entropy

IMPACT FACTOR 2.738

Indexed in: PubMed

Article

Computing the Integrated Information of a Quantum Mechanism

Larissa Albantakis, Robert Prentner and Ian Durham

Topic Collection
Advances in Integrated Information Theory

Edited by
Dr. Larissa Albantakis, Dr. Matteo Grasso and Dr. Andrew Haun

MDPI

https://doi.org/10.3390/e25030449

entropy

MDPI

Article

Computing the Integrated Information of a Quantum Mechanism

Larissa Albantakis [1,2,*], Robert Prentner [2,3] and Ian Durham [2,4]

[1] Department of Psychiatry, University of Wisconsin-Madison, Madison, WI 53719, USA
[2] Association for Mathematical Consciousness Science, 80539 Munich, Germany
[3] Munich Center for Mathematical Philosophy, Ludwig-Maximilians-University, 80539 Munich, Germany
[4] Department of Physics, Saint Anselm College, Manchester, NH 03102, USA
[*] Correspondence: albantakis@wisc.edu

Abstract: Originally conceived as a theory of consciousness, integrated information theory (IIT) provides a theoretical framework intended to characterize the compositional causal information that a system, in its current state, specifies about itself. However, it remains to be determined whether IIT as a theory of consciousness is compatible with quantum mechanics as a theory of microphysics. Here, we present an extension of IIT's latest formalism to evaluate the mechanism integrated information (φ) of a system subset to discrete, finite-dimensional quantum systems (e.g., quantum logic gates). To that end, we translate a recently developed, unique measure of intrinsic information into a density matrix formulation and extend the notion of conditional independence to accommodate quantum entanglement. The compositional nature of the IIT analysis might shed some light on the internal structure of composite quantum states and operators that cannot be obtained using standard information-theoretical analysis. Finally, our results should inform theoretical arguments about the link between consciousness, causation, and physics from the classical to the quantum.

Keywords: causal analysis; causation; quantum information theory; entanglement structure; multivariate interaction

1. Introduction

Integrated information theory [1–4] stands out as one theory of consciousness that explicitly proposes a formal framework for identifying conscious systems. Specifically, IIT provides requirements about the intrinsic causal structure of a system that supports consciousness based on the essential ("phenomenal") properties of experience. Its formal framework evaluates the causal powers that a set of interacting physical units exerts on itself in a compositional manner [1,4–7].

IIT does not presuppose that consciousness arises at the level of neurons rather than atoms, molecules, or larger brain areas but assumes causation to be a central concept for analyzing a physical system across the hierarchy from the microphysical to the macroscopic [2,8–11]. One prediction of IIT is that consciousness appears at the level of organization at which the intrinsic causal powers of a system are maximized [2]. To that end, IIT offers a formal framework for causal emergence that compares the amount of integrated information of macroscopic causal models to their underlying microscopic system descriptions [9,10].

Nevertheless, IIT's causal framework has been formalized for discrete dynamical systems with macroscopic, possibly irreversible, cognitive/computational systems in mind [1,4,12–14]. Accordingly, in prior studies [8–10], micro-level systems corresponded to classical causal networks [15,16], constituted of individual, conditionally independent physical units that can (in principle) be manipulated and whose states can be observed. Thus, it remains to be determined whether IIT is compatible with quantum mechanics [17,18], especially because it is still contested whether causality plays a fundamental role in physics, and particularly in quantum physics [19,20].

check for updates

Citation: Albantakis, L.; Prentner, R.; Durham, I. Computing the Integrated Information of a Quantum Mechanism. *Entropy* **2023**, *25*, 449. https://doi.org/10.3390/e25030449

Academic Editor: Georg Northoff

Received: 3 January 2023
Revised: 22 February 2023
Accepted: 1 March 2023
Published: 3 March 2023

Here, we are interested in the question of whether it is possible to apply or extend the causal framework of IIT to quantum systems, starting with IIT's measure of mechanism integrated information (φ) [4,6]. Several attempts to apply the general principles of IIT to quantum systems have recently been proposed [21–23]. Of these, the work by Zanardi et al. [22] comes closest to a direct translation of the previous version of the theory ("IIT 3.0") [1] into a quantum-mechanical framework. However, this translation is not unique, does not converge to the classical formalism for essentially classical state updates, and also does not explicitly take the philosophical grounding of IIT as a theory of consciousness into account.

Our objective is to accurately transform the various steps of the IIT formalism in its latest iteration ("IIT 4.0") [4,6] to be applicable to both (macroscopic) classical and (microscopic) quantum systems. As a first step, here we propose an extension of the IIT formalism to evaluate the integrated information (φ) of a mechanism within a system [6] to quantum mechanisms (e.g., quantum logic gates). To enable a direct quantitative comparison between macroscopic and microscopic systems, quantum integrated information should converge to the classical formulation if the quantum system under consideration has a classical analog. (This means that we should get the same quantitative results when we analyze, e.g., a reversible logic gate applied to a classical basis state using the quantum or classical formalism.) Our main contributions, of merit beyond the scope of IIT, are (1) the translation of a newly defined, unique measure of intrinsic information [6,24] to a quantum density matrix formalism, and (2) a formulation of the causal constraints specified by a partial quantum state. To that end, we extend the notion of conditional independence and causal marginalization [16] to accommodate quantum entanglement. In the results section, we will apply our theoretical developments to classical computational gates and their quantum analogs (such as the CNOT gate), as well as quantum states and gates without a classical counterpart. The additional challenges of evaluating the integrated information of an entire quantum system will be outlined in the discussion.

While our investigation is based on IIT's formal framework, it raises questions that apply to any theory of consciousness and its relation to (micro) physics [21,25,26]. However, we also want to emphasize that this work is not concerned with the question of whether biological systems (in particular, the brain) should be treated quantum-theoretically or classically. The question of whether a theory of consciousness, such as IIT, is generally applicable across microscopic and macroscopic scales and thus consistent with our knowledge of microphysics is important in either case.

Our work is also not directly related to the potential role of consciousness in quantum measurements and the operational collapse of the wave function [27–29], although we briefly discuss several difficulties in applying IIT's causal analysis to measurement dynamics. In contrast to quantum theories of consciousness, such as "Orch OR" [30,31] we do not mean to suggest that consciousness depends on quantum-specific phenomena, such as the collapse or "orchestrated reduction" of the wave function, nor on entanglement, and there are arguments against a significant role of quantum effects in macroscopic brain processes, including consciousness [32,33] (but see [34–36]). Irrespective of this, at the finest level of description, the brain (and everything else) is a quantum system. However, the contents of our experiences seem to correlate with macroscopic neural mechanisms rather than microphysical processes. Our objective is to provide the tools to investigate and compare candidate classical and quantum systems within the framework of IIT, but also more generally, in terms of their informational, computational, and causal properties. At the very least, our results should inform theoretical arguments about the link between consciousness, causation, and physics from the classical to the quantum [37]. Finally, the compositional nature of the IIT analysis might also shed some light on the internal structure of composite quantum states and operators that cannot be obtained using standard information-theoretical analysis. To that end, we provide python code to analyze quantum mechanisms of two and three qubits, available at https://github.com/Albantakis/QIIT (accessed on 30 December 2022).

