_c, s'_e\}$. The integrated information φ_s is a measure of the irreducibility of a cause–effect state, compared to the directional system partition θ that affects the maximal cause–effect state the least (minimum partition, or MIP). Systems for which integrated information is maximal (φ_s^*) compared to any competing candidate system with overlapping units are called maximal substrates, or complexes.

The IIT 4.0 formalism to measure a system’s integrated information φ_s and to identify maximal substrates was first presented in [13]. An example of how to identify complexes in a simple system is given in Fig 1, while a comparison with prior accounts (IIT 1.0, IIT 2.0, and IIT 3.0) can be found in S2 Text. An outline of the IIT algorithm is included in S1 Fig.

Existence, intrinsicity, and information: Determining the maximal cause–effect state of a candidate system

Given a causal model with corresponding TPMs \mathcal{T}_e (3) and \mathcal{T}_c (4), we wish to identify the maximal cause–effect state specified by a system in its current state over itself and to quantify the causal power with which it does so. In this way, we quantify the cause–effect power of a system from its intrinsic perspective, rather than from the perspective of an outside observer (see Box 1).

System intrinsic information ii . Intrinsic information $ii(s, \bar{s})$ measures the causal power of a system S over itself, for its current state s , over a specific cause or effect state \bar{s} . Intrinsic information depends on interventional conditional probabilities and unconstrained probabilities of cause or effect states and is the product of selectivity and informativeness.

On the effect side, intrinsic effect information ii_e of the current state s over a possible effect state \bar{s} is defined as:

$$ii_e(s, \bar{s}) = p_e(\bar{s} | s) \log \left(\frac{p_e(\bar{s} | s)}{p_e(\bar{s})} \right), \tag{5}$$

where $p_e(\bar{s} | s)$ (3) is the interventional conditional probability that the current state s produces the effect state \bar{s} , as indicated by \mathcal{T}_e .

The interventional unconstrained probability $p_e(\bar{s})$

$$p_e(\bar{s}) = |\Omega_s|^{-1} \sum_{s \in \Omega_s} p_e(\bar{s} | s), \tag{6}$$

is defined as the marginal probability of \bar{s} , averaged across all possible current states of S with equal probability (where $|\Omega_s|$ denotes the cardinality of the state space Ω_s).

On the cause side, intrinsic cause information ii_c of the current state s over a possible cause state \bar{s} is defined as:

$$ii_c(s, \bar{s}) = p_c^-(\bar{s} | s) \log \left(\frac{p_c(s | \bar{s})}{p_c(s)} \right), \tag{7}$$

where $p_c(s, \bar{s})$ (4) is the interventional conditional probability that the cause state \bar{s} produces the current state s , as indicated by \mathcal{T}_c , and the interventional unconstrained probability is again defined as the marginal probability of s , averaged across all possible cause states of S with

equal probability,

$$p_c(s) = |\Omega_s|^{-1} \sum_{\bar{s} \in \Omega_s} p_c(s | \bar{s}). \tag{8}$$

Moreover, $p_c^-(\bar{s} | s)$ (4) is the interventional conditional probability that the current state $s \in \Omega_s$ was produced by \bar{s} ; it is derived from \mathcal{T}_c using Bayes' rule, where we again assign a uniform prior to the possible cause states \bar{s} ,

$$p_c^-(\bar{s} | s) = \frac{p_c(s | \bar{s}) \cdot |\Omega_s|^{-1}}{p_c(s)} = \frac{p_c(s | \bar{s})}{\sum_{\hat{s} \in \Omega_s} p_c(s | \hat{s})}. \tag{9}$$

Informativeness (over chance). In (5) and (7), the logarithmic term (in base 2 throughout) is called *informativeness*. Note that informativeness is expressed in terms of ‘forward’ probabilities (probability of a subsequent state given the current state) for both ii_e (5) and ii_c (7). However, ii_e (5) evaluates the increase in probability of the effect state due to the current state based on \mathcal{T}_e , while ii_c (7) evaluates the increase in probability of the current state due to the cause state based on \mathcal{T}_c .

In line with the existence postulate, a system S in state s has cause–effect power (it takes and makes a difference) if it raises the probability of a possible effect state compared to chance, which is to say compared to its unconstrained probability,

$$\log \left(\frac{p_e(\bar{s} | s)}{p_e(\bar{s})} \right) > 0, \tag{10}$$

and if the probability of the current state is raised above chance by a possible cause state,

$$\log \left(\frac{p_c(s | \bar{s})}{p_c(s)} \right) > 0. \tag{11}$$

Informativeness is additive over the number of units: if a system specifies a cause or effect state with probability $p = 1$, its causal power increases additively with the number of units whose states it fully specifies (*expansion*), given that the chance probability of all states decreases exponentially.

Selectivity (over states). From the intrinsic perspective of a system, cause–effect power over a specific cause or effect state depends not only on the deviation from chance it produces, but also on how its probability is concentrated on that state, rather than being diluted over other states. This is measured by the *selectivity* term in front of the logarithmic term in (5) and (7), corresponding to the conditional probability $p_c^-(\bar{s} | s)$ or $p_e(\bar{s} | s)$ of that specific cause or effect state. (Note that here, on the cause side, we use the ‘backward’ probability (probability of a prior state given the current state) obtained through Bayes’ rule, while we use the ‘forward’ probability of the effect state \bar{s} given s on the effect side.) Selectivity means that if $p < 1$, the system’s causal power becomes subadditive (*dilution*) (see [14] for details). For example, as shown in [12], if an unconstrained unit is added to a fully specified unit, intrinsic information does not just stay the same, but decreases exponentially. From the intrinsic perspective of the system, the informativeness of a specific cause or effect state is diluted because it is spread over multiple possible states, yet the system must select only one state.

Altogether, taking the product of informativeness and selectivity leads to a tension between expansion and dilution: a larger system will tend to have higher informativeness than a smaller

system because it will deviate more from chance, but it will also tend to have lower selectivity because it will have a larger repertoire of states to select from.

Because of the selectivity term, intrinsic information is reduced by indeterminism and degeneracy. As shown in [13], indeterminism decreases the probability of the selected effect state because it implies that the same state can lead to multiple states. In turn, degeneracy decreases the probability of the selected cause state because it implies that multiple states can lead to the same state, even in a deterministic system.

The intrinsic information ii is quantified in units of *intrinsic bits*, or *ibits*, to distinguish it from standard information-theoretic measures (which are typically additive). Formally, the *ibit* corresponds to a point-wise information value (measured in bits) weighted by a probability.

The maximal cause–effect state. Taking the product of informativeness and selectivity on the system’s cause and effect sides captures the postulates of existence (taking and making a difference) and intrinsicity (taking and making a difference over itself) for each possible cause or effect state, as measured by intrinsic information. However, the information postulate further requires that the system selects a specific cause or effect state. The selection is determined by the principle of maximal existence (Box 1): the cause or effect specified by the system should be the one that maximizes intrinsic information. On the effect side (and similarly for the cause side, see S1 Fig),

$$\begin{aligned}
 s'_e(\mathcal{T}_e, s) &= \operatorname{argmax}_{\bar{s} \in \Omega_S} ii_e(s, \bar{s}) \\
 &= \operatorname{argmax}_{\bar{s} \in \Omega_S} p_e(\bar{s} | s) \log \left(\frac{p_e(\bar{s} | s)}{p_e(\bar{s})} \right).
 \end{aligned}
 \tag{12}$$

The system’s intrinsic effect information is the value of ii_e (5) for its maximal effect state:

$$ii_e(\mathcal{T}_e, s) := ii_e(s, s'_e) = \max_{\bar{s} \in \Omega_S} p_e(\bar{s} | s) \log \left(\frac{p_e(\bar{s} | s)}{p_e(\bar{s})} \right).
 \tag{13}$$

We have made the dependency of s' and ii_e on \mathcal{T}_e explicit in (12) and (13) to highlight that, for intrinsic information to properly assess cause–effect power, all probabilities must be derived from the system’s interventional transition probability function, while imposing a uniform prior distribution over all possible system states. If $ii_e(\mathcal{T}_e, s) = 0$, the system S in state s has no causal power. This is the case if and only if $p_e(\bar{s} | s) = p_e(\bar{s})$ for every \bar{s} [14] (and likewise, it can be shown that $ii_c(\mathcal{T}_c, s) = 0$ if and only if $p_c(s | \bar{s}) = p_c(s)$ for every \bar{s} .) It is worthwhile to mention that when $ii_e(\mathcal{T}_e, s) \neq 0$, the system state s always increases the probability of the intrinsic effect state compared to chance. Similarly, when $ii_c(\mathcal{T}_c, s) \neq 0$ the intrinsic cause state increases the probability of the system state, satisfying (11). Note also that a system’s intrinsic cause–effect state does not necessarily correspond to the actual cause and effect states (what actually happened before / will happen after) in the dynamical evolution of the system, which typically also depends on extrinsic influences. (For an account of actual causation according to the causal principles of IIT, see [10].)

Intrinsic difference. Because consciousness is the way it is, the formulation of its properties in physical, operational terms should be unique and based on quantities that uniquely satisfy the postulates [12, 32]. Intrinsic information is formulated as a product of selectivity and informativeness based on the notion of intrinsic difference (ID) [14]. This is a measure of the difference between two probability distributions which uniquely satisfies three properties (causality, intrinsicity, and specificity) that align with the postulates of IIT (but also have independent justification):

causality (Existence): the measure is zero if and only if the system does not make a difference

intrinsicity (Intrinsicity): the measure increases if the system is expanded without noise (expansion) and decreases if the system is expanded without signal (dilution)

specificity (Information): the measure reflects the cause–effect power of a specific state over a specific cause and effect state.

The properties uniquely satisfied by the ID are described in a general mathematical context in [14], as well as some additional discussion in S2 Text.

Note that, on the effect side, ii_e is formally equivalent to the ID between the constrained effect repertoire $p_e(\bar{s} | s)$ and the unconstrained effect repertoire $p_e(\bar{s})$. On the cause side, the application of Bayes rule to compute $p_c^-(\bar{s} | s)$ as the selectivity term means that ii_c is not strictly equivalent to the ID between two probability distributions. However, analogously to the effect formulation, it is defined as the product of selectivity and informativeness of causes.

Integration: Determining the irreducibility of a candidate system

Having identified the maximal cause–effect state $s' = \{s'_c, s'_e\}$ of a candidate system S in its current state s , the next step is to evaluate whether the system specifies the cause–effect state of its units in a way that is *irreducible*, as required by the integration postulate: a candidate system can only be a substrate of consciousness if it is *one* system—that is, if it cannot be subdivided into subsets of units that exist separately from one another.

Directional system partitions. To that end, we define a set of *directional* system partitions $\Theta(S)$ that divide S into $k \geq 2$ parts $\{S^{(i)}\}_{i=1}^k$, such that

$$S^{(i)} \neq \emptyset, S^{(i)} \cap S^{(j)} = \emptyset, \text{ and } \bigcup_{i=1}^k S^{(i)} = S. \tag{14}$$

In words, each part $S^{(i)}$ must contain at least one unit, there must be no overlap between any two parts $S^{(i)}$ and $S^{(j)}$, and every unit of the system must appear in exactly one part. For each part $S^{(i)}$, the partition removes the causal connections of that part with the rest of the system in a directional manner: either the part’s inputs, outputs, or both are replaced by independent “noise” (they are “cut” by the partition in the sense that their causal powers are substituted by chance). Directional partitions are necessary because, from the intrinsic perspective of a system, a subset of units that cannot affect the rest of the system, or cannot be affected by it, cannot truly be a part of the system. In other words, to be a part of a system, a subset of units must be able to interact with the rest of the system in both directions (*cause and effect*).

A partition $\theta \in \Theta(S)$ thus has the form

$$\theta = \{S_{\delta_1}^{(1)}, S_{\delta_2}^{(2)}, \dots, S_{\delta_k}^{(k)}\}, \tag{15}$$

where $\delta_i \in \{\leftarrow, \rightarrow, \leftrightarrow\}$ indicates whether the inputs (\leftarrow), outputs (\rightarrow), or both (\leftrightarrow) are cut for a given part. For each part $S^{(i)}$, we can then identify a set of units $X^{(i)} \subseteq S$ whose inputs to $S^{(i)}$ have been cut by the partition, and the complementary set $Y^{(i)} = S \setminus X^{(i)}$ whose inputs to $S^{(i)}$ are

left intact. Specifically,

$$X^{(i)} = \begin{cases} S \setminus S^{(i)} & \text{if } \delta_i \in \{\leftarrow, \leftrightarrow\} \\ \bigcup_{\substack{j \neq i: \\ \delta_j \in \{\rightarrow, \leftrightarrow\}}} S^{(j)} & \text{if } \delta_i \in \{\rightarrow\}. \end{cases} \tag{16}$$

In the first case, if $\delta_i \in \{\leftarrow, \leftrightarrow\}$, all inputs to $S^{(i)}$ from $S \setminus S^{(i)}$ are cut. In the second case, if $\delta_i \in \{\rightarrow\}$, there may still be inputs to $S^{(i)}$ that are cut, which correspond to the outputs of all $S^{(j)}$ with $\delta_j \in \{\rightarrow, \leftrightarrow\}$.

Given a partition $\theta \in \Theta(S)$, we define partitioned transition probability matrices \mathcal{T}_e^θ and \mathcal{T}_c^θ in which all connections affected by the partition are “noised.” This is done by combining the independent contributions of each unit $S_j \in S$ in line with the conditional independence assumption (2). For the effect TPM (and analogously for the cause TPM)

$$\mathcal{T}_e^\theta \equiv p_e^\theta(\bar{s} | s) = \prod_{j=1}^n p_e^\theta(\bar{s}_j | s), \quad \bar{s}, s \in \Omega_S, \tag{17}$$

where the partitioned probability of a unit $S_j \in S^{(i)}$ is defined as

$$p_e^\theta(\bar{s}_j | s) = |\Omega_{X^{(i)}}|^{-1} \sum_{x^{(i)} \in \Omega_{X^{(i)}}} p_e(\bar{s}_j | x^{(i)}, y^{(i)}), \tag{18}$$

and $y^{(i)} = s \setminus x^{(i)}$. This means that all connections to unit S_j that are affected by the partition are *causally marginalized* (replaced by independent noise).