2. Theory

The purpose of IIT's formal analysis is to evaluate the irreducible causal information that a system in a particular state specifies about itself. Notably, IIT's notion of causal information differs from other information-theoretical measures in multiple ways: it is intrinsic (evaluated from the perspective of a mechanism within the system), state-dependent (evaluated for particular states, not state averages), causal (evaluated against all possible counterfactuals of a system transition [15,16]), and irreducible (evaluated against a partition of the mechanism into independent parts). Moreover, the IIT analysis is *compositional* [5]: instead of only analyzing the system as a whole or only its elementary components, any system subset counts as a candidate *mechanism* that may specify its own irreducible cause and effect within the system. The IIT analysis thus evaluates the irreducible cause-effect information (φ) of every subset of units within the system [6], which amounts to "unfolding" the system's cause-effect structure.

In the following, we will extend IIT's φ-measure, the integrated information of a mechanism, to be applicable to finite-dimensional quantum systems. While the full IIT analysis assumes a dynamical system of interacting units, mechanism integrated information (φ) can be evaluated in a straightforward manner for any type of input-output logic, such as sets of logic gates or whole computational circuits, as well as information channels (see Figure 1, as an example). For a classical template of our quantum version of mechanism integrated information (φ) we follow Barbosa et al. [6], including minor updates within the most recent formulation "IIT 4.0" [4], which is briefly reviewed in the following. As a result, the quantum integrated information of a mechanism, as defined below coincides, with the classical measure [4,6] if the quantum system under consideration has a classical analog.

2.1. Classical Systems

In the canonical IIT formalism, a (classical) physical system S of n interacting units is defined as a stochastic system $S = \{S_1, S_2, \dots, S_n\}$ with finite, discrete state space $\Omega_S = \prod_i \Omega_{S_i}$ and current state $s_t \in \Omega_S$ [6] that evolves according to a transition probability function

$$\mathcal{T}_S \equiv p(s_{t+1} \mid s_t) = \Pr(S_{t+1} = s_{t+1} \mid S_t = s_t), \quad s_t, s_{t+1} \in \Omega_S, \tag{1}$$

with the additional requirement that S corresponds to a causal network [6]. This implies that the conditional probabilities $p(s_{t+1}|s_t)$ are well-defined for all possible states

$$\exists\, p(s_{t+1}|s_t) \,\forall\, s_t, s_{t+1} \in \Omega_S, \tag{2}$$

with $p(s_{t+1}|s_t) = p(s_{t+1}|do(s_t))$ [15,16,38,39], where the "do-operator" $do(s_t)$ indicates that s_t is imposed by intervention. Moreover, the individual random variables $S_i \in S$ are assumed to be conditionally independent of each other given the preceding state of S,

$$p(s_{t+1} \mid s_t) = \prod_{i=1}^{n} p(s_{i,t+1}|s_t), \tag{3}$$

which has to be revisited in the quantum case. The canonical IIT formalism does not extend to systems with infinite state space described in continuous time. In discrete systems, instantaneous interactions are associated with classical uncertainty (due to incomplete knowledge) and are discounted in IIT because those are not intrinsic to the system. Therefore, Equation (3) holds from the intrinsic perspective [10].

If S is an open system within a larger universe U with current state $u_t \in \Omega_U$, variables $W = U \setminus S$ are treated as fixed background conditions throughout the causal analysis (see [4,40] for details).

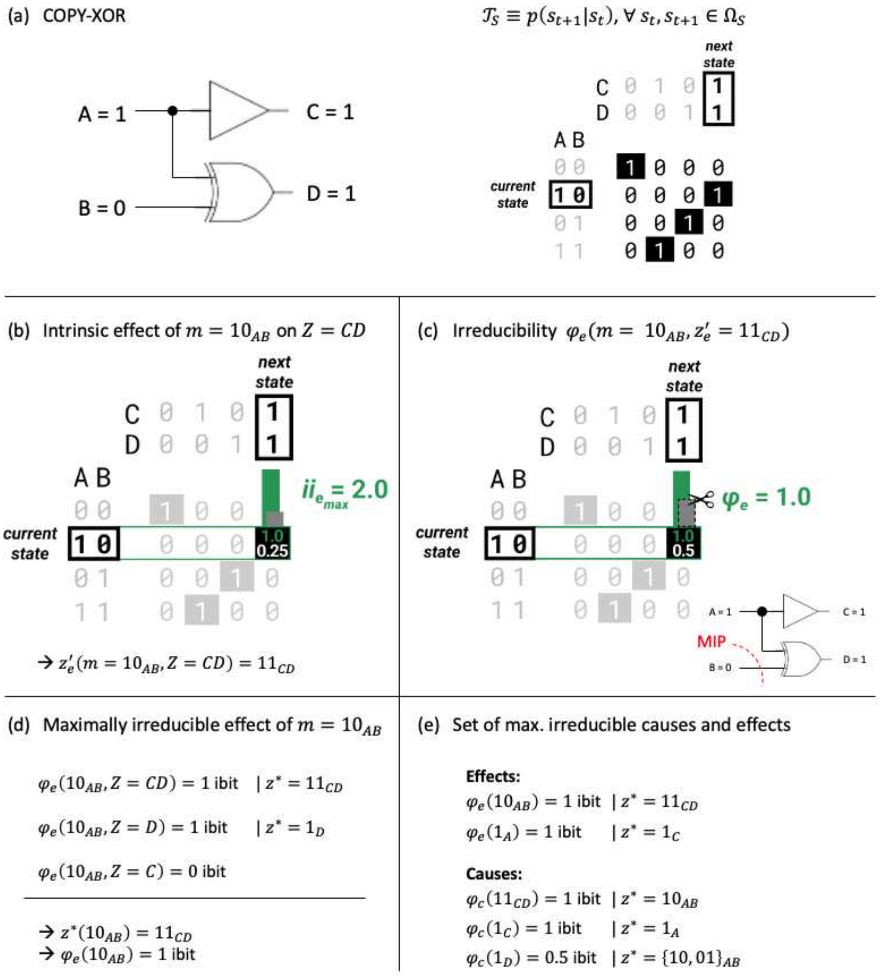
(a) COPY-XOR
$\mathcal{T}_S \equiv p(s_{t+1}|s_t), \forall s_t, s_{t+1} \in \Omega_S$
A = 1
B = 0
C = 1
D = 1
next state
current state
C
D
A B
(b) Intrinsic effect of $m = 10_{AB}$ on $Z = CD$
$ii_{e_{max}} = 2.0$
1.0
0.25
$\rightarrow z'_e(m = 10_{AB}, Z = CD) = 11_{CD}$
(c) Irreducibility $\varphi_e(m = 10_{AB}, z'_e = 11_{CD})$
$\varphi_e = 1.0$
1.0
0.5
MIP
(d) Maximally irreducible effect of $m = 10_{AB}$
$\varphi_e(10_{AB}, Z = CD) = 1$ ibit $| z^* = 11_{CD}$
$\varphi_e(10_{AB}, Z = D) = 1$ ibit $| z^* = 1_D$
$\varphi_e(10_{AB}, Z = C) = 0$ ibit
$\rightarrow z^*(10_{AB}) = 11_{CD}$
$\rightarrow \varphi_e(10_{AB}) = 1$ ibit
(e) Set of max. irreducible causes and effects
Effects:
$\varphi_e(10_{AB}) = 1$ ibit $| z^* = 11_{CD}$
$\varphi_e(1_A) = 1$ ibit $| z^* = 1_C$
Causes:
$\varphi_c(11_{CD}) = 1$ ibit $| z^* = 10_{AB}$
$\varphi_c(1_C) = 1$ ibit $| z^* = 1_A$
$\varphi_c(1_D) = 0.5$ ibit $| z^* = \{10, 01\}_{AB}$