System integrated information φ_s . The integrated effect information φ_e measures how much the partition $\theta \in \Theta_S$ reduces the probability with which a system S in state $s \in \Omega_S$ specifies its effect state s'_e (12),

$$\varphi_e(\mathcal{T}_e, s, \theta) = p_e(s'_e | s) \left| \log \left(\frac{p_e(s'_e | s)}{p_e^\theta(s'_e | s)} \right) \right|_+. \tag{19}$$

Note that φ_e has the same form as the intrinsic information $ii_e(s, \bar{s})$ (5), with the partitioned effect probability taking the place of the unconstrained (marginal) probability. Here, $|\cdot|_+$ represents the positive part operator, which sets the negative values to 0. This ensures that the system as a whole raises the probability of the effect state compared to the partitioned probability. Likewise, the integrated cause information φ_c is defined as

$$\varphi_c(\mathcal{T}_c, s, \theta) = p_c^-(s'_c | s) \left| \log \left(\frac{p_c(s | s'_c)}{p_c^\theta(s | s'_c)} \right) \right|_+. \tag{20}$$

(By the principle of maximal existence, if two or more cause–effect states are tied for maximal intrinsic information, the system specifies the one that maximizes $\varphi_{c/e}$).

By the zeroth postulate, existence requires cause *and* effect power, and the integration postulate requires that its cause–effect power be irreducible. By the principle of minimal existence (Box 2), then, system integrated information for a given partition is the minimum of its irreducibility on the cause *and* effect sides:

$$\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s, \theta) = \min\{\varphi_c(\mathcal{T}_c, s, \theta), \varphi_e(\mathcal{T}_e, s, \theta)\}. \tag{21}$$

Moreover, again by the principle of minimal existence, the integrated information of a system is given by its irreducibility over its minimum partition (MIP) $\theta \in \Theta_S$, such that

$$\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s) := \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s, \theta'). \tag{22}$$

The MIP is defined as the partition $\theta \in \Theta_S$ that minimizes the system’s integrated information, relative to the maximum possible value it could take for arbitrary TPMs $\mathcal{T}'_e, \mathcal{T}'_c$ over the units of system S

$$\theta' = \operatorname{argmin}_{\theta \in \Theta(S)} \frac{\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s, \theta)}{\max_{\mathcal{T}'_e, \mathcal{T}'_c} \varphi_s(\mathcal{T}'_e, \mathcal{T}'_c, s, \theta)}. \tag{23}$$

Accordingly, the system is reducible if at least one partition $\theta \in \Theta_S$ makes no difference to the cause or effect probability. The normalization term in the denominator of (23) ensures that $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s)$ is evaluated fairly over a system’s fault lines by assessing integration relative to its maximum possible value over a given partition. Using the *relative* integrated information quantifies the strength of the interactions between parts in a way that does not depend on the number of parts and their size. As proven in [13], the maximal value of $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s, \theta)$ for a given partition θ is the normalization factor $\max_{\mathcal{T}'_e, \mathcal{T}'_c} \varphi_s(\mathcal{T}'_e, \mathcal{T}'_c, s, \theta) = \sum_{i=1}^k |S^{(i)}| |X^{(i)}|$, which corresponds to the maximal possible number of “connections” (pairwise interactions) affected by θ . For example, as shown in [13], the MIP will correctly identify the fault line dividing a system into two large subsets of units linked through a few interconnected units (a “bridge”), rather than defaulting to partitions between individual units and the rest of the system. Once the minimum partition has been identified, the integrated information across it is an *absolute* quantity, quantifying the loss of intrinsic information due to cutting the minimum partition of the system. (If two or more partitions $\theta \in \Theta(S)$ minimize Eq (23), we select the partition with the largest unnormalized φ_s value as θ' , applying the principle of maximal existence.) Defining θ' as in (23), moreover, ensures that $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s) = 0$ if the system is not *strongly connected* in graph-theoretic terms (see (10) in S1 Notes).

In summary, the system integrated information ($\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s)$, also called ‘*small phi*’, quantifies the extent to which system S in state s has cause–effect power over itself as *one* system (i.e., irreducibly). $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s)$ is thus a quantifier of irreducible existence.

Exclusion: Determining maximal substrates (complexes)

In general, multiple candidate systems with overlapping units may have positive values of $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s)$. By the exclusion postulate, the substrate of consciousness must be definite; that is, it must comprise a definite set of units. But which one? Once again, we employ the principle of maximal existence (Box 2): among candidate systems competing over the same substrate with respect to an essential requirement for existence, in this case irreducibility, the one that exists is the one that exists the most. Accordingly, the maximal substrate, or complex, is the candidate substrate with the maximum value of system integrated information (φ_s^*), and overlapping substrates with lower φ_s are thus excluded from existence.

Determining maximal substrates recursively. Within a universal substrate U_0 in state u_0 , subsets of units that specify maxima of irreducible cause–effect power (complexes) can be identified iteratively: the substrate with maximum φ_s^* is identified as a complex, the corresponding units are excluded from further consideration, the remaining units are searched for the next maximal substrate. Formally, an iterative search is performed to find a sequence of

systems $S_k^* \subseteq U_k$ with

$$\varphi_s^*(\mathcal{T}_e, \mathcal{T}_c, u_k) = \max_{S \subseteq U_k} \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s), \tag{24}$$

such that

$$S_k^* = \operatorname{argmax}_{S \subseteq U_k} \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s), \tag{25}$$

and $U_{k+1} = U_k \setminus S_k^*$ until $U_{k+1} = \emptyset$ or $U_{k+1} = U_k$ (the units in $U_0 \setminus U_{k+1}$ still serve as background conditions, for details see [13]). If the maximal substrate S_k^* is not unique, and all tied systems overlap, the next best system that is unique is chosen instead (see S1 Text).

For any complex S^* in its corresponding state $s^* \in \Omega_{S^*}$, overlapping substrates that specify less integrated information ($\varphi_s < \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s^*)$) are excluded. Consequently, specifying a maximum of integrated information φ_s^* compared to all overlapping systems

$$S \cap \tilde{S} \neq \emptyset \Rightarrow \varphi_s(s) > \varphi_s(\tilde{s}), \quad \forall S \neq \tilde{S} \subseteq U \tag{26}$$

is a sufficient requirement for a system $S \subseteq U$ to be a complex.

As described in [13], this recursive search for maximal substrates “condenses” the universe U_0 in state $u_0 \in \Omega_{U_0}$ into a disjoint (non-overlapping) and exhaustive set of complexes—the first complex, second complex, and so on.

Determining maximal unit grains. Above, we presented how to determine the borders of a complex within a larger system U , assuming a particular grain for the units $U_i \in U$. In principle, however, all possible grains should be considered [33, 34]. In the brain, for example, the grain of units could be brain regions, groups of neurons, individual neurons, sub-cellular structures, molecules, atoms, quarks, or anything finer, down to hypothetical atomic units of cause–effect power [3, 4]. For any unit grain—neurons, for example—the grain of updates could be minutes, seconds, milliseconds, micro-seconds, and so on. However, by the exclusion postulate, the units that constitute a system S must also be definite, in the sense of having a definite grain.

Once again, the grain is defined by the principle of maximal existence: across the possible micro- and macroscopic levels, the “winning” grain is the one that ensures maximally irreducible existence (φ_s^*) for the entity to which the units belong [33, 34].

To evaluate integrated information across grains requires a mathematical framework for defining coarser (macro) units from finer (micro) units. Such a framework has been developed in previous work [33–35], and is updated here to fully align with the postulates.

Supposing that $U = u$ is a universe of micro units in a state, a macro unit $J = j$ is a combination of a set of micro units $\hat{S} \subseteq U$, and a mapping g from the state \hat{S} to the state of J ,

$$j = g(\hat{s}),$$

where

$$g : \Omega_{\hat{s}} \rightarrow \Omega_j.$$

As constituents of a complex upon which its cause–effect power rests, the units themselves should comply with the postulates of IIT. Otherwise it would be possible to “make something out of nothing.” Accordingly, units themselves must also be maximally irreducible, as measured by the integrated information of the units when they are treated as candidate systems (φ_s); otherwise, they would not be units but “disintegrate” into their constituents. However, in contrast to systems, units only need to be maximally irreducible *within*, because they do not

exist as complexes in their own right: a unit J with substrate \hat{S} qualifies as a candidate unit of a larger system S if its integrated information when treated as a candidate system (φ_s) is higher than that of any system of units (including potential macro units) that can be defined using a subset of \hat{S} . Out of all possible sets of such candidate units, the set of (macro) units that define a complex is the one that maximizes the existence of the complex to which the units belong, rather than their own existence.

In practice, the search for the maximal grain should be an iterative process, starting from micro units: identify potential substrates for macro units (\hat{S}) that are maximally irreducible within, identify mappings g that maximize the integrated information of systems of macro units, then consider additional potential substrates for macro units, and so on iteratively, until a global maximum is found. The iterative approach is necessary for establishing that a substrate is maximally irreducible within, as this criterion requires consideration not only of micro units, but also of all finer grains (potential meso units defined from subsets of \hat{S}).

Here we outlined an overall framework for identifying macro units consistent with the postulates. Additional details about the nature of the mapping g , and how to derive the transition probabilities for a system of macro units are also informed by the postulates (see (11) in [S1 Notes](#)).

Unfolding the cause–effect structure of a complex

Once a maximal substrate and the associated maximal cause–effect state have been identified, we must unfold its cause–effect power to reveal its cause–effect structure of distinctions and relations, in line with the composition postulate. As components of the cause–effect structure, distinctions and relations must also satisfy the postulates of IIT (save for composition).

Composition and causal distinctions

Causal distinctions capture how the cause–effect power of a substrate is structured by subsets of units that specify irreducible causes and effects over subsets of its units. A candidate distinction $d(m)$ consists of (1) a mechanism $M \subseteq S$ in state $m \in \Omega_M$ inherited from the system state $s \in \Omega_S$; (2) a maximal cause–effect state $z^* = \{z_c^*, z_e^*\}$ over the cause and effect purviews ($Z_c, Z_e \subseteq S$) linked by the mechanism; and (3) an associated value of irreducibility ($\varphi_d > 0$). A distinction $d(m)$ is thus represented by the tuple

$$d(m) = (m, z^*, \varphi_d). \quad (27)$$

For a given mechanism m , our goal is to identify its maximal cause Z_c^* in state $z_c^* \in \Omega_{Z_c^*}$ and its maximal effect Z_e^* in state $z_e^* \in \Omega_{Z_e^*}$ within the system, where $Z_c^*, Z_e^* \subseteq S$.

As above, in line with existence, intrinsicality, and information, we determine the maximal cause or effect state specified by the mechanism over a candidate purview within the system based on the value of intrinsic information $ii(m, z)$. Next, in line with integration, we determine the value of integrated information $\varphi_d(m, Z, \theta)$ over the minimum partition θ . In line with exclusion, we determine the maximal cause–effect purviews for that mechanism over all possible purviews $Z \subseteq S$ based on the associated value of irreducibility $\varphi_d(m, Z, \theta)$. Finally, we determine whether the maximal cause–effect state specified by the mechanism is congruent with the system’s overall cause–effect state ($z_c^* \subseteq s_c^*, z_e^* \subseteq s_e^*$), in which case we conclude that it contributes a distinction to the overall cause–effect structure.

The updated formalism to identify causal distinctions within a system S in state s was first presented in [12]. Here we provide a summary with minor adjustments on selecting z_c^* and z_e^* ,

the cause integrated information $\varphi_c(m, Z)$, and the requirement that causal distinctions must be congruent with the system’s maximal cause–effect state (see S2 Text).

Existence, intrinsicity, and information: Determining the cause and effect state specified by a mechanism over candidate purviews. Like the system as a whole, its subsets must comply with existence, intrinsicity, and information. As for the system, we begin by quantifying, in probabilistic terms, the difference a subset of units $M \subseteq S$ in its current state $m \subseteq s$ takes and makes from and to subsets of units $Z \subseteq S$ (cause and effect purview). As above, we start by establishing the interventional conditional probabilities and unconstrained probabilities from the TPMs \mathcal{T}_c and \mathcal{T}_e .

When dealing with a mechanism constituted by a subset of system units, it is important to capture the constraints on a purview state z that are exclusively due to the mechanism in its state (m), removing any potential contribution from other system units. This is done by causally marginalizing all variables in $X = S \setminus M$, which corresponds to imposing a uniform distribution as $p(X)$ [8, 10, 12] (see (12) in S1 Notes). The effect probability of a single unit $Z_i \in Z$ conditioned on the current state m is thus defined as

$$p_e(z_i | m) = |\Omega_X|^{-1} \sum_{x \in \Omega_X} p(z_i | m, x), \quad z_i \in \Omega_{Z_i}. \tag{28}$$

In addition, product probabilities $\pi(z|m)$ are used instead of conditional probabilities $p_e(z|m)$ to discount correlations from units in $X = S \setminus M$ with divergent outputs to multiple units in $Z \subseteq S$ [8, 10, 36]. Otherwise, X might introduce correlations in Z that would be wrongly considered as effects of M . Based on the appropriate TPM, the probability over a set Z of $|Z|$ units is thus defined as the product of the probabilities over individual units

$$\pi_e(z | m) = \prod_{i=1}^{|Z|} p_e(z_i | m), \quad z \in \Omega_Z, \tag{29}$$

and

$$\pi_c(m | z) = \prod_{i=1}^{|M|} p_c(m_i | z), \quad m \in \Omega_M. \tag{30}$$

Note that for a single unit purview $\pi_e(z|m) = p_e(z|m)$, and for a single unit mechanism $\pi_c(m|z) = p_c(m|z)$. By using product probabilities, causal marginalization maintains the conditional independence between units (2) because independent noise is applied to individual connections. The assumption of conditional independence distinguishes IIT’s causal powers analysis from standard information-theoretic analyses of information flow [10, 27] and corresponds to an assumption that variables are “physical” units in the sense that they are irreducible within and can be observed and manipulated independently.

From Eqs (29) and (30) we can also define unconstrained probabilities

$$\pi_e(z; M) = |\Omega_M|^{-1} \sum_{m \in \Omega_M} \pi_e(z | m), \quad z \in \Omega_Z, \tag{31}$$

and

$$\pi_c(m; Z) = |\Omega_Z|^{-1} \sum_{z \in \Omega_Z} \pi_c(m | z), \quad m \in \Omega_M. \tag{32}$$

Given the set $Y = S \setminus Z$, the backward cause probability (selectivity) for a mechanism m with $|M|$ units is computed using Bayes' rule over the product distributions

$$\pi_c^-(z | m) = \frac{\pi_c(m | z) \cdot |\Omega_Z|^{-1}}{\pi_c(m; Z)} = \frac{\prod_{i=1}^{|M|} p_c(m_i | z)}{\sum_{\hat{z} \in \Omega_Z} \prod_{i=1}^{|M|} p_c(m_i | \hat{z})}, \quad z \in \Omega_Z, \quad (33)$$

where $p_c(m_i | z) = |\Omega_Y|^{-1} \sum_{y \in \Omega_Y} p_c(m_i | z, y)$ in line with (28).