Figure 1. Outline of the IIT analysis applied to a classical COPY-XOR gate. (**a**) The COPY-XOR gate and its (deterministic) transition probability function $\mathcal{T}_S$ depicted by a probability matrix. To avoid a proliferation of subscripts, in the following we use different letters to denote inputs and outputs. For conceptual ease, A/C and B/D can (but do not have to) be interpreted as the same physical units before and after the update. Unit C is a copy of the input bit A, and D corresponds to an XOR function of both input bits (A, B). For input state $AB = 10$ (also denoted by 10_{AB}), the COPY-XOR gate outputs $CD = 11$ (denoted by 11_{CD}). (**b**) Based on $\mathcal{T}_S$, we can identify the intrinsic effect of a mechanism M in its current state m over a purview Z as the effect state z'_e with maximal intrinsic effect information ii_e. For $m = 10_{AB}$ and $Z = CD$, the intrinsic effect is $z'_e = 11_{CD}$. (**c**) Next, we assess the irreducibility of the intrinsic effect by computing the integrated information $\varphi_e(m, Z)$ over the minimum partition (MIP). (**d**) To identify the maximally irreducible effect of a mechanism m, we compare $\varphi_e(m, Z)$ across all possible effect purviews Z. Here, the maximally irreducible effect of $m = 10_{AB}$ is $z^*_e = 11_{CD}$ because it specifies a maximum of φ_e and is the largest purview that does so (see text for details). (**e**) For a given system, we identify all maximally irreducible causes and effects. Given the input state $AB = 10$, the classical IIT analysis identifies two irreducible effects; the first-order mechanism 1_A specifies the effect 1_C, and the second-order mechanism 10_{AB} specifies the effect 11_{CD}. Given the output state $CD = 11$, the IIT analysis identifies three irreducible mechanisms, including mechanism 1_D with purview 10_{AB} or 01_{AB} (which are tied). Both intrinsic information (ii) and integrated information (φ) are quantified in "ibit" units (see text below).

A mechanism $M \subseteq S$ is a subset of the system S with current state $m_t \in \Omega_M$. The intrinsic information that a mechanism M in state m_t specifies over a "purview" $Z_{t\pm1} \subseteq S$, is defined by a difference measure $ii(m_t, Z_{t\pm1})$, which quantifies how much m_t constrains the state of $Z_{t\pm1}$ compared to chance, but also takes its *selectivity* into account (how much the mechanism specifies a particular state of $Z_{t\pm1}$) [4,6]. The mechanism's integrated information $\varphi(m_t, Z, \theta)$ is then evaluated over the maximal cause and effect states $z'_{c/e}$ identified by the intrinsic information measure. It quantifies how much the mechanism m_t constrains $z'_{c/e}$ as *one* mechanism, compared to a partition θ

$$\theta = \{(M^{(1)}, Z^{(1)}), (M^{(2)}, Z^{(2)}), \ldots, (M^{(k)}, Z^{(k)})\}, \tag{4}$$

of the mechanism and purview into k independent parts [1,6]. Below we will define all relevant quantities for computing the mechanism integrated information $\varphi(m_t)$ following Barbosa et al. [6] with minor updates from [4]. Figure 1 outlines the steps of IIT's causal analysis for a simple example system, a COPY-XOR gate.

2.1.1. Cause and Effect Repertoires

How the state of a mechanism $M = m$ constrains the possible states of a purview Z is captured by a product probability distribution $\pi(Z|m)$, which can be computed from the system's transition probability function (Equation (1)) [1,6,16]. Specifically, $\pi_c(Z|m) = \pi(Z_{t-1}|m_t)$ is the "cause repertoire" of m over Z, and $\pi_e(Z|m) = \pi(Z_{t+1}|m_t)$ is the "effect repertoire". Without loss of generality, in what follows, we will focus on the effects of m_t on purviews $Z = Z_{t+1}$ and omit update indices ($t-1, t, t+1$) unless necessary.

To capture the constraints on Z that are due to the mechanism in its state ($M = m$) and nothing else, it is important to remove any contributions to the repertoire from outside the mechanism. This is performed by "causally marginalizing" all variables in $X = S \backslash M$ [1,6,16]. When evaluating the constraints of m onto a single unit $Z_i \in Z$, causal marginalization amounts to imposing a uniform distribution as $p(X_t)$. The effect repertoire of a single unit $Z_i \in Z$ is thus defined as

$$\pi_e(Z_i \mid m) = |\Omega_X|^{-1} \sum_{x_t \in \Omega_X} p(Z_{i,t+1} \mid m_t, x_t). \tag{5}$$

In the general case of an effect repertoire over a set Z of $|Z|$ units (where $|Z|$ denotes the cardinality of the set of units Z), each $Z_i \in Z$ must receive independent inputs from units in $X = S \setminus M$ to discount correlations from units in X with divergent outputs to multiple units in Z (see Figure 2). Formally, this amounts to using product probabilities $\pi(Z|m)$ instead of standard conditional probabilities $p(Z|m)$ (again imposing a uniform interventional distribution). The effect repertoire over a set Z of $|Z|$ units Z_i is thus defined as the product of the effect repertoires over individual units

$$\pi_e(Z \mid m) = \bigotimes_{i=1}^{|Z|} \pi_e(Z_i \mid m), \tag{6}$$

where $\bigotimes$ is the Kronecker product of the probability distributions. As in [4], we define the unconstrained effect repertoire as the marginal distribution

$$\pi_e(Z; M) = |\Omega_M|^{-1} \sum_{m \in \Omega_M} \pi_e(Z \mid m). \tag{7}$$

The cause repertoire $\pi_c(Z|m)$ is obtained using Bayes' rule over the product distributions of the corresponding effect repertoire (for details, see [4,6]). The unconstrained cause repertoire $\pi_c(Z)$ is simply the uniform distribution over the states of Z.

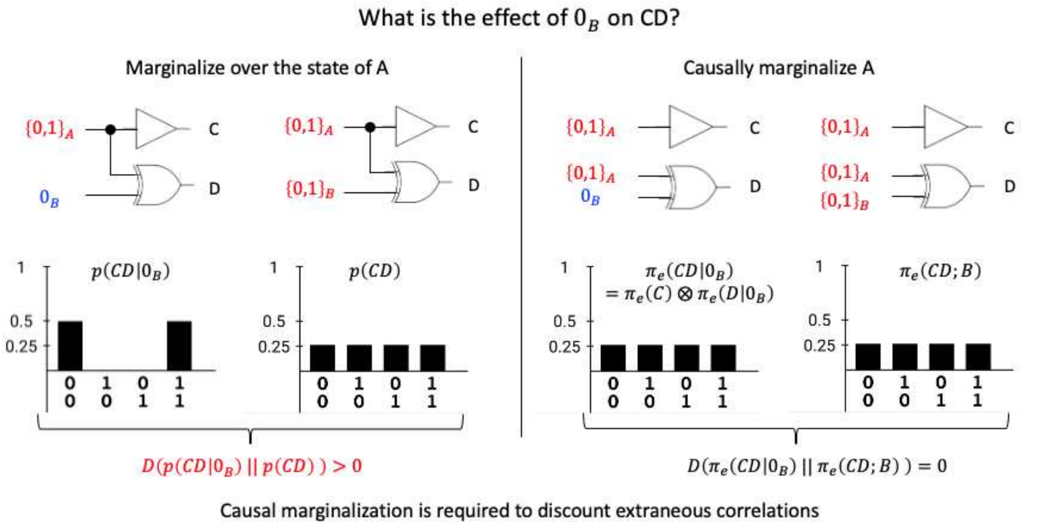
What is the effect of 0_B on CD?
Marginalize over the state of A
Causally marginalize A
$\{0,1\}_A$
0_B
$\{0,1\}_B$
C
D
$p(CD|0_B)$
$p(CD)$
$\pi_e(CD|0_B)$
$= \pi_e(C) \otimes \pi_e(D|0_B)$
$\pi_e(CD;B)$
1
0.5
0.25
0 1 0 1
0 0 1 1
$D(p(CD|0_B)\ ||\ p(CD)) > 0$
$D(\pi_e(CD|0_B)\ ||\ \pi_e(CD;B)) = 0$
Causal marginalization is required to discount extraneous correlations

Figure 2. Causal marginalization. Let us assume we want to identify the effect of the input bit $B = 0$ (or 0_B) on the output CD in the COPY-XOR system of Figure 1. Intuitively, by itself, 0_B does not have an effect on C, as it does not input into C. It also has no effect on D because, by itself, it specifies no information about the output state of the XOR D. However, simply marginalizing the input A (averaging over all possible input states of A while maintaining the common inputs from A to C and D) would result in a "spurious" correlation between the output bits that is not due to B, but instead due to the common inputs from A. Capturing the fact that 0_B, by itself, has no effect on CD, requires causal marginalization (independent marginal inputs to each unit in the effect purview).

2.1.2. Intrinsic Difference (ID)

The classical version of mechanism integrated information (φ) evaluates the difference between two probability distributions $P = [p_1, \ldots, p_N]$ and $Q = [q_1, \ldots, q_N]$ based on a newly developed information measure, the "intrinsic difference" (ID) [6,24]. The ID measure is uniquely defined based on three desired properties: *causality*, *specificity*, and *intrinsicality*, which align with the postulates of IIT [4,6,24]. Specifically,

$$\mathrm{ID}(P, Q) = \max_{\alpha}\left(p_\alpha \log\left(\frac{p_\alpha}{q_\alpha}\right)\right), \tag{8}$$

where α denotes a particular state in the distribution.

Formally, the ID is related to the Kullback–Leibler Divergence (KLD) or "relative entropy" measure,

$$\mathrm{KLD}(P, Q) = \sum_{\alpha} p_\alpha \log\left(\frac{p_\alpha}{q_\alpha}\right). \tag{9}$$

While the KLD can be viewed as an average of the point-wise mutual information $\log\left(\frac{p_\alpha}{q_\alpha}\right)$ across states, the ID is instead defined based on the state that maximizes the difference between distributions (specificity property). For fully selective distributions (there is one state with probability one), the ID thus coincides with the KLD and is additive. Otherwise, the ID is subadditive and decreases with indeterminism (intrinsicality property). As argued in [6], this allows the ID to capture the information specified by a mechanism within a particular system. (We refer to [4,6] for further discussion of the ID as the proper difference measure in the context of IIT.) From the perspective of a mechanism, the system has to be taken *as is* (intrinsic perspective), while the KLD evaluates information from the perspective of a channel designer with the possibility to perform error correction (extrinsic perspective) [24]. To highlight this difference, the unit assigned to the ID measure is labeled an "ibit" or "intrinsic bit". Logarithms are evaluated with base 2 throughout. Formally, the "ibit" corresponds to a point-wise information value measured in bits weighted by a probability. Note that while the ID (like the KLD) is, in principle, unbounded for arbitrary distributions, all IIT measures based on the ID are bounded, as demonstrated in [41].