To correctly quantify intrinsic causal constraints, the marginal probability of possible cause states (for computing $\pi_c^-(z | m)$ or $\pi_c(m; Z)$) is again set to the uniform distribution. As above, all probabilities are obtained from the TPMs T_e (3) and T_c (4) and thus correspond to *interventional* probabilities throughout.

Having defined cause and effect probabilities, we can now evaluate the intrinsic information of a mechanism m over a purview state $z \in \Omega_Z$ analogously to the system intrinsic information (5) and (7). The intrinsic effect information that a mechanism in a state m specifies about a purview state z is

$$ii_e(m, z) = \pi_e(z | m) \log \left(\frac{\pi_e(z | m)}{\pi_e(z; M)} \right). \quad (34)$$

The intrinsic cause information that a mechanism in a state m specifies about a purview state z is

$$ii_c(m, z) = \pi_c^-(z | m) \log \left(\frac{\pi_c(m | z)}{\pi_c(m; Z)} \right). \quad (35)$$

As with system intrinsic information, the logarithmic term is the informativeness, which captures how much causal power is exerted by the mechanism m on its potential effect z (how much it increases the probability of that state above chance), or by the potential cause z on the mechanism m . The term in front of the logarithm corresponds to the mechanism's selectivity, which captures how much the causal power of the mechanism m is concentrated on a specific state of its purview (as opposed to other states). In the following we will again focus on the effect side, but an equivalent procedure applies on the cause side (see S1 Fig).

Based on the principle of maximal existence, the maximal effect state of m within the purview Z is defined as

$$z'_e(m, Z) = \operatorname{argmax}_{z \in \Omega_Z} ii_e(m, z), \quad (36)$$

which corresponds to the specific effect of m on Z . Note that z'_e is not always unique (see S1 Text). The maximal intrinsic information of mechanism m over a purview Z is then

$$ii_e(m, Z) := ii_e(m, z'_e) = \max_{z \in \Omega_Z} ii_e(m, z). \quad (37)$$

Note that, by this definition, if $ii_e(m, Z) \neq 0$, mechanism m always raises the probability of its maximal effect state compared to the unconstrained probability. This is because there is at least one state $z \in \Omega_Z$ such that $\pi_e(z|m) > \pi_e(z; M)$.

The intrinsic information of a candidate distinction, like that of the system as a whole, is sensitive to indeterminism (the same state leading to multiple states) and degeneracy (multiple states leading to the same state) because both factors decrease the probability of the selected

state. Moreover, the product of selectivity and informativeness leads to a tension between expansion and dilution: larger purviews tend to increase informativeness because conditional probabilities will deviate more from chance, but they also tend to decrease selectivity because of the larger repertoire of states.

Integration: Determining the irreducibility of a candidate distinction. To comply with integration, we must next ask whether the specific effect of m on Z is irreducible. As for the system, we do so by evaluating the integrated information $\varphi_e(m, Z)$. To that end, we define a set of “disintegrating” partitions $\Theta(M, Z)$ as

$$\Theta(M, Z) = \left\{ \left\{ (M^{(i)}, Z^{(i)}) \right\}_{i=1}^k : k \in \{2, 3, 4, \dots\}, M^{(i)} \in \mathbb{P}(M), Z^{(i)} \in \mathbb{P}(Z), \right. \\ \left. \bigcup M^{(i)} = M, \bigcup Z^{(i)} = Z, Z^{(i)} \cap Z^{(j)} = M^{(i)} \cap M^{(j)} = \emptyset \forall i \neq j, M^{(i)} = M \Rightarrow Z^{(i)} = \emptyset \right\}, \tag{38}$$

where $\{M^{(i)}\}$ is a partition of M and $\{Z^{(i)}\}$ is a partition of Z , but the empty set may also be used as a part (\mathbb{P} denotes the power set). As introduced in [10, 12], a disintegrating partition $\theta \in \Theta(M, Z)$ either “cuts” the mechanism into at least two independent parts if $|M| > 1$, or it severs all connections between M and Z , which is always the case if $|M| = 1$ (we refer to [10, 12] for details). Note that disintegrating partitions differ from system partitions (23), which divide the system into two or more parts in a directed manner to evaluate whether and to what extent the system is integrated in terms of its cause–effect power. Instead, disintegrating partitions apply to mechanism–purview pairs within the system, which are already directed, to evaluate the cause or effect power specified by the mechanism over its purview.

Given a partition $\theta \in \Theta(M, Z)$, we can define the partitioned effect probability

$$\pi_e^\theta(z'_e | m) = \prod_{i=1}^k \pi_e(z_e^{(i)} | m^{(i)}), \tag{39}$$

with $\pi_e(\emptyset | m^{(i)}) = \pi_e(\emptyset) = 1$. In the case of $m^{(i)} = \emptyset$, $\pi_e(z_e^{(i)} | \emptyset)$ corresponds to the fully partitioned effect probability

$$\pi_e(z | \emptyset) = \prod_{i=1}^{|Z|} \sum_{s \in \Omega_S} p_e(z_i | s) |\Omega_S|^{-1}. \tag{40}$$

The integrated effect information of mechanism m over a purview $Z \subseteq S$ with effect state z'_e for a particular partition $\theta \in \Theta(M, Z)$ is then defined as

$$\varphi_e(m, Z, \theta) = \pi_e(z'_e | m) \left| \log \left(\frac{\pi_e(z'_e | m)}{\pi_e^\theta(z'_e | m)} \right) \right|_+. \tag{41}$$

The effect of m on z'_e is reducible if at least one partition $\theta \in \Theta(M, Z)$ makes no difference to the effect probability or increases it compared to the unpartitioned probability. In line with the principle of minimal existence, the total integrated effect information $\varphi_e(m, Z)$ again has to be evaluated over θ' , the minimum partition (MIP)

$$\varphi_e(m, Z) := \varphi_e(m, Z, \theta'), \tag{42}$$

which requires a search over all possible partitions $\theta \in \Theta(M, Z)$:

$$\theta' = \operatorname{argmin}_{\theta \in \Theta(M, Z)} \frac{\varphi_e(m, Z, \theta)}{\max_{\theta'} \varphi_e(m, Z, \theta)}. \tag{43}$$

As in (23), the minimum partition is evaluated against its maximum possible value across all possible systems TPMs \mathcal{T}' , which again corresponds to the number of possible pairwise interactions affected by the partition.

The integrated cause information is defined analogously, as

$$\varphi_c(m, Z) := \varphi_c(m, Z, \theta') = \pi_c^-(z'_c | m) \left| \log \left(\frac{\pi_c(m | z'_c)}{\pi_c^{\theta'}(m | z'_c)} \right) \right|_+, \tag{44}$$

where the partitioned probability $\pi_c^0(m | z)$ is again a product distribution over the parts in the partition, as in (39).

Taken together, the intrinsic information (37) determines what cause or effect state the mechanism m specifies. Its integrated information quantifies to what extent m specifies its cause or effect in an irreducible manner. Again, $\varphi(m, Z)$ is a quantifier of irreducible existence.

Exclusion: Determining causal distinctions. Finally, to comply with exclusion, a mechanism must select a definite effect purview, as well as a cause purview, out of a set of candidate purviews. Resorting again to the principle of maximal existence, the mechanism’s effect purview and associated effect is the one having the maximum value of integrated information across all possible purviews $Z \subseteq S$ in state $z'_c(m, Z)$ (36)

$$z_c^*(m) = \operatorname{argmax}_{Z \subseteq S} \varphi_e(m, z'_c(m, Z)). \tag{45}$$

The integrated effect information of a mechanism m within S is then

$$\varphi_e(m) := \varphi_e(m, z_c^*(m)) = \max_{Z \subseteq S} \varphi_e(m, z'_c(m, Z)). \tag{46}$$

The integrated cause information $\varphi_c(m)$ and the maximally irreducible cause $z_c^*(m)$ are defined in the same way (see S1 Fig). Based again on the principle of minimal existence, the irreducibility of the distinction specified by a mechanism is given by the minimum between its integrated cause and effect information

$$\varphi_d(m) = \min(\varphi_c(m), \varphi_e(m)). \tag{47}$$

Determining the set of causal distinctions that are congruent with the system cause–effect state. As required by composition, unfolding the full cause–effect structure of the system S in state s requires assessing the irreducible cause–effect power of every subset of units within S (Fig 2). Any $m \subseteq s$ with $\varphi_d > 0$ specifies a candidate distinction $d(m) = (m, z^*, \varphi_d)$ (27) within the system S in state s . However, in order to contribute to the cause–effect structure of a system, distinctions must also comply with intrinsicity and information at the system level. Thus, the fact that the system must select a specific cause–effect state implies that the cause–effect state they specify over subsets of the system ($z^* = \{z_c^*, z_e^*\}$) must be congruent with the cause–effect state specified over itself by the system as a whole s' .

We thus define the set of all causal distinctions within S in state s as

$$D(\mathcal{T}_e, \mathcal{T}_c, s) = \{d(m) : m \subseteq s, \varphi_d(m) > 0, z_c^*(m) \subseteq s'_c, z_e^*(m) \subseteq s'_e\}. \tag{48}$$

Altogether, distinctions can be thought of as irreducible “handles” through which the system can take and make a difference to itself by linking an intrinsic cause to an intrinsic effect over subsets of itself. As components within the system, causal distinctions have no inherent structure themselves. Whatever structure there may be between the units that make up a distinction is not a property of the distinction but due to the structure of the system, and thus captured already by its compositional set of distinctions. Similarly, from an extrinsic perspective,

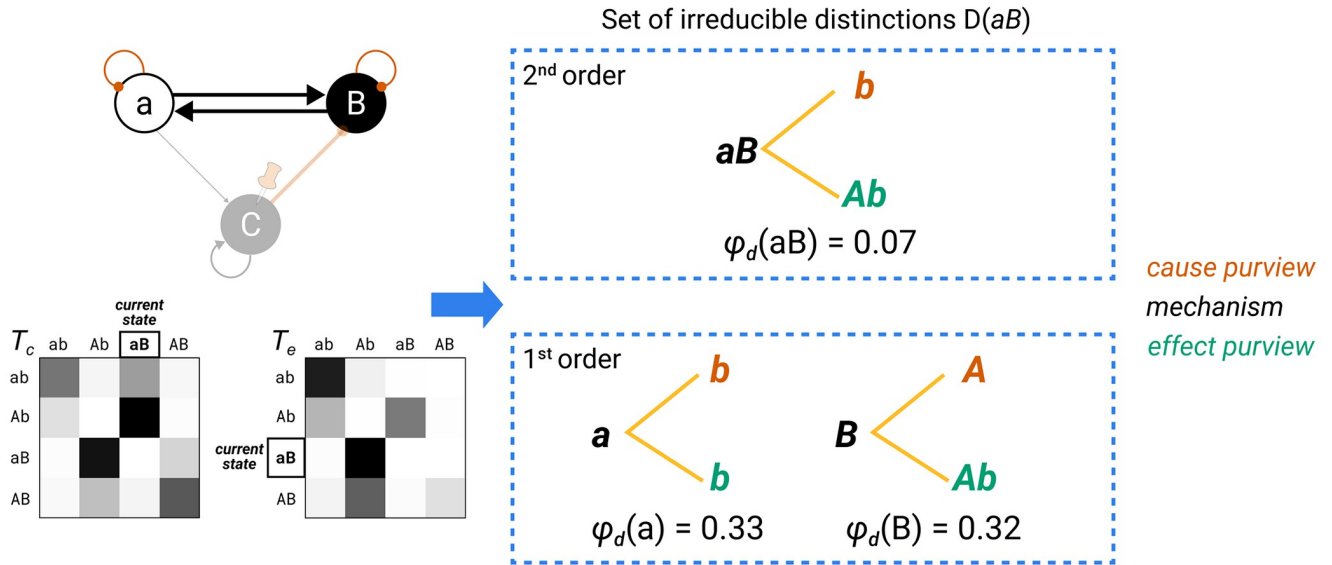


Fig 2. Composition and causal distinctions. Identifying the irreducible causal distinctions specified by a substrate in a state requires evaluating the specific causes and effects of every system subset. The candidate substrate is constituted of two interacting units $S = aB$ (see Fig 1) with TPMs T_c and T_e as shown. In addition to the two first-order mechanisms a and B , the second-order mechanism aB specifies its own irreducible cause and effect, as indicated by $\varphi_d > 0$.

<https://doi.org/10.1371/journal.pcbi.1011465.g002>

one may uncover additional causes and effects, both within the system and across its borders, at either macro or micro grains. However, from the intrinsic perspective of the system causes and effects that are excluded from its cause–effect structure do not exist [17, 29].

For example, as shown in Fig 3(A), a system may have a mechanism through which it specifies, in a maximally irreducible manner, the effect state of a triplet of units (e.g., $z_c^* = abc$, a third-order purview; again lowercase letters for units indicate state “–1,” uppercase letters state “+1”). However, if the system lacks a mechanism through which it can specify the effect state of single units, each taken individually (say, unit a , a first-order effect purview), then, from its intrinsic perspective, that unit does not exist as a single unit. By the same token, if the system can specify individually the state of unit a , b , and c , but lacks a way to specify irreducibly the state of abc together, then, from its intrinsic perspective, the triplet abc does not exist as a triplet (see Fig 3(B)). Finally, even if the system can distinguish the single units a , b , and c , as well as the triplet abc , if it lacks handles to distinguish pairs of units such as ab and bc , it cannot order units in a sequence.

Composition and causal relations

Causal relations capture how the causes and/or effects of a set of distinctions within a complex overlap with each other. Just as a distinction specifies which units/states constitute a cause purview and the linked effect purview, a relation specifies which units/states correspond to which units/states among the purviews of a set of distinctions. Relations thus reflect how the cause–effect power of its distinctions is “bound together” within a complex. The irreducibility due to this binding of cause–effect power is measured by the relations’ irreducibility ($\varphi_r > 0$). Relations between distinctions were first described in [11] (for differences with the initial presentation see S2 Text).