2.1.3. Identifying Intrinsic Causes and Effects

Based on the intrinsic difference (8), the intrinsic effect information that the mechanism $M = m$ specifies over a purview Z can be quantified by comparing its effect repertoire $\pi_e(Z|m)$ to chance, that is, to the unconstrained effect repertoire $\pi_e(Z; M)$ (7),

$$ii_e(m, Z) = \mathrm{ID}(\pi_e(Z|m), \pi_e(Z; M)) \tag{10}$$

The specific state $z'_e \in \Omega_Z$ over which (10) is maximized corresponds to the intrinsic effect of the mechanism $M = m$ on the purview Z,

$$z'_e(m, Z) = \operatorname*{argmax}_{z \in \Omega_Z} \left(\pi_e(Z|m) \log \left(\frac{\pi_e(Z|m)}{\pi_e(Z; M)} \right) \right). \tag{11}$$

The intrinsic cause $z'_c(m, Z)$ is defined in the same way based on the respective cause repertoires. (Note that the definition of the intrinsic cause information ii_c and, consequently, also the integrated cause information φ_c, has been updated in [4] compared to [6]. However, this update of the classical formulation is of no consequence in the quantum case and is thus not further discussed herein).

2.1.4. Disintegrating Partitions

The integrated effect information $\varphi_e(m, Z, \theta)$ quantifies how much the mechanism m specifies the intrinsic effect $z'_e(m, Z)$ as *one* mechanism and is assessed by comparing the effect probability $\pi(z'_e \mid m)$ to a partitioned effect probability $\pi^\theta_e(z'_e \mid m)$ in which certain connections from M to Z are severed (causally marginalized).

Barbosa et al. [6] (see also [4,16]) define the set of possible partitions $\theta \in \Theta(M, Z)$ as

$$\Theta(M, Z) = \left\{ \{(M^{(i)}, Z^{(i)})\}_{i=1}^{k} \middle| k \in \{2, 3, 4, \dots\}, M^{(i)} \in \mathbb{P}(M), Z^{(i)} \in \mathbb{P}(Z), \right.$$

$$\left. \bigcup M^{(i)} = M, \bigcup Z^{(i)} = Z, Z^{(i)} \cap Z^{(j)} = M^{(i)} \cap M^{(j)} = \varnothing \; \forall \, i \neq j, M^{(i)} = M \Longrightarrow Z^{(i)} = \varnothing \right\}. \tag{12}$$

In words, for each $\theta \in \Theta(M, Z)$, it holds that $\{M^{(i)}\}$ is a partition of M and $\{Z^{(i)}\}$ is a partition of Z (as indicated in Equation (4)), but the empty set may also be used as a part ($\mathbb{P}$ denotes the powerset). However, if the whole mechanism is one part ($M^{(i)} = M$), then it must be cut away from the entire purview. This definition guarantees that any $\theta \in \Theta(M, Z)$ is a "disintegrating partition" of $\{M, Z\}$: it 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 $|M| = 1$, where again $|M|$ denotes the cardinality of the set of units M.

Given a partition $\theta \in \Theta(M, Z)$ constituted of k parts (see Equation (12)), we can define the partitioned effect repertoire

$$\pi^\theta_e(Z \mid m) = \bigotimes_{i=1}^{k} \pi_e(Z^{(i)} \mid m^{(i)}), \tag{13}$$

with $\pi(\varnothing|m^{(i)}) = \pi(\varnothing) = 1$. In the case of $m^{(i)} = \varnothing$, $\pi_e(Z^{(i)}|\varnothing)$ corresponds to the fully partitioned effect repertoire

$$\pi_e(Z \mid \varnothing) = \bigotimes_{i=1}^{|Z|} \sum_{s_t \in \Omega_S} p(Z_{i,t+1} \mid s_t) |\Omega_S|^{-1}. \tag{14}$$

2.1.5. Mechanism Integrated Information

In all, the general form of $\varphi_e(m, Z, \theta)$ corresponds to that of the intrinsic difference ID (8), albeit over the specific effect state z'_e

$$\varphi_e(m, Z, \theta) = \varphi_e(m, z'_e, \theta) = \pi_e(z'_e \mid m) \log\left(\frac{\pi_e(z'_e \mid m)}{\pi^{\theta}_e(z'_e \mid m)}\right). \tag{15}$$

Quantifying the integrated effect information of a mechanism m_t within a system S, moreover, requires optimization across all possible partitions $\theta \in \Theta$ to identify the minimum partition (MIP)

$$\theta' = \underset{\theta \in \Theta(M,Z)}{\operatorname{argmin}} \frac{\varphi_e(m, Z, \theta)}{\max\limits_{\mathcal{T}'_S} \varphi_e(m, Z, \theta)}. \tag{16}$$

The normalization factor $\max_{\mathcal{T}'_S} \varphi_e(m, Z, \theta)$ ensures that the minimum partition is evaluated against its maximum possible value across all possible system $\mathcal{T}'_S$ of the same dimensions as the original system. It was introduced in [4] and shown to correspond to the number of possible pairwise interactions affected by the partition.

The integrated effect information of a mechanism over a particular purview Z then corresponds to $\varphi_e(m, Z) = \varphi_e(m, Z, \theta')$ (which is not normalized, see [4]). Within system S, $\varphi_e(m)$ is then defined as the integrated effect information of m evaluated across all possible purviews, $Z \subseteq S$ with $\varphi_e(m) = \max\limits_{Z} \varphi_e(m, Z)$.

The effect purview

$$Z^*_e(m) = \underset{Z \subseteq S}{\operatorname{argmax}}\, \varphi_e(m, Z), \tag{17}$$

in state

$$z^*_e(m) = \underset{\{z'_e \mid Z \subseteq S\}}{\operatorname{argmax}}\, \varphi(m, Z = z'_e) = \underset{\{z'_e \mid Z \subseteq S\}}{\operatorname{argmax}} \left(\pi_e(z'_e \mid m) \log\left(\frac{\pi_e(z'_e \mid m)}{\pi^{\theta'}_e(z'_e \mid m)}\right)\right) \tag{18}$$

corresponds to the maximally irreducible intrinsic effect of $M = m$ within S.

To summarize,

$$\varphi_e(m) = \varphi(m, z^*_e) = \max_{Z \subseteq S} \left(\pi_e(z'_e \mid m) \log\left(\frac{\pi_e(z'_e \mid m)}{\pi^{\theta'}_e(z'_e \mid m)}\right)\right), \tag{19}$$

with θ' as in (16) and analogously for $\varphi_c(m)$.

Finally, the set of all irreducible causes and effects $\{z^*_{c/e} : m \subseteq s,\ \varphi_{c/e}(m) > 0\}$ within a system S in state s forms the basis of the system's state-dependent cause-effect structure.

(While the value $\varphi_e(m)$ is unique, there may be multiple purviews Z^*_e, or multiple states z^*_e within a purview Z^*_e, that maximize $\varphi_e(m)$ [4,6,42,43]. As outlined in IIT 4.0 [4], such ties in z^*_e are resolved according to a congruence requirement with the overall cause-effect state of the system and further eliminated by the "maximum existence principle" applied at the system level, selecting the z^*_e that maximizes the amount of structured information Φ within the system. Here, we apply the simplified criterion that larger purviews are selected in the case of ties across purviews with different numbers of units $|Z_e|$, as larger purviews typically allow for larger Φ values. Any remaining ties are reported in the examples below.)

2.2. Quantum Systems

Our objective is to define a quantum version of IIT's mechanism integrated information $\varphi(m)$ that is applicable to composite quantum systems and coincides with the classical measure [4,6] if there is a classical analog to the quantum system. To that end, we start with a discrete, composite quantum system Q in state $\rho = \sum_s |\psi_s\rangle\langle\psi_s|$, which can be pure or mixed and is described by its density matrix [22,23].

Q consists of n units $\mathcal{H}_1, \ldots, \mathcal{H}_n$, which are each described by a finite dimensional Hilbert space such that $\mathcal{H}_Q = \bigotimes_{i=1}^{n} \mathcal{H}_i$. Without loss of generality [14], we will focus on systems constituted of n qubits. The system's time evolution is defined by a completely positive (trace-preserving) linear map $\mathcal{T} = \{T_\alpha\}$ [44], as

$$\rho_{t+1} = \mathcal{T}(\rho_t) = \sum_\alpha T_\alpha \rho_t T_\alpha^\dagger. \tag{20}$$

Rather than evaluating quantum systems with specific observables, these "CPTP" maps can be interpreted as general, but finite, quantum information channels (where the Planck constant is absorbed in the evolution operator $\mathcal{T}$).

We will mainly consider unitary transformations (U)

$$\rho_{t+1} = U\rho_t U^\dagger, \tag{21}$$

where $U^\dagger U = 1$, which means that U is reversible and the inverse of U corresponds to its adjoint ($U^{-1} = U^\dagger$). However, we will also address quantum measurements $\mathcal{F} = \{F_\alpha\}$ with $\sum_\alpha F_\alpha^\dagger F_\alpha = I$, where the probability of obtaining the result α is given by $\Pr(\alpha) = tr(F_\alpha^\dagger F_\alpha \rho_t)$ in the discussion section. If Q is an open system with environment E, such that the joint system evolves under a unitary transformation, we can evaluate the subsystem Q by treating the environment E in its current state e_t as a fixed background condition (see Section 4.3 below).

A mechanism $M \subseteq Q$ is a subset of Q with current state $m = \rho_t^M = tr_{M'}(\rho_t)$ within the corresponding Hilbert space $\mathcal{H}_M = \bigotimes_{i \in M} \mathcal{H}_i$, where $M' = Q \setminus M$ and $tr_{M'}$ denotes the trace over the Hilbert space $\mathcal{H}_{M'}$.

The quantum integrated information of a mechanism M should quantify how much the state ρ_t^M constrains the state of a purview, a system subset $Z_{t\pm1} \subseteq Q$, before or after an update $\mathcal{T}$ of the system, compared to a partition θ of the mechanism and purview into k independent parts (Equation (4)). As above, we will omit the update indices $(t-1, t, t+1)$ unless necessary and focus on effects.

2.2.1. Quantum Cause and Effect Repertoires

To translate the cause and effect repertoires into a density matrix description, we first treat the special case of a single purview node $Z = Z_i$ with $|Z| = 1$, for which $\pi_e(Z|m) = p(Z_{t+1}|m_t)$ in the classical case. Replacing the probability distributions with the corresponding density matrices, we obtain

$$\pi_e(Z_i|m) = \rho_{t+1}^{Z_i|m} = tr_{Z_i'}\Big(\mathcal{T}(\rho^M \otimes \rho_{mm}^{M'})\Big), \tag{22}$$

where $'$ denotes the complement of a set in Q and $\rho_{mm}^{M'}$ is the maximally mixed state of $M' = Q \setminus M$ (see also [22,23]).

Next, we consider the case of purviews comprised of multiple units ($|Z| > 1$). In the classical case, units in M' may induce correlations between units in Z, as shown in Figure 2 by example of the COPY-XOR gate. The quantum equivalent of a classical COPY-XOR gate is the CNOT gate (Figure 3). For classical inputs, the CNOT behaves identically to the COPY-XOR gate and thus the same considerations apply. This means that, also in quantum systems, extraneous correlations should be discounted when evaluating the causal constraints of a system subset M, since they do not correspond to constraints due to the mechanism M itself. In the following, we will use $\rho_{t+1}^{Z|m}$ to denote $tr_{Z'}\Big(\mathcal{T}(\rho^M \otimes \rho_{mm}^{M'})\Big)$, while $\pi_e(Z|m)$ corresponds to the final effect repertoire, after discounting extraneous correlations.