A set of distinctions $d \subseteq D(s)$ is related if the cause–effect state of each distinction $d \in d$ overlaps congruently over a set of shared units, which may be part of the cause, the effect, or

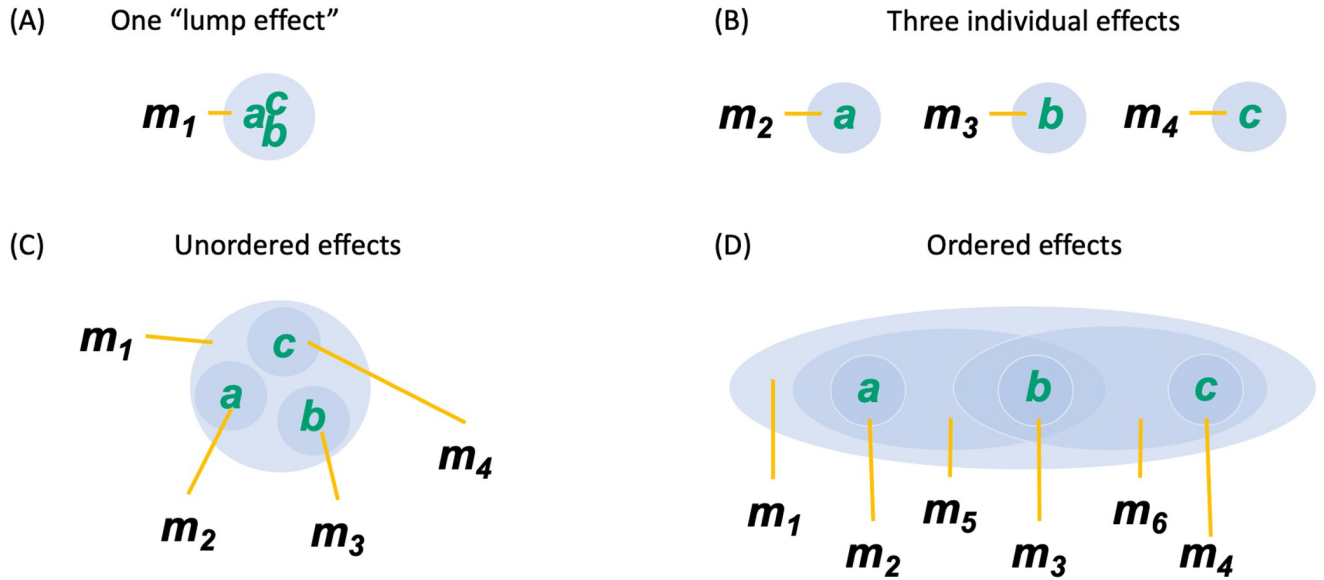


Fig 3. Composition of intrinsic effects. From the intrinsic perspective of the system, a specific cause or effect is only available to the system if it is selected by a causal distinction $d \in D(s)$. In (A), only the top-order effect is specified. From the intrinsic perspective, the system cannot distinguish the individual units. In (B), only first-order effects are specified. The system has no “handle” to select all three units together. (C) If both first- and third-order effects are specified, but no second-order effects, the system can distinguish individual units and select them together, but has no way of ordering them sequentially. (D) The system can distinguish individual units, select them altogether, as well as order them sequentially, in the sense that it has a handle for ab and bc , but not ac . The ordering becomes apparent once the relations among the distinctions are considered (see below, Fig 5).

<https://doi.org/10.1371/journal.pcbi.1011465.g003>

both the cause and the effect of each distinction. Below we will denote the cause of a distinction d as $z_c^*(d)$ and its effect as $z_e^*(d)$. For a given set of distinctions $\mathbf{d} \subseteq D(s)$, there are potentially many “relating” sets of causes and/or effects \mathbf{z} such that

$$\mathbf{z} : \mathbf{z} \cap \{z_c^*(d), z_e^*(d)\} \neq \emptyset \quad \forall d \in \mathbf{d}, \quad \bigcap_{z \in \mathbf{z}} z \neq \emptyset, \quad |\mathbf{z}| > 1 \tag{49}$$

with maximal overlap

$$o^*(\mathbf{z}) = \bigcap_{z \in \mathbf{z}} z \neq \emptyset. \tag{50}$$

Since $z_c^*(m) \subseteq s'_c$ and $z_e^*(m) \subseteq s'_e$ are sets of tuples containing both the units and their states, the intersection operation considers both the units and the state of the units.

All possible sets \mathbf{z} specify unique aspects about a relation $r(\mathbf{d})$ and constitute the various “faces” of the relation (Fig 4). The maximal overlap $o^*(\mathbf{z})$ (50) is also called the “face purview.” The set of faces associated with a relation thus specifies which type of relation it is (e.g., a single-faceted relation that only relates the causes of the set of distinctions, or a multi-faceted relation, which requires some of the distinctions to overlap on both the cause and effect side). Note that (49) includes the case $\mathbf{z} = \{z_c^*(d), z_e^*(d)\}$, which indicates a “self-relation” over the cause and effect of a single distinction $d \in D(s)$.

A relation $r(\mathbf{d})$ thus consists of a set of distinctions $\mathbf{d} \in D(s)$, with an associated set of faces $\mathbf{f}(\mathbf{d}) = \{f(\mathbf{z})\}_a$ and irreducibility $\varphi_r > 0$,

$$r(\mathbf{d}) = (\mathbf{d}, \mathbf{f}(\mathbf{d}), \varphi_r). \tag{51}$$

A relation that binds together $h = |\mathbf{d}|$ distinctions is a h -degree relation. A relation face $f(\mathbf{z}) \in \mathbf{f}(\mathbf{d})$

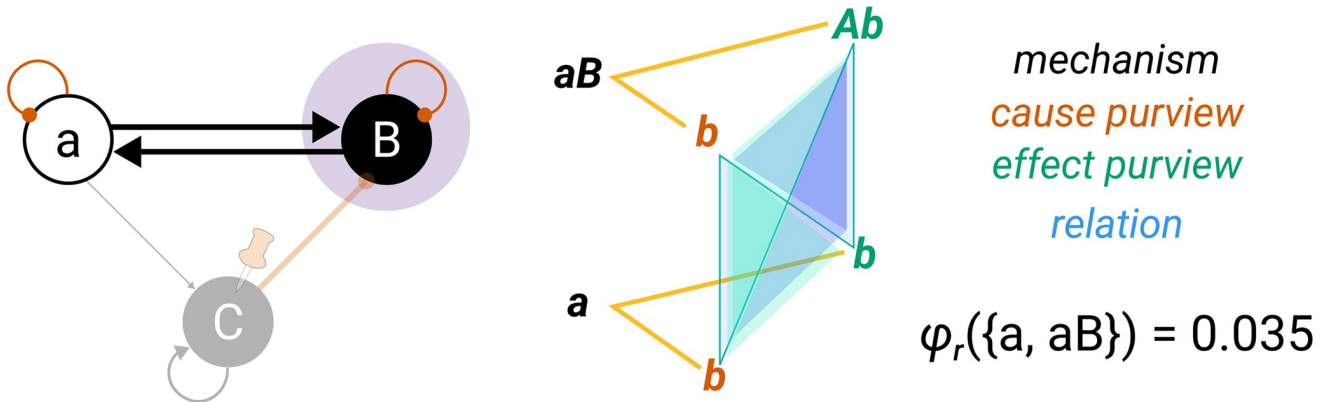


Fig 4. Composition and causal relations. Relations between distinctions specify joint causes and/or effects. The two distinctions $d(a)$ and $d(aB)$ each specify their own cause and effect. In this example, their cause and effect purviews overlap over the unit b and are congruent, which means that they all specify b to be in state “-1.” The relation $r(\{a, aB\})$ thus binds the two distinctions together over the same unit. Relation faces are indicated by the blue lines and surfaces between the distinctions’ causes and/or effects (different shades are used to individuate the faces). Because all four purviews overlap over the same unit, all nine possible faces exist. Note that the fact that the two distinctions overlap irreducibly can only be captured by a relation and not by a high-order distinction.

<https://doi.org/10.1371/journal.pcbi.1011465.g004>

consists of a set of causes and effects \mathbf{z} (as in 49), with associated face purview $o^*(\mathbf{z})$ (50)

$$f(\mathbf{z}) = (\mathbf{z}, o^*(\mathbf{z})). \tag{52}$$

A relation face over $k = |\mathbf{z}|$ purviews is a k -degree face. The set of faces includes all the ways in which the set of distinctions \mathbf{d} counts as related according to (49). Because \mathbf{z} may include either the cause, or the effect, or both the cause and effect of a distinction $d \in \mathbf{d}$, a relation $r(\mathbf{d})$ with $|\mathbf{d}| > 1$ may comprise up to $3^{|\mathbf{d}|}$ faces. If a set of distinctions $\mathbf{d} \in D(s)$ does not overlap congruently, it is not related (in that case $o^*(\mathbf{z}) = \emptyset$ for all possible $f(\mathbf{z}) \in f(\mathbf{d})$) (Fig 5).

Causal relations inherit existence from the cause–effect power of the distinctions that compose them. They inherit intrinsicality because the causes and effects that compose their faces are specified within the substrate. Moreover, relations are specific because the joint purviews of their faces must be congruent for all causes and effects $z^* \in \mathbf{z}$. Note that relation purviews are necessarily congruent with the overall cause and effect state specified by the system as a whole, because the causes and effects of the distinctions composing a relation must themselves be congruent.

The irreducibility of a causal relation is measured by “unbinding” distinctions from their joint purviews, taking into account all faces of the relation. Distinctions $d \in D(s)$ are already established as maximally irreducible components, characterized by their value of integrated information φ_d . To assess the irreducibility of a relation, we thus assume that the integrated information φ_d of a distinction is distributed uniformly across unique cause and effect purview units, such that

$$\frac{\varphi_d}{|z_c^*(d) \cup z_e^*(d)|} \tag{53}$$

is the average irreducible information φ_d per unique purview unit for an individual distinction $d \in \mathbf{d}$ with cause–effect state $z^*(d) = \{z_c^*(d), z_e^*(d)\}$. Since the union operator takes the states of the units into account, incongruent units are counted separately, while congruent units on the cause and effect side count as one.

Since distinctions are related by specifying common units into common states, the effect of “unbinding” a distinction must be proportional to the number of units jointly specified in the

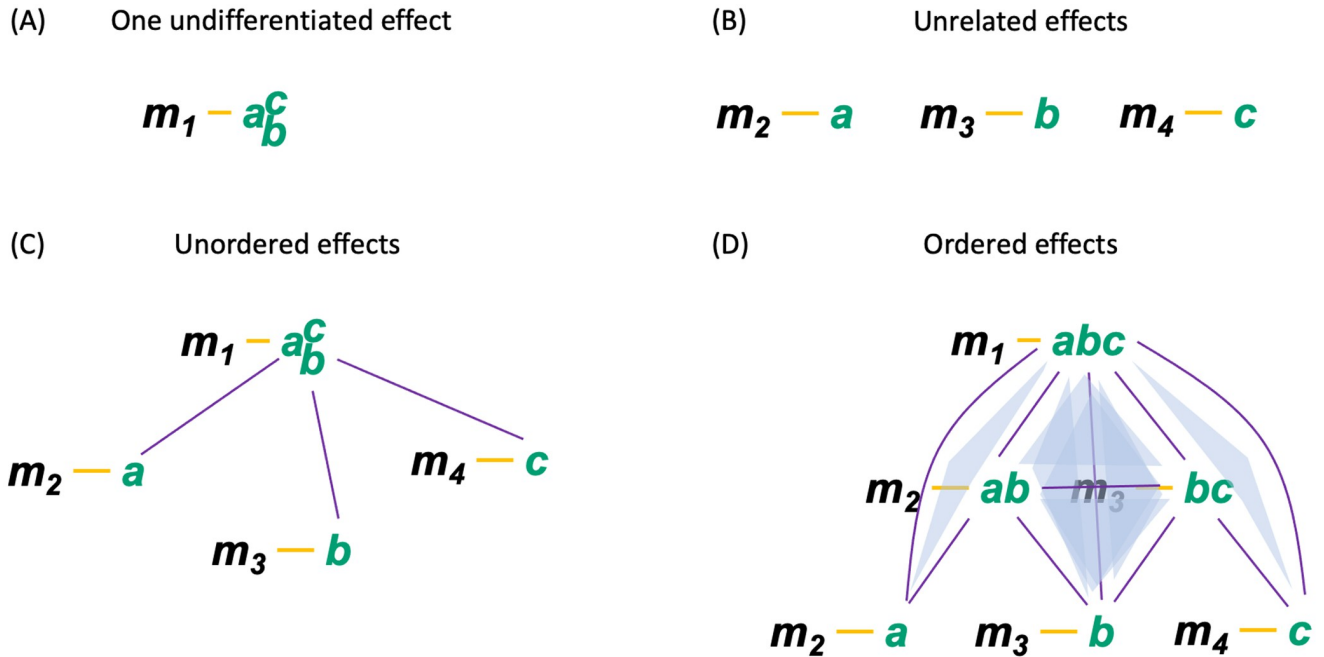


Fig 5. Structuring of intrinsic effects by relations. (A) A single undifferentiated effect has no relations. (B) Likewise, there are no relations among multiple non-overlapping effects. (C) The set of three first-order effects and one third-order effect supports three relations, which bind the effects together. (D) The set of first, second, and third-order effects supports a large number of relations (ten 2-relations (between two effects), six 3-relations, and one 4-relation), which bind the effects in a structure that is ordered sequentially.

<https://doi.org/10.1371/journal.pcbi.1011465.g005>

relation, *i.e.* the number of distinct units over the joint purviews of all faces in the relation:

$$\left| \bigcup_{f \in f(d)} o_f^* \right|. \tag{54}$$

This union of the face purviews o_f^* is also called the “relation purview” or the “joint purview” of the relation. While any partition of one or more distinctions from the relation will “unbind” the set of distinctions d , by the principle of minimal existence, a relation can only be as irreducible as the minimal amount of integrated information specified by any one distinction in the relation. Therefore, the relation integrated information $\varphi_r(d)$ is defined as

$$\varphi_r(d) = \min_{d \in d} \left| \bigcup_{f \in f(d)} o_f^* \right| \frac{\varphi_d}{|z_c^*(d) \cup z_e^*(d)|}. \tag{55}$$

In words, for each distinction, we take the average integrated information per distinct purview element (53), multiply it by the number of units across all faces of the relation (54), and then find the distinction that contributes the least integrated information per overlap unit as the minimum partition of the relation (with corresponding integrated information φ_r). Defining φ_r in this way guarantees that the integrated information of a relation cannot exceed the integrated information of its weakest distinction. For a given set of distinctions, the maximum value of φ_r occurs for a relation in which the cause and effect of each distinction is fully overlapped by all other distinctions in the relation (in that case, $\varphi_r = \min_{d \in d} \varphi_d$). Note also that a relation satisfies exclusion (distinctions overlap on *this whole* set of units) in that its integrated information is naturally maximized (per the principle of maximal existence) over the maximal

congruent overlap o_j^* for each relation face (50) (taking subsets of these overlaps could only reduce the integrated information of the relation).

In summary, just as distinctions link a cause with an effect, relations bind various combinations of causes and effects that are congruent over the same units (Fig 4). And just as a distinction captures the irreducibility of an individual cause–effect linked by a mechanism, a relation captures the irreducibility of a set of distinctions bound by the joint purviews of their causes and/or effects.

For a set of distinctions D , we define the set of all relations among them as

$$R(D) = \{r(\mathbf{d}) : \varphi_r(\mathbf{d}) > 0\}, \forall \mathbf{d} \subseteq D. \tag{56}$$

In practice, the total number of relations and their $\sum_{R(D)} \varphi_r$ can be determined analytically for a given set of distinctions D , which greatly reduces the necessary computations (see S3 Text). Together, a set of distinctions D and its associated set of relations $R(D)$ compose a cause–effect structure.

Cause–effect structures and Φ -structures

A cause–effect structure is defined as the union of the distinctions specified by a substrate and the relations binding them together:

$$C(D) = D \cup R(D). \tag{57}$$

The cause–effect structure specified by a maximal substrate—a complex—is also called a Φ -structure:

$$C(\mathcal{T}_e, \mathcal{T}_c, s^*) = \left\{ \{d(m) = \{m, z^*, \varphi_d\} \in \mathcal{T}_e, \mathcal{T}_c, s^*\} \cup \{r(\mathbf{d}) = \{\mathbf{d}, \mathbf{f}(\mathbf{d}), \varphi_r\} \in R(D(\mathcal{T}_e, \mathcal{T}_c, s^*))\} \right\}. \tag{58}$$

The sum of the values of integrated information of a substrate’s distinctions and relations, called Φ (“big Phi,” “structure Phi”) corresponds to the *structure integrated information* of the Φ -structure,

$$\Phi(\mathcal{T}_e, \mathcal{T}_c, s^*) = \sum_{C(\mathcal{T}_e, \mathcal{T}_c, s^*)} \varphi. \tag{59}$$

Note that Φ is not computed based on a partition (as system phi), but rather a sum of the integrated information within the structure (where each term of the sum was computed by partitioning). Within a Φ -structure, various types of meaningful sub-structures can be specified, which we term Φ -folds. A Φ -fold is composed of a subset of the distinctions and relations that compose the overall cause–effect structure. A special case is the *distinction Φ -fold*, denoted $C(\{d\})$, a sub-structure composed of a single distinction and the relations bound to it, which form its *context* [11] (see (13) in S1 Notes). A *compound Φ -fold* is a sub-structure composed of the distinction Φ -folds specified by a subset of units. A *compound Φ -fold* is a relevant part of a Φ -structure because it can be accessed or manipulated by changing the state, connections, or functioning of a part of the substrate. Finally, a *content Φ -fold*, or simply *content*, is composed of a subset of distinctions that are highly interrelated (regardless of the mechanisms and units that specify them).

In conclusion, a maximal substrate or complex is a set of units $S^* = s^*$ that satisfies all of IIT’s postulates: its cause–effect power is intrinsic, specific, irreducible, definite, and structured. By IIT, a complex S^* does not exist as such, but exists “unfolded” into its Φ -structure, with all the causal distinctions and relations that compose it. In other words, a substrate is what can be observed and manipulated “operationally” from the extrinsic perspective. From

the intrinsic perspective, what truly exists is a complex with all its causal powers unfolded—an *intrinsic entity* that exists for itself, absolutely, rather than relative to an external observer.

According to the explanatory identity of IIT, an experience is identical to the Φ -structure of an intrinsic entity: every property of the experience should be accounted for by a corresponding property of the Φ -structure, with no additional ingredients. If a system S in state s is a complex, then its Φ -structure corresponds to the quality of the experience of S in state s , while its Φ value corresponds to its quantity—in other words, to the nature and amount of intrinsic content.

Results and discussion

In this section, we apply the mathematical framework of IIT 4.0 to several example systems. The goal is to illustrate three critical implications of IIT's postulates:

1. **Consciousness and connectivity:** how the way units interact determines whether a substrate can support a Φ -structure of high Φ .
2. **Consciousness and activity:** how changes in the state of a substrate's units change Φ -structures.
3. **Consciousness and functional equivalence:** how substrates that are functionally equivalent may not be equivalent in terms of their Φ -structures, and thus in terms of consciousness.

The following examples will feature very simple networks constituted of binary units $U_i \in U$ with $\Omega_{U_i} = \{-1, 1\}$ for all U_i and a logistic (sigmoidal) activation function

$$p(U_{i,t} = 1 \mid u_{t-1}) = \frac{1}{1 + \exp(-k \sum_{j=1}^n w_{j,i} u_{j,t-1})}, \quad (60)$$

where $k > 0$ and

$$\sum_{j=1}^n w_{j,i} = 1 \quad \forall i. \quad (61)$$

In Eq (60), the parameter k defines the slope of the logistic function and allows one to adjust the amount of noise or determinism in the activation function (higher values signify a steeper slope and thus more determinism). The units U_i can thus be viewed as noisy linear threshold units with weighted connections among them, where k determines the connection strength.

As in Figs 1 and 2, units denoted by uppercase letters are in state '1' (ON, depicted in black), units denoted by lowercase letters are in state '-1' (OFF, depicted in white). Cause-effect structures are illustrated as geometrical shapes projected into 3D space (Fig 6). Distinctions are depicted as mechanisms (black labels) tying a cause (red labels) and an effect (green labels) through a link (orange edges, thickness indicating φ_d). Relation faces of second- and third-degree relations are depicted as edges or triangular surfaces between the causes and effects of the related distinctions. While edges always bind pairs of distinctions (a second-degree relation), triangular surfaces may bind the causes and effects of two or three distinctions (second- or third-degree relation). Relations of higher degrees are not depicted.

All examples were computed using the "iit-4.0" feature branch of PyPhi [37]. This branch will be available in the next official release of the software. An example notebook [available here](#) recreates the analysis of Fig 1 (identifying complexes), Fig 2 (computing distinctions), and Fig 4 (computing relations).

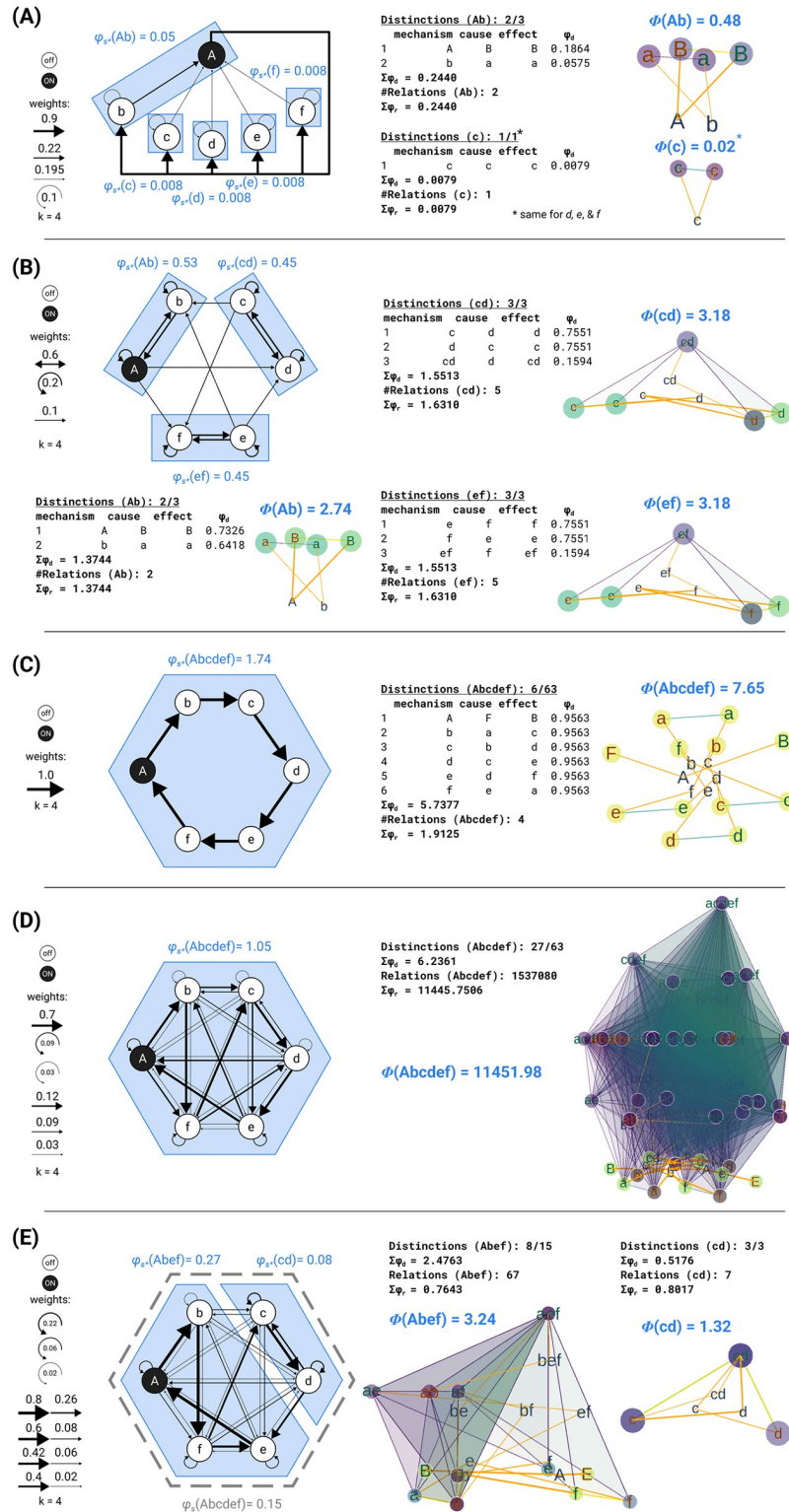


Fig 6. Causal powers analysis of various network architectures. Each panel shows the network's causal model and weights on the left. Blue regions indicate complexes with their respective ϕ_s values. In all networks, $k = 4$ and the state is $Abcdef$. The Φ -structure(s) specified by the network's complexes are illustrated to the right (with only second- and third-degree relation faces depicted) with a list of their distinctions for smaller systems and their $\Sigma\phi$ values for those systems with many distinctions and relations. All integrated information values are in bits. (A) A degenerate network

in which unit *A* forms a bottleneck with redundant inputs from and outputs to the remaining units. The first-maximal complex is *Ab*, which excludes all other subsets with $\varphi_s > 0$ except for the individual units *c*, *d*, *e*, and *f*. (B) The modular network condenses into three complexes along its fault lines (which exclude all subsets and supersets), each with a maximal φ_s value, but low Φ , as the modules each specify only two or three distinctions and at most five relations. (C) A directed cycle of six units forms a six-unit complex with $\varphi_s = 1.74$ ibits, as no other subset is integrated. However, the Φ -structure of the directed cycle is composed of only first-order distinctions and few relations. (D) A specialized lattice also forms a complex (which excludes all subsets), but specifies 27 first- and high-order distinctions, with many relations ($>1.5 \times 10^6$) among them. Its Φ value is 11452 ibits. (E) A slightly modified version of the specialized lattice in which the first-maximal complex is *Abef*. The full system is not maximally irreducible and is excluded as a complex, despite its positive φ_s value (indicated in gray).

<https://doi.org/10.1371/journal.pcbi.1011465.g006>

Consciousness and connectivity

The first set of examples highlights how the organization of connections among units impacts the ability of a substrate to support a cause–effect structure with high structure integrated information (high Φ). Fig 6 shows five systems, all in the same state $s = Abcdef$ with the same number of units, but with different connectivity among the units.

Degenerate systems, indeterminism, and specificity. Fig 6A shows a network with medium indeterminism ($k = 4$) and high degeneracy, due to the fact that unit *A* forms a “bottleneck” with inputs and outputs to and from the remaining units. The network condenses into one complex of two units *Ab* and four complexes corresponding to the individual units *c*, *d*, *e*, and *f* (also called “monads”).

The causes and effects of the causal distinctions for the two types of complexes are shown in the middle, and the corresponding cause–effect structures are illustrated on the right. In this case, degeneracy (coupled with indeterminism) undermines the ability of the maximal substrate to grow in size, which in turn limits the richness of the Φ -structure that can be supported. Because of the bottleneck architecture, the current state of candidate system *Abcdef* has many possible causes and effects, leading to an exponential decrease in selectivity (the conditional probabilities of cause and effect states). This dilutes the value of intrinsic information (ii) for larger subsets of units, which in turn reduces their value of system integrated information φ_s . Consequently, the maximal substrates are small, and their Φ values are necessarily low.

This example suggests that to grow and achieve high values of Φ , substrates must be constituted of units that are specialized (low degeneracy) and interact very effectively (low indeterminism).

Notably, the organization of the cerebral cortex, widely considered as the likely substrate of human consciousness, is characterized by extraordinary specialization of neural units at all levels [38–40]. Moreover, if the background conditions are well controlled, neurons are thought to interact in a highly reliable, nearly deterministic manner [41–43].

Modular systems, fault lines, and irreducibility. Fig 6B shows a network comprising three weakly interconnected modules, each having two strongly connected units ($k = 4$). In this case, the weak inter-module connections are clear fault lines. Properly normalized, partitions along these fault lines separating modules yield values of φ_s that are much smaller than those yielded by partitions that cut across modules. As a consequence, the 6-unit system condenses into three complexes (*Ab*, *cd*, and *ef*), as determined by their maximal φ_s values. Again, because the modules are small, their Φ values are low. Intriguingly, a brain region such as the cerebellum, whose anatomical organization is highly modular, does not contribute to consciousness [44, 45], even though it contains several times more neurons than the cerebral cortex (and is indirectly connected to it).

Note that fault lines can be due not just to neuroanatomy but also to neurophysiological factors. For example, during early slow-wave sleep, the dense interconnections among neuronal groups in cerebral cortical areas may break down, becoming causally ineffective due to the

bistability of neuronal excitability. This bistability, brought about by neuromodulatory changes [46], is associated with the loss of consciousness [47].

Directed cycles, structural sparseness, and composition. Fig 6C shows a directed cycle in which six units are unidirectionally connected with weight $w = 1.0$ and $k = 4$. Each unit copies the state of the unit before it, and its state is copied by the unit after it, with some indeterminism. The copy cycle constitutes a 6-unit complex with a maximal $\varphi_s = 1.74$ ibits. However, despite the “large” substrate, the Φ -structure it specifies has low structure integrated information ($\Phi = 7.65$). This is because the system’s Φ -structure is composed exclusively of first-order distinctions, and consequently of a small number of relations.

Highly deterministic directed cycles can easily be extended to constitute large complexes, being more irreducible than any of their subsets. However, the lack of cross-connections (“chords” in graph-theoretic terms) greatly limits the number of components of the Φ -structures specified by the complexes, and thus their structure integrated information (Φ). (Note also that increasing the number of units that constitute the directed cycle would not change the amount of φ_s specified by the network as a whole.)

The brain is rich in partially segregated, directed cycles, such as those originating in cortical areas, sequentially reaching stations in the basal ganglia and thalamus, and cycling back to cortex [48, 49]. These cycles are critical for carrying out many cognitive and other functions, but they do not appear to contribute directly to experience [4].