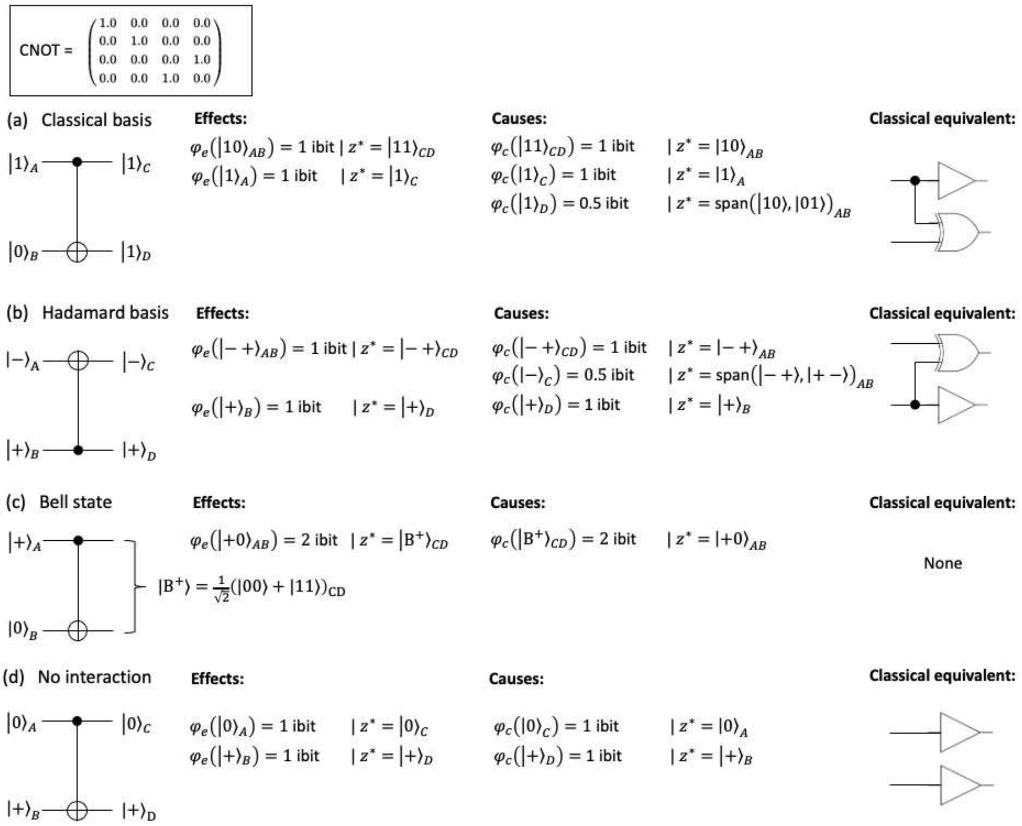
CNOT = $\begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \end{pmatrix}$
(a) Classical basis
$|1\rangle_A$
$|0\rangle_B$
$|1\rangle_C$
$|1\rangle_D$
Effects:
$\varphi_e(|10\rangle_{AB}) = 1$ ibit | $z^* = |11\rangle_{CD}$
$\varphi_e(|1\rangle_A) = 1$ ibit | $z^* = |1\rangle_C$
Causes:
$\varphi_c(|11\rangle_{CD}) = 1$ ibit | $z^* = |10\rangle_{AB}$
$\varphi_c(|1\rangle_C) = 1$ ibit | $z^* = |1\rangle_A$
$\varphi_c(|1\rangle_D) = 0.5$ ibit | $z^* = \mathrm{span}(|10\rangle, |01\rangle)_{AB}$
Classical equivalent:
(b) Hadamard basis
$|-\rangle_A$
$|+\rangle_B$
$|-\rangle_C$
$|+\rangle_D$
Effects:
$\varphi_e(|-+\rangle_{AB}) = 1$ ibit | $z^* = |-+\rangle_{CD}$
$\varphi_e(|+\rangle_B) = 1$ ibit | $z^* = |+\rangle_D$
Causes:
$\varphi_c(|-+\rangle_{CD}) = 1$ ibit | $z^* = |-+\rangle_{AB}$
$\varphi_c(|-\rangle_C) = 0.5$ ibit | $z^* = \mathrm{span}(|-+\rangle, |+-\rangle)_{AB}$
$\varphi_c(|+\rangle_D) = 1$ ibit | $z^* = |+\rangle_B$
Classical equivalent:
(c) Bell state
$|+\rangle_A$
$|0\rangle_B$
$|B^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{CD}$
Effects:
$\varphi_e(|+0\rangle_{AB}) = 2$ ibit | $z^* = |B^+\rangle_{CD}$
Causes:
$\varphi_c(|B^+\rangle_{CD}) = 2$ ibit | $z^* = |+0\rangle_{AB}$
Classical equivalent:
None
(d) No interaction
$|0\rangle_A$
$|+\rangle_B$
$|0\rangle_C$
$|+\rangle_D$
Effects:
$\varphi_e(|0\rangle_A) = 1$ ibit | $z^* = |0\rangle_C$
$\varphi_e(|+\rangle_B) = 1$ ibit | $z^* = |+\rangle_D$
Causes:
$\varphi_c(|0\rangle_C) = 1$ ibit | $z^* = |0\rangle_A$
$\varphi_c(|+\rangle_D) = 1$ ibit | $z^* = |+\rangle_B$
Classical equivalent:

Figure 3. CNOT gate. The CNOT operator is shown in the top box. (**a**) For a pure input state in the classical basis, we obtain the same results as in the classical case (Figure 1). (**b**) For a pure input state in the Hadamard basis, the role of the "control" (here B) and "target" (here A) is reversed compared to (**a**) (as indicated in the circuit diagram). (**c**) The CNOT is often used to produce a "Bell state" of two maximally entangled qubits. In this exclusively quantum scenario, only the second-order mechanisms $|+0\rangle_{AB}$ and $|B^+\rangle_{CD}$ specify an effect or cause, respectively. None of the subsets has any cause or effect information ($\varphi = 0$ ibit). (**d**) Conversely, given the input state $|0+\rangle_{AB}$, all second-order mechanisms are fully reducible ($\varphi = 0$ ibit) and only the first-order mechanisms specify causes and effects.

In the quantum case, units in Z may be correlated due to entanglement, which means quantum systems may violate the conditional independence assumption imposed for classical systems (Equation (3)). (Note that incomplete knowledge or a coarse-grained temporal scale can lead to a violation of conditional independence in a classical system, but those "instantaneous interactions" are not considered intrinsic to the system and are thus ignored in IIT's causal analysis [10]). Simply inserting Equation (22) into Equation (6) would inadvertently destroy correlations in Z that are due to entanglement (either preserved or produced during the transformation $\mathcal{T}$). In order to correctly capture correlations due to entanglement and discount extraneous correlations due to correlated "noise" from units in M', the entanglement structure of $\rho_{t+1}^{Z|m}$ must be taken into account.

The multipartite entanglement structure of an n-qubit *pure* state $|\psi\rangle$ can be identified through partial traces. Following [45], we define a partition $\mathcal{P}^r(V) = \{V^{(1)}, \ldots, V^{(r)}\}$ with $r = |\mathcal{P}^r| \leq n$, $\bigcup V^{(i)} = V$ and $V^{(i)} \bigcap V^{(j)} = \emptyset$ if $i \neq j$.

Definition 1. *An n-qubit pure state $|\psi\rangle$ is $\mathcal{P}^r$-separable iff it can be written as $|\psi\rangle = \bigotimes_{i=1}^{r} \left|\psi^{(i)}\right\rangle$.*

In the general case that $\rho_{t+1}^{Z|m}$ is a mixed state, it has to be decomposed into a convex mixture of pure states to identify its entanglement structure.

Definition 2. *An n-qubit mixed state ρ is $\mathcal{P}^r$-separable iff it can be decomposed into a convex mixture $\rho = \sum_s p_s |\psi_s\rangle\langle\psi_s|$, with $p_s \geq 0, \forall s$ and $\sum_s p_s = 1$, such that every $|\psi_s\rangle$ in the mixture is a $\mathcal{P}^r$-separable pure state $|\psi_s\rangle = \bigotimes_{i=1}^r \left|\psi_s^{(i)}\right\rangle$ under the same partition $\mathcal{P}^r$.*

Note that Definition 2 differs from that in [45], as we require the same partition $\mathcal{P}^r$ for all $|\psi_s\rangle$ in the mixture.