Specialized lattices and Φ -structures with high structure integrated information. Fig 6D shows a network consisting of six heterogeneously connected units—a “specialized” lattice, again with $k = 4$. While many subsystems within the specialized network have positive values of system integrated information φ_s , the full 6-unit system is the maximal substrate (excluding all its subsets from being maximal substrates). Out of 63 possible distinctions, the Φ -structure comprises 27 distinctions with causes and effects congruent with the system’s maximal cause-effect state. Consequently, the full 6-unit system also specifies a much larger number of causal relations compared to the copy cycle system.

Preliminary work indicates that lattices of specialized units, implementing different input-output functions, but partially overlapping in their inputs (receptive field) and outputs (projective fields), are particularly well suited to constituting large substrates that unfold into extraordinarily rich Φ -structures. The number of distinctions specified by an optimally connected, specialized system is bounded above by $2^n - 1$, and that of the relations among as many distinctions is bounded by $2^{(2^n - 1)} - 1$. The structure integrated information of such structures is correspondingly large [50].

In the brain, a large part of the cerebral cortex, especially its posterior regions, is organized as a dense, divergent-convergent hierarchical 3D lattice of specialized units, which makes it a plausible candidate for the substrate of human consciousness [4, 11, 51, 52]. Note that directed cycles originating and ending in such lattices typically remain excluded from the first-maximal complex because minimal partitions across such cycles yield a much lower value of φ_s compared to minimal partitions across large lattices.

Near-maximal substrates, extrinsic entities, and exclusion. Finally, Fig 6E shows a network of six units, four of which (*Abef*) constitute a specialized lattice that corresponds to the first complex. Though integrated, the full set of 6 units happens to be slightly less irreducible ($\varphi_s = 0.15$) than one of its 4-unit subsets ($\varphi_s = 0.27$). From the extrinsic perspective, the 6-unit system undoubtedly behaves as a highly integrated whole (nearly as much as its 4-unit subset), one that could produce complex input-output functions due to its rich internal structure. From the intrinsic perspective of the system, however, only the 4-unit subset satisfies all the postulates of existence, including maximal irreducibility (accounting for the definite nature of

experience). In this example, the remaining units form a second complex with low φ_s and serve as background conditions for the first complex.

A similar situation may occur in the brain. The brain as a whole is undoubtedly integrated (not to mention that it is integrated with the body as a whole), and neural “traffic” is heavy throughout. However, its anatomical organization may be such that a subset of brain regions, arranged in a dense 3D lattice primarily located in posterior cortex, may achieve a much higher value of integrated information than any other subset. Those regions would then constitute the first complex (the “main complex,” [4]), and the remaining regions might condense into a large number of much smaller complexes.

Taken together, the examples in Fig 6 demonstrate that the connectivity among the units of a system has a strong impact on what set of units can constitute a complex and thereby on the structure integrated information it can specify. The examples also demonstrate the role played by the various requirements that must be satisfied by a substrate of consciousness: existence (causal power), intrinsicity, specificity, maximal irreducibility (integration and exclusion), and composition (structure).

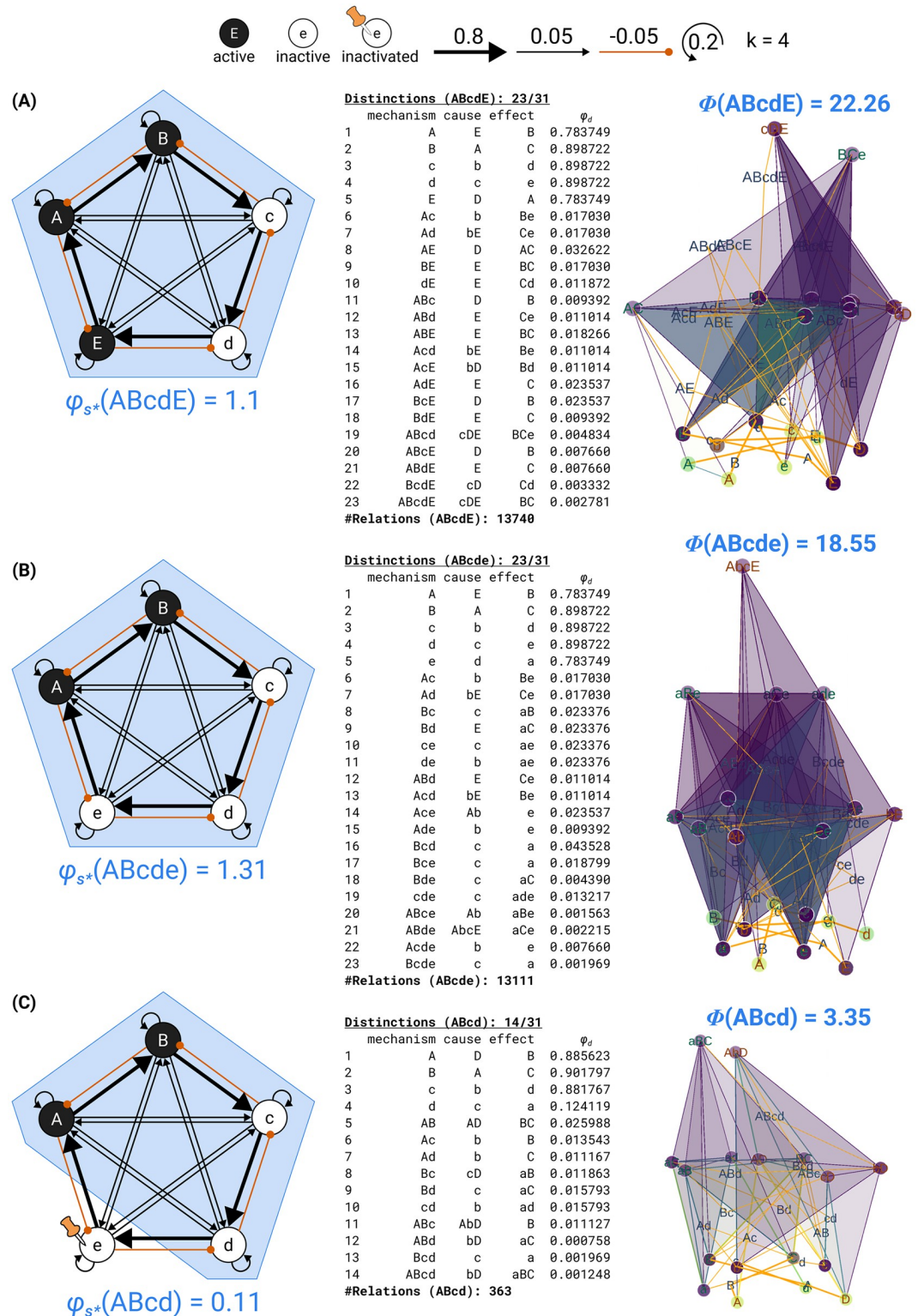
Consciousness and activity: Active, inactive, and inactivated units

A substrate exerts cause–effect power in its current state. For the same substrate, changing the state of even one unit may have major consequences on the distinctions and relations that compose its Φ -structure: many may be lost, or gained, and many may change their value of irreducibility (φ_d and φ_r).

Fig 7 shows a network of five binary units that interact through excitatory and inhibitory connections (weights indicated in the figure). The system is initially in state $s = ABCdE$ (Fig 7A) and is a maximal substrate with $\varphi_s = 1.1$ ibits and a Φ -structure composed of 23 distinctions and their 13740 relations.

If we change the state of unit E from ON to OFF (in neural terms, the unit becomes inactive), the distinctions that the unit contributes to when ON, as well as the associated relations, may change (Fig 7B). In the case illustrated by the Figure, what changes are the purviews and irreducibility of several distinctions and associated relations, the number of distinctions stays the same, φ_s changes only slightly, but the number of relations is lower, leading to a lower Φ value. In other words, what a single unit contributes to intrinsic existence is not some small “bit” of information. Instead, a unit contributes an entire sub-structure, composed of a very large number of distinctions and relations. The set of distinctions to which a subset of units contributes as a mechanism, either alone or in combination with other units, together with their associated relations, forms a compound Φ -fold. With respect to the neural substrate of consciousness in the brain, this means that even a change in the state of a single unit is typically associated with a change in an entire Φ -fold within the overall Φ -structure, with a corresponding change in the structure of the experience. (Note, however, that in larger systems such changes will typically be less extreme, see also [11].).

In Fig 7C, we see what happens if unit E , instead of just turning inactive (OFF) is *inactivated* (abolishing its cause–effect power because it no longer has any counterfactual states and thus cannot be intervened upon). In this case, all the distinctions and relations to which that unit contributes as a mechanism would cease to exist (its compound Φ -fold collapses). Moreover, all the distinctions and relations to whose purviews that unit contributes—its purview Φ -fold—would also collapse or change. In fact, the complex shrinks because it cannot include that unit. With respect to the neural substrate of consciousness, this means that while an inactive unit contributes to a different experience, an inactivated unit ceases to contribute to experience altogether. The fundamental difference between inactive and inactivated units leads to the



The same system in state $ABcde$, where unit E is inactive (“OFF”) also forms a complex with the same number of distinctions, but a somewhat lower Φ value due to a lower number of relations between distinctions. In addition, the system’s Φ -structure differs from that in (A) , as the system now specifies a different set of compositional causes and effects. (C) If instead of being inactive, unit E is inactivated (fixed into the “OFF” state), the inactivated unit cannot contribute to the complex or Φ -structure anymore. The complex is now constituted of four units ($ABcd$), with only 14 distinctions and markedly reduced structure integrated information ($\Phi = 3.35$).

<https://doi.org/10.1371/journal.pcbi.1011465.g007>

following corollary of IIT: unlike a fully inactivated substrate which, as would be suspected, cannot support any experience, an inactive substrate can. If a maximal substrate in an inactive state is in working order and specifies a large Φ -structure, it will support a highly structured experience, such as the experience of empty space [11] or the feeling of “pure presence” (see (14) in S1 Notes).

Consciousness and functional equivalence: Being is not doing

By the intrinsicity postulate, the Φ -structure of a complex depends on the causal interactions between system subsets, not on the system’s interaction with its environment (except for the role of the environment in triggering specific system states). In general, different physical systems with different internal causal structure may perform the same input–output functions.

Fig 8 shows three simple deterministic systems with binary units (here the “OFF” state is 0, and “ON” is 1) that perform the same input–output function, treating the internal dynamics of the system as a black box. The function could be thought of, for example, as an electronic toll-booth “counting 8 valid coins” (8 times input $I = 1$) before opening the gate [53]. Each system receives one binary input (I) and has one binary output (O). The output unit switches “ON” on a count of eight positive inputs $I = 1$ (when the global state with label ‘0’ is reached in the cycle), upon which the system resets (Fig 8A).

In addition to being functionally equivalent in their outward behavior, the three systems share the same internal global dynamics, as their internal states update according to the same global state-transition diagram (Fig 8B). Given an input $I = 1$, the system updates its state, cycling through all its 8 global states (labeled 0–7) over 8 updates. For an input of $I = 0$, the system remains in its present state. Moreover, all three systems are constituted of three binary units whose joint states map one-to-one onto the systems’ global state labels (0–7). However, the mapping is different for different systems (Fig 8C, left). This is because the internal binary update sequence depends on the interactions among the internal units [29, 53], which differ in the three cases, as can easily be determined through manipulations and observations.

For consistency in the causal powers analysis, in all three cases, the global state “0” that activates the output unit if $I = 1$ is selected such that it corresponds to the binary state “all OFF” (000), which is followed by $1 := 100$ and $2 := 010$. Also, the Φ -structure of each system is unfolded in state $1 := 100$ in all three cases.

Despite their functional equivalence and equivalent global dynamics, the systems differ in how they condense into complexes and in the cause–effect structures they specify.

As shown in Fig 8C, the first system forms a 3-unit complex with a relatively rich Φ -structure ($\Phi = 21.01$ ibits). While the second system also forms a 3-unit complex with the same $\varphi_s = 2$ ibits, it specifies a completely different set of distinctions and has much lower structure integrated information ($\Phi = 3.64$ ibits).

Finally, the third system is reducible ($\varphi_s = 0$ ibits)—in this case, because there are only feed-forward connections from unit A to units B and C —and it condenses into three complexes with small Φ -structures.

These examples illustrate a simple scenario of functional equivalence of three systems characterized by a different architecture. The equivalence is with respect to a simple input–output

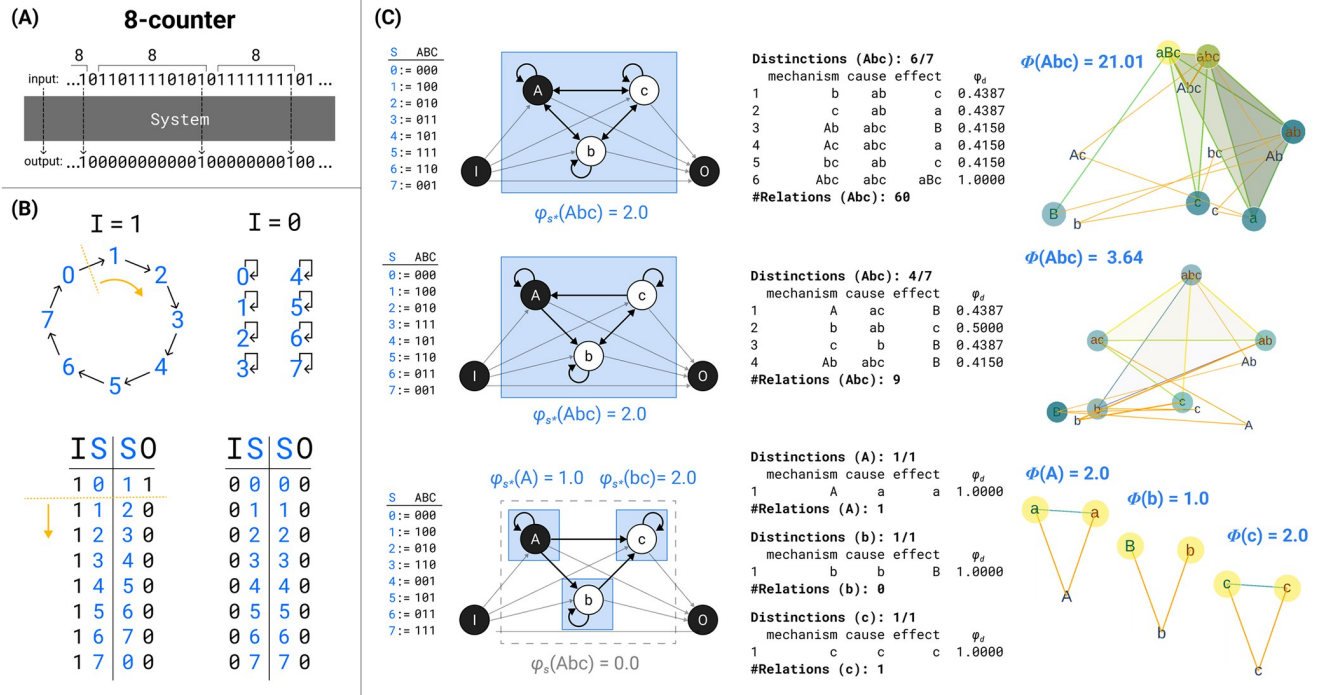


Fig 8. Functionally equivalent networks with different Φ -structures. (A) The input–output function realized by three different systems (shown in (C)): a count of eight instances of input $I = 1$ leads to output $O = 1$. (B) The global state–transition diagram is also the same for the three systems: if $I = 0$, the systems will remain in their current global state, labeled as 0–7; if $I = 1$, the systems will move one state forward, cycling through their global states, and activate the output if $S = 0$. (C) Three systems constituted of three binary units but differing in how the units are connected and interact. As a consequence, the one-to-one mapping between the 3-bit binary states and the global state labels differ. However, all three systems initially transition from 000 to 100 to 010. Analyzed in state 100, the first system (top) turns out to be a single complex that specifies a Φ -structure with six distinctions and many relations, yielding a high value of Φ . The second system (middle) is also a complex, with the same φ_s value, but it specifies a Φ -structure with fewer distinctions and relations, yielding a lower value of Φ . Finally, the third system (bottom) is reducible ($\varphi_s = 0$) and splits into three smaller complexes (entities) with minimal Φ -structures and low Φ .

<https://doi.org/10.1371/journal.pcbi.1011465.g008>

function, in this case coin counting, which they multiply realize. The systems are also equivalent in terms of their global system dynamics, in the sense that they go through a globally equivalent sequence of internal states. However, because of their different substrates, the three systems specify different cause–effect structures. Therefore, based on the postulates of IIT, they are not phenomenally equivalent. In other words, they are equivalent in what they *do* extrinsically, but not in what they *are* intrinsically.

This dissociation between phenomenal and functional equivalence has important implications. As we have seen, a purely feed-forward system necessarily has $\varphi_s = 0$. Therefore, it cannot support a cause–effect structure and cannot be conscious, whereas systems with a recurrent architecture can. On the other hand, the behavior (input–output function) of any (discrete) recurrent system can also be implemented by a system with a feed-forward architecture [54]. This implies that any behavior performed by a conscious system supported by a recurrent architecture can also be performed by an unconscious system, no matter how complex the behavior is. More generally, digital computers implementing programs capable of artificial general intelligence may in principle be able to emulate any function performed by conscious humans and yet, because of the way they are physically organized, they would do so without experiencing anything, or at least anything resembling, in quantity and quality, what each of us experiences [20] (see also (15) in S1 Notes).

The examples also show that the overall system dynamics, while often revealing relevant aspects of a system's architecture, typically do not and cannot exhaust the richness of its current cause–effect structure. For example, a system in a fixed point is dynamically “dead” (and “does” nothing), but it may be phenomenally quite “alive,” for example, experiencing “pure presence” (see (14) in [S1 Notes](#)). Of course, the system's causal powers can be fully unfolded, and revealed dynamically, by extensive manipulations and observations of subsets of system units because they are implicitly captured by the system's causal model and ultimately by its transition probability matrix [29].

Conclusions

IIT attempts to account for the presence and quality of consciousness in physical terms. It starts from the existence of experience, and proceeds by characterizing its essential properties—those that are immediate and irrefutably true of every conceivable experience (axioms). These are then formulated as essential properties of physical existence (postulates), the necessary and sufficient conditions that a substrate must satisfy to support an experience—to constitute a complex. Note that “substrate” is meant in purely operational terms—as a set of units that a conscious observer can observe and manipulate. Likewise, “physical” is understood in purely operational terms as cause–effect power—the power to take and make a difference.

The postulates can be assessed based purely on a substrate's transition probability matrix, as was illustrated by a few idealized causal models. Thus, a substrate of consciousness must be able to take and make a difference upon itself (existence and intrinsicity), it must be able to specify a cause and an effect state that are highly informative and selective (information), and it must do so in a way that is both irreducible (integration) and definite (exclusion). Finally, it must specify its cause and effect in a structured manner (composition), where the causal powers of its subsets over its subsets compose a cause–effect structure of distinctions and relations—a Φ -structure. Thus, a complex does not exist as such but only “unfolded” as a Φ -structure—an *intrinsic entity* that exists for itself, absolutely, rather than relative to an external observer.

As shown above, these requirements constrain what substrates can and cannot support consciousness. Substrates that lack in specificity, due to indeterminism and/or degeneracy, cannot grow to be large complexes. Substrates that are weakly integrated, due to architectural or functional fault lines in their interactions, are less integrated than some of their subsets. Because they are not maximally irreducible, they do not qualify as complexes. This is the case even though they may “hang together” well enough from an extrinsic perspective (having a respectable value of φ_s). Furthermore, even substrates that are maximally integrated may support Φ -structures that are extremely sparse, as in the case of directed cycles. Based on the postulates of IIT, a universal substrate ultimately “condenses” into a set of disjoint (non-overlapping) complexes, each constituted of a set of macro or micro units.

The physical account of consciousness provided by IIT should be understood as an explanatory identity: every property of an experience should ultimately be accounted for by a property of the cause–effect structure specified by a substrate that satisfies its postulates, with no additional ingredients. The identity is not between two different substances or realms—the phenomenal and the physical—but between intrinsic (subjective) existence and extrinsic (objective) existence. Intrinsic existence is immediate and irrefutable, while extrinsic existence is defined operationally as cause–effect power discovered through observation and manipulation. The primacy of intrinsic existence (of experience) in IIT contrasts with standard attempts at accounting for consciousness as something “generated by” or “emerging from” a substrate constituted of matter and energy and following physical laws.

The physical correspondent of an experience is not the substrate as such but the Φ -structure specified by the substrate in its current state. Therefore, minor changes in the substrate state can correspond to major changes in the specified Φ -structure. For example, if the state of a single unit changes, an entire Φ -fold within the Φ -structure will change, and if a single inactive unit is inactivated, its associated Φ -fold will collapse, even though the current state of the substrate appears the same (Fig 7).

Each experience corresponds to a Φ -structure, not a set of functions, processes, or computations. Said otherwise, consciousness is about being, not doing [1, 29, 55]. This means that systems with different architectures may be functionally equivalent—both in terms of global input–output functions and global intrinsic dynamics—but they will not be phenomenally equivalent. For example, a feed-forward system can be functionally equivalent to a recurrent system that constitutes a complex, but feed-forward systems cannot constitute complexes because they do not satisfy maximal irreducibility. Accordingly, artificial systems powered by super-intelligent computer programs, but implemented by feed-forward hardware or encompassing critical bottlenecks, would experience nothing (or nearly nothing) because they have the wrong kind of physical architecture, even though they may be behaviorally indistinguishable from human beings [20].

Even though the entire framework of IIT is based on just a few axioms and postulates, it is not possible in practice to exhaustively apply the postulates to unfold the cause–effect power of realistic systems [32, 56]. It is not feasible to perform all possible observations and manipulations to fully characterize a universal TPM, or to perform all calculations on the TPM that would be necessary to condense it exhaustively into complexes and unfold their cause–effect power in full. The number of possible systems, of system partitions, of candidate distinctions—each with their partitions and relations—is the result of multiple, nested combinatorial explosions. Moreover, these observations, manipulations, and calculations would need to be repeated at many different grains, with many rounds of maximizations. For these reasons, a full analysis of complexes and their cause–effect structure can only be performed on idealized systems of a few units [37].

On the other hand, we can simplify the computation considerably by using various assumptions and approximations, as with the “cut one” approximation described in [37]. Also, while the number of relations vastly exceeds the number of units and of distinctions (its upper bound for a system of n units is $2^{(2^n - 1)} - 1$), it can be determined analytically, and so can $\Sigma\varphi_r$ for a given set of distinctions S3 Text. Developing tight approximations, as well as bounded estimates of a system’s integrated information (φ_s and Φ), is one of the main areas of ongoing research related to IIT [50].

Despite the infeasibility of an exhaustive calculation of the relevant quantities and structures for a realistic system, IIT already provides considerable explanatory and predictive power in many real-world situations, making it eminently testable [4, 57, 58]. A fundamental prediction is that Φ should be high in conscious states, such as wakefulness and dreaming, and low in unconscious states, such as dreamless sleep and anesthesia. This prediction has already found substantial support in human studies that have applied measures of complexity inspired by IIT to successfully classify subjects as conscious vs. unconscious [4, 22, 23, 59]. IIT can also account mechanistically for the loss of consciousness in deep sleep and anesthesia [4, 47]. Furthermore, it can provide a principled account of why certain portions of the brain may constitute an ideal substrate of consciousness and others may not, why the borders of the main complex in the brain should be where they are, and why the units of the complex should have a particular grain (the one that yields a maximum of φ_s). A stringent prediction is that the location of the main complex, as determined by the overall maximum of φ_s within the brain, should correspond to its location as determined through clinical and experimental evidence.

Another prediction that follows from first principles is that constituents of the main complex can support conscious contents even if they are mostly inactive, but not if they are inactivated [4, 11]. Yet another prediction is that the complete inactivation of constituents of the main complex should lead to absolute agnosia (unawareness that anything is missing).

IIT further predicts that the quality of experience should be accounted for by the way the Φ -structure is composed, which in turn depends on the architecture of the substrate specifying it. This was demonstrated in a recent paper showing how the fundamental properties of spatial experiences—those that make space feel “extended”—can be accounted for by those of Φ -structures specified by 2D grids of units, such as those found in much of posterior cortex [11]. This prediction is in line with neurological evidence of their role in supporting the experience of space [11]. Ongoing work aims at accounting for the quality of experienced time and that of experienced objects (see (16) in [S1 Notes](#)). A related prediction is that changes in the strength of connections within the neural substrate of consciousness should be associated with changes in experience, even if neural activity does not change [60]. Also, similarities and dissimilarities in the structure of experience should be accounted for by similarities and dissimilarities among Φ -structures and Φ -folds specified by the neural substrate of consciousness.

While the listed predictions may appear largely qualitative in nature, many of them rest on specific features of the accompanying quantitative analysis. This is the case for predictions regarding the borders (and grain) of the main complex in the brain, which depend on the relative φ_s values of potential substrates of interest, and even more so for predictions regarding the quality and richness of certain experiences and the predicted features of their underlying substrates. IIT's postulates, and the mathematical framework proposed to evaluate them, rest on “inferences to a good explanation” ([Box 1](#)). While we have aimed for maximal consistency, specificity, and simplicity at every junction in formulating IIT's mathematical implementation, some of the algorithmic choices remain open to further evaluation. These include, for example, the proper treatment of background conditions and the resolution of ties given symmetries in the TPMs of specific systems (see [S1 Text](#)). More generally, further validation of IIT will depend on a systematic back-and-forth between phenomenology, theoretical inferences, and neuroscientific evidence [1].

In addition to empirical work aimed at validating the theory, much remains to be done at the theoretical level. According to IIT, the meaning of an experience is its feeling—whether those of spatial extendedness, of temporal flow, or of objects, to name but a few (“the meaning is the feeling”). This means that every meaning is identical to a sub-structure within a current Φ -structure—a content of experience—whether it is triggered by extrinsic inputs or it occurs spontaneously during a dream. Therefore, all meaning is ultimately intrinsic. Ongoing work aims at providing a self-consistent explanation of how intrinsic meanings can capture relevant features of causal processes in the environment (see (17) in [S1 Notes](#)). It will also be important to explain how intersubjectively validated knowledge can be obtained despite the intrinsic and partially idiosyncratic nature of meaning.

To the extent that the theory is validated through empirical evidence obtained from the human brain, IIT can then offer a plausible inferential basis for addressing several questions that depend on an explicit theory of consciousness. As indicated in the section on phenomenal and functional equivalence, and argued in ongoing work [20], one consequence of IIT is that typical computer architectures are not suitable for supporting consciousness, no matter whether their behavior may resemble ours. By the same token, it can be inferred from IIT that animal species that may look and behave quite differently from us may be highly conscious, as long as their brains have a compatible architecture. Other inferences concern our own experience and whether it plays a causal role, or is simply “along for the ride” while our brain performs its functions. As recently argued, IIT implies that we have true free will—that we have

true alternatives, make true decisions, and truly cause. Because only what truly exists (intrinsically, for itself) can truly cause, we, rather than our neurons, cause our willed actions and are responsible for their consequences [18].

Finally, an ontology that is grounded in experience as intrinsic existence—an intrinsic ontology—must not only provide an account of subjective existence in objective, operational terms, but also offer a path toward a unified view of nature—of all that exists and happens. One step in this direction is the application of the same postulates that define causal powers (existence) to the evaluation of actual causes and effects (“what caused what” [10]). Another is to unify classical accounts of information (as communication and storage of signals) with IIT’s notion of information as derived from the properties of experience—that is, information as causal, intrinsic, specific, maximally irreducible, and structured (meaningful) [8] (see also (18) in [S1 Notes](#)). Yet another is the study of the evolution of a substrate’s causal powers as conditional probabilities that update themselves [61].

Even so, there are many ways in which IIT may turn out to be inadequate or wrong. Are some of its assumptions, including those of a discrete, finite set of “atomic” units of cause–effect power, incompatible with current physics [32, 62] (but see [63–66])? Are its axiomatic basis and the formulation of axioms as postulates sound and unique? And, most critically, can IIT survive the results of empirical investigations assessing the relationship between the quantity and quality of consciousness and its substrate in the brain?

Supporting information

S1 Text. Resolving ties in the IIT algorithm. Operational process for resolving ties due to maxima / minima in the IIT algorithm.

(PDF)

S2 Text. Comparison to IIT 1.0–3.0 and subsequent publications. Summary of the changes in IIT 4.0 relative to earlier versions of the theory.

(PDF)

S3 Text. Analytical results for the number and integrated information of relations. Statement and proof of theorems describing the number of relations and the sum of their integrated information, $\Sigma\varphi_r$.

(PDF)

S1 Fig. IIT Algorithm. Visual summary of the algorithm for identifying complexes and unfolding cause–effect structures.

(PDF)

S1 Notes. Footnotes.

(PDF)

Author Contributions

Conceptualization: Larissa Albantakis, Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M. Haun, William Marshall, Alireza Zaeemzadeh, Melanie Boly, Bjørn E. Juel, Jeremiah Hendren, Jonathan P. Lang, Giulio Tononi.

Formal analysis: Larissa Albantakis, Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M. Haun, William Marshall, William G. P. Mayner, Alireza Zaeemzadeh.

Funding acquisition: Larissa Albantakis, William Marshall, Giulio Tononi.

Investigation: Larissa Albantakis, Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M. Haun, William Marshall, William G. P. Mayner, Alireza Zaeemzadeh, Bjørn E. Juel, Shuntaro Sasai, Keiko Fujii, Isaac David.

Methodology: Larissa Albantakis, Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M. Haun, William Marshall, William G. P. Mayner, Alireza Zaeemzadeh, Shuntaro Sasai, Keiko Fujii, Giulio Tononi.

Project administration: Jonathan P. Lang, Giulio Tononi.

Software: William G. P. Mayner, Isaac David.

Supervision: Larissa Albantakis, Giulio Tononi.

Validation: Larissa Albantakis.

Visualization: Larissa Albantakis, Matteo Grasso.

Writing – original draft: Larissa Albantakis, Giulio Tononi.

Writing – review & editing: Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M. Haun, William Marshall, William G. P. Mayner, Alireza Zaeemzadeh, Bjørn E. Juel, Isaac David, Jeremiah Hendren, Jonathan P. Lang.

References

1. Ellia F, Hendren J, Grasso M, Kozma C, Mindt G, P Lang J, M Haun A, Albantakis L, Boly M, and Tononi G. Consciousness and the fallacy of misplaced objectivity. *Neuroscience of Consciousness*. 2021; 2021(2):1–12. <https://doi.org/10.1093/nc/niab032> PMID: 34667639
2. Nagel T. What is it like to be a bat? *The philosophical review*. 1974; 83(4):435–450. <https://doi.org/10.2307/2183914>
3. Tononi G. Integrated information theory. *Scholarpedia*. 2015; 10(1):4164. <https://doi.org/10.4249/scholarpedia.4164>
4. Tononi G, Boly M, Massimini M, Koch C. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*. 2016; 17(7):450–461. <https://doi.org/10.1038/nrn.2016.44> PMID: 27225071
5. Tononi G, Sporns O. Measuring information integration. *BMC neuroscience*. 2003; 4(31):1–20. <https://doi.org/10.1186/1471-2202-4-31> PMID: 14641936
6. Tononi G. An information integration theory of consciousness. *BMC neuroscience*. 2004; 5:42. <https://doi.org/10.1186/1471-2202-5-42> PMID: 15522121
7. Balduzzi D, Tononi G. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput Biol*. 2008; 4(6):e1000091. <https://doi.org/10.1371/journal.pcbi.1000091> PMID: 18551165
8. Oizumi M, Albantakis L, Tononi G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*. 2014; 10(5):e1003588. <https://doi.org/10.1371/journal.pcbi.1003588> PMID: 24811198
9. Balduzzi D, Tononi G. Qualia: the geometry of integrated information. *PLoS computational biology*. 2009; 5(8):e1000462. <https://doi.org/10.1371/journal.pcbi.1000462> PMID: 19680424
10. Albantakis L, Marshall W, Hoel E, Tononi G. What caused what? A quantitative account of actual causation using dynamical causal networks. *Entropy*. 2019; 21(5):459. <https://doi.org/10.3390/e21050459> PMID: 33267173
11. Haun AM, Tononi G. Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy*. 2019; 21(12):1160. <https://doi.org/10.3390/e21121160>
12. Barbosa LS, Marshall W, Albantakis L, Tononi G. Mechanism Integrated Information. *Entropy*. 2021; 23(3):362. <https://doi.org/10.3390/e23030362> PMID: 33803765
13. Marshall W, Grasso M, Mayner WG, Zaeemzadeh A, Barbosa LS, Chastain E, et al. System Integrated Information. *Entropy*. 2023; 25. <https://doi.org/10.3390/e25020334> PMID: 36832700
14. Barbosa LS, Marshall W, Streipert S, Albantakis L, Tononi G. A measure for intrinsic information. *Scientific Reports*. 2020; 10(1):18803. <https://doi.org/10.1038/s41598-020-75943-4> PMID: 33139829

15. Intrinsic Ontology Wiki;. Available from: <https://centerforsleepandconsciousness.psychiatry.wisc.edu/intrinsic-ontology-wiki/>.
16. Albantakis L. Integrated information theory. In: Overgaard M, Mogensen J, Kirkeby-Hinrup A, editors. *Beyond Neural Correlates of Consciousness*. Routledge; 2020. p. 87–103.
17. Grasso M, Albantakis L, Lang JP, Tononi G. Causal reductionism and causal structures. *Nature Neuroscience*. 2021; 24(10):1348–1355. <https://doi.org/10.1038/s41593-021-00911-8> PMID: 34556868
18. Tononi G, Albantakis L, Boly M, Cirelli C, Koch C. Only what exists can cause: An intrinsic view of free will. 2022.
19. Tononi G, Koch C. Consciousness: here, there and everywhere? *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2015; 370:20140167–. <https://doi.org/10.1098/rstb.2014.0167> PMID: 25823865
20. Findlay G, Marshall W, Albantakis L, Mayner WGP, Koch C, Tononi G. Dissociating Intelligence from Consciousness in Artificial Systems – Implications of Integrated Information Theory. In: *Proceedings of the 2019 Towards Conscious AI Systems Symposium, AAAI SSS19; 2019 and forthcoming*.
21. Albantakis L, Prentner R, Durham I. Measuring the integrated information of a quantum mechanism. *Entropy*. 2023; 25.
22. Massimini M, Ferrarelli F, Huber R, Esser SK, Singh H, Tononi G. Breakdown of cortical effective connectivity during sleep. *Science*. 2005; 309(5744):2228–2232. <https://doi.org/10.1126/science.1117256> PMID: 16195466
23. Casarotto S, Comanducci A, Rosanova M, Sarasso S, Fecchio M, Napolitani M, et al. Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of Neurology*. 2016; 80(5):718–729. <https://doi.org/10.1002/ana.24779> PMID: 27717082
24. Comolatti, R et al. Why does time feel flowing?; in preparation.
25. Grasso, M et al. How do phenomenal objects bind general concepts with particular features?; in preparation.
26. Janzing D, Balduzzi D, Grosse-Wentrup M, Schölkopf B. Quantifying causal influences. *The Annals of Statistics*. 2013; 41(5):2324–2358. <https://doi.org/10.1214/13-AOS1145>
27. Ay N, Polani D. Information Flows in Causal Networks. *Advances in Complex Systems*. 2008; 11(01):17–41. <https://doi.org/10.1142/S0219525908001465>
28. Pearl J. *Causality: models, reasoning and inference*. vol. 29. Cambridge Univ Press; 2000.
29. Albantakis L, Tononi G. Causal Composition: Structural Differences among Dynamically Equivalent Systems. *Entropy* 2019, Vol 21, Page 989. 2019; 21(10):989.
30. Cooper J. *Plato: Complete Works*. Hackett; 1997.
31. Tillemans T. Dharmak?rti. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*. Spring 2021 ed. Metaphysics Research Lab, Stanford University; 2021.
32. Barrett AB, Mediano PAM. The phi measure of integrated information is not well-defined for general physical systems. *Journal of Consciousness Studies*. 2019; 26(1-2):11–20.
33. Hoel EP, Albantakis L, Marshall W, Tononi G. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*. 2016; 2016(1). <https://doi.org/10.1093/nc/niw012> PMID: 30788150
34. Marshall W, Albantakis L, Tononi G. Black-boxing and cause-effect power. *PLOS Computational Biology*. 2018; 14(4):e1006114. <https://doi.org/10.1371/journal.pcbi.1006114> PMID: 29684020
35. Hoel EP, Albantakis L, Tononi G. Quantifying causal emergence shows that macro can beat micro. *PNAS*. 2013; 110(49):19790–19795. <https://doi.org/10.1073/pnas.1314922110> PMID: 24248356
36. Krohn S, Ostwald D. Computing integrated information. *Neuroscience of Consciousness*. 2017; 2017(1). <https://doi.org/10.1093/nc/nix017> PMID: 30042849
37. Mayner WGP, Marshall W, Albantakis L, Findlay G, Marchman R, Tononi G. PyPhi: A toolbox for integrated information theory. *PLoS Computational Biology*. 2018; 14(7):e1006343. <https://doi.org/10.1371/journal.pcbi.1006343> PMID: 30048445
38. Kanwisher N. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*. 2010; 107(25):11163–11170. <https://doi.org/10.1073/pnas.1005062107>
39. Ponce CR, Xiao W, Schade PF, Hartmann TS, Kreiman G, Livingstone MS. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*. 2019; 177(4):999–1009. <https://doi.org/10.1016/j.cell.2019.04.005> PMID: 31051108
40. Khosla M, Wehbe L. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*. 2022.

41. Mainen ZF, Sejnowski TJ. Reliability of spike timing in neocortical neurons. *Science*. 1995; 268(5216):1503–1506. <https://doi.org/10.1126/science.7770778> PMID: 7770778
42. Hires SA, Gutnisky DA, Yu J, O'Connor DH, Svoboda K. Low-noise encoding of active touch by layer 4 in the somatosensory cortex. *eLife*. 2015; 4:e06619. <https://doi.org/10.7554/eLife.06619> PMID: 26245232
43. Nolte M, Reimann MW, King JG, Markram H, Muller EB. Cortical reliability amid noise and chaos. *Nature Communications*. 2019; 10(1):1–15. <https://doi.org/10.1038/s41467-019-11633-8> PMID: 31439838
44. Lemon R, Edgley S. Life without a cerebellum. *Brain*. 2010; 133(3):652–654. <https://doi.org/10.1093/brain/awq030> PMID: 20305277
45. Yu F, Jiang QJ, Sun XY, Zhang RW. A new case of complete primary cerebellar agenesis: clinical and imaging findings in a living patient. *Brain*. 2015; 138(6):e353–e353. <https://doi.org/10.1093/brain/awu239> PMID: 25149410
46. Steriade M, Nunez A, Amzica F. A novel slow (< 1 Hz) oscillation of neocortical neurons in vivo: depolarizing and hyperpolarizing components. *Journal of Neuroscience*. 1993; 13(8):3252–3265. <https://doi.org/10.1523/JNEUROSCI.13-08-03252.1993> PMID: 8340806
47. Pigorini A, Sarasso S, Proserpio P, Szymanski C, Arnulfo G, Casarotto S, et al. Bistability breaks-off deterministic responses to intracortical stimulation during non-REM sleep. *Neuroimage*. 2015; 112:105–113. <https://doi.org/10.1016/j.neuroimage.2015.02.056> PMID: 25747918
48. Middleton FA, Strick PL. Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Research Reviews*. 2000; 31(2-3):236–250. [https://doi.org/10.1016/S0165-0173\(99\)00040-5](https://doi.org/10.1016/S0165-0173(99)00040-5) PMID: 10719151
49. Foster NN, Barry J, Korobkova L, Garcia L, Gao L, Becerra M, et al. The mouse cortico–basal ganglia–thalamic network. *Nature*. 2021; 598(7879):188–194. <https://doi.org/10.1038/s41586-021-03993-3> PMID: 34616074
50. Zaeemzadeh, A et al. Upper Bounds for Integrated Information; in preparation.
51. Boly M, Massimini M, Tsuchiya N, Postle BR, Koch C, Tononi G. Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. *Journal of Neuroscience*. 2017; 37(40):9603–9613. <https://doi.org/10.1523/JNEUROSCI.3218-16.2017> PMID: 28978697
52. Watakabe A, Skibbe H, Nakae K, Abe H, Ichinohe N, Rachmadi MF, et al. Local and long-distance organization of prefrontal cortex circuits in the marmoset brain. *bioRxiv*. 2022.
53. Hanson JR, Walker SI. Formalizing falsification for theories of consciousness across computational hierarchies. *Neuroscience of Consciousness*. 2021; 2021(2). <https://doi.org/10.1093/nc/niab014> PMID: 34377534
54. Krohn K, Rhodes J. Algebraic Theory of Machines. I. Prime Decomposition Theorem for Finite Semi-groups and Machines. *Transactions of the American Mathematical Society*. 1965; 116:450. <https://doi.org/10.1090/S0002-9947-1965-0188316-1>
55. Albantakis L, Tononi G. The Intrinsic Cause-Effect Power of Discrete Dynamical Systems—From Elementary Cellular Automata to Adapting Animats. *Entropy*. 2015; 17(8):5472–5502. <https://doi.org/10.3390/e17085472>
56. Moyal R, Fekete T, Edelman S. Dynamical Emergence Theory (DET): A Computational Account of Phenomenal Consciousness. *Minds and Machines*. 2020; 30(1):1–21. <https://doi.org/10.1007/s11023-020-09516-9>
57. Melloni L, Mudrik L, Pitts M, Koch C. Making the hard problem of consciousness easier. *Science*. 2021; 372(6545):911–912. <https://doi.org/10.1126/science.abj3259> PMID: 34045342
58. Sarasso S, Casali AG, Casarotto S, Rosanova M, Sinigaglia C, Massimini M, et al. Consciousness and complexity: a consilience of evidence. *Neuroscience of Consciousness*. 2021; 7(2):1–24.
59. Sarasso S, D'Ambrosio S, Fecchio M, Casarotto S, Viganò A, Landi C, et al. Local sleep-like cortical reactivity in the awake brain after focal injury. *Brain*. 2020; 143(12):3672–3684. <https://doi.org/10.1093/brain/awaa338> PMID: 33188680
60. Song C, Haun AM, Tononi G. Plasticity in the structure of visual space. *Eneuro*. 2017; 4(3). <https://doi.org/10.1523/ENEURO.0080-17.2017> PMID: 28660245
61. Albantakis L, Hintze A, Koch C, Adami C, Tononi G. Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. *PLoS computational biology*. 2014; 10(12):e1003966. <https://doi.org/10.1371/journal.pcbi.1003966> PMID: 25521484
62. Carroll S. Consciousness and the Laws of Physics. *Journal of Consciousness Studies*. 2021; 28(9):16–31. <https://doi.org/10.53765/20512201.28.9.016>

63. Zanardi P, Tomka M, Venuti LC. Towards Quantum Integrated Information Theory. arXiv. 2018;1806.01421.
64. Kleiner J, Tull S. The Mathematical Structure of Integrated Information Theory. *Frontiers in Applied Mathematics and Statistics*. 2021; 6:74. <https://doi.org/10.3389/fams.2020.602973>
65. Esteban FJ, Galadí JA, Langa JA, Portillo JR, Soler-Toscano F. Informational structures: A dynamical system approach for integrated information. *PLOS Computational Biology*. 2018; 14(9):e1006154. <https://doi.org/10.1371/journal.pcbi.1006154> PMID: 30212467
66. Kalita P, Langa JA, Soler-Toscano F. Informational Structures and Informational Fields as a Prototype for the Description of Postulates of the Integrated Information Theory. *Entropy*. 2019; 21(5):493. <https://doi.org/10.3390/e21050493> PMID: 33267207