Definition 3. *Out of the set of partitions $\{\mathcal{P}^r\}_\rho = \{\mathcal{P}^r | \rho$ is $\mathcal{P}^r$-separable$\}$, we define the maximal partition $\mathcal{P}^*(\rho)$ as the one with the maximal number of parts $r^* = \max_{\mathcal{P}^r} r$ and $r^* = |\mathcal{P}^*| \leq n$.*

Definition 4. *Given the maximal partition $\mathcal{P}^*$ of $\rho_{t+1}^{Z|m}$, we can define the quantum effect repertoire of mechanism m over purview Z as*

$$\pi_e(Z \mid m) = \bigotimes_{i=1}^{r^*} \pi_e(Z^{(i)} \mid m) = \bigotimes_{i=1}^{r^*} \rho_{t+1}^{Z^{(i)}|m}. \tag{23}$$

The product in (23) is thus taken over the reduced density matrices of all subsets $Z^{(i)} \subseteq Z$ that are entangled within themselves but not entangled with the other qubits in Z. Note that $\mathcal{P}^*$ is a simple set partition and should not be confused with the disintegrating partitions $\Theta(M, Z)$ (12) used to evaluate the integrated information $\varphi(m, Z, \theta)$. Identifying the entanglement structure for multipartite mixed states remains an area of active research [46–48]. For two-qubit mixed states, separability can be determined using the Peres–Horodecki criterion of the positive partial transform [49,50]. For general bipartite systems, however, this criterion is only a necessary condition for separability [50] and may thus miss certain complex forms of entanglement [51]. See [46] for methods to detect entanglement in multipartite mixed states.

Several implications follow from the definition of the effect repertoire (23):

1. If $\rho_{t+1}^{Z|m}$ corresponds to a pure state, the purview qubits are fully determined by the mechanism qubits. Thus, $\rho_{t+1}^{Z|m}$ is not influenced by qubits outside of m. It follows that $\pi_e(Z|m) = \rho_{t+1}^{Z|m}$ if the latter is pure. This is analogous to the classical case, where $\pi_e(Z|m) = p(Z_{t+1}|m_t)$ if $p(Z_{t+1}|m_t)$ is deterministic.
2. Conceptually, entangled subsets are treated as indivisible units in the effect repertoire. If a purview is fully entangled, then $\pi_e(Z|m) = \rho_{t+1}^{Z|m}$.
3. Extraneous classical correlations are successfully discounted, which means they will not contribute to the integrated information of a mechanism (Figure 3).

The cause repertoire of a mechanism in state m over a purview Z also requires causal marginalization (independent noise applied to conditionally independent subsets) to isolate the causal constraints of m over Z. In the classical case, the cause repertoire is obtained by applying Bayes' rule to the effect product probabilities. The quantum case is more complex as the entanglement structure of ρ^M might need to be taken into account.

If $\mathcal{T}$ is a unitary transformation (21), the cause repertoire for any subset $m^{(i)} \in \mathcal{P}^*(\rho^M)$ that is, itself, mutually entangled (e.g., the subset could consist of an entangled pair of qubits) but is not entangled with units of other subsets (e.g., other qubits) can be obtained by applying the adjoint operator $\mathcal{T}^\dagger$

$$\pi_c(Z \mid m^{(i)}) = \rho_{t-1}^{Z|m^{(i)}} = tr_{Z'}\Big(\mathcal{T}^\dagger(\rho^{M^{(i)}} \otimes \rho_{mm}^{M'^{(i)}})\Big). \tag{24}$$

Definition 5. *Given the maximal partition $\mathcal{P}^*$ of ρ^M, we can define the quantum cause repertoire of mechanism m over purview Z as*

$$\pi_c(Z \mid m) = \frac{\prod_{i=1}^{r^*} \pi_c(Z \mid m^{(i)})}{tr\left(\prod_{i=1}^{r^*} \pi_c(Z \mid m^{(i)})\right)}. \tag{25}$$

Note that the product here is over parts of ρ^M, not of $\rho_{t-1}^{Z|m}$. This introduces an asymmetry in the formulation of cause and effect repertoires, as in the classical case [1,16]. This asymmetry is a direct implication of treating non-entangled subsets as "physical" causal units rather than abstract statistical variables. Causal units are conditionally independent in the present given the past, but not vice versa. This means that in the effect repertoire, purview subsets that are not entangled with other units are conditionally independent given the mechanism and independent noise from outside the mechanism (due to causal marginalization). By contrast, the cause repertoire is inferred from the conditionally independent mechanism subsets but is not itself conditionally independent. The set of effects specified by a quantum state ρ_t undergoing a unitary transformation (U) may thus differ from the set of causes specified by $\rho_{t+1} = U\rho_t U^\dagger$ (Figure 3). (The assumption of conditional independence, paired with causal marginalization, distinguishes IIT's causal analysis from standard information-theoretical analyses of information flow [16,39]).

As pointed out in [23], the quantum IIT formalism proposed by Zanardi et al. [22] does not include causal marginalization (which was formulated in terms of "virtual units" in [1]). We will show below that causal marginalization (Equations (23) and (25)) is necessary to isolate the causes and effects of system subsets in the quantum case—an observation that should be of relevance to the causal analysis of quantum systems beyond IIT.

2.2.2. Quantum Intrinsic Information (QID)

Our goal is to define a quantum version of the intrinsic difference measure, which coincides with the classical measure (8) [6] in the classical case. In quantum information theory, the classical definition of the KLD (9), or relative entropy, is extended from probability distributions to density matrices based on the von Neumann entropy. The quantum relative entropy of the density matrix ρ with respect to another density matrix σ is then defined as:

$$S(\rho||\sigma) = \mathrm{Tr}\rho \log \rho - \mathrm{Tr}\rho \log \sigma, \tag{26}$$

which coincides with the classical case if $\rho\sigma = \sigma\rho$. Unitary operations, including a change in basis, leave $S(\rho||\sigma)$ invariant [44]. Specifically, if ρ and σ are expressed as orthonormal decompositions $\rho = \sum_i p_i |i\rangle\langle i|$ and $\sigma = \sum_j q_j |j\rangle\langle j|$, we can write [52]

$$S(\rho||\sigma) = \sum_i p_i \left(\log(p_i) - \sum_j P_{ij} \log(q_j) \right), \tag{27}$$

where $P_{ij} = \langle i|j\rangle\langle j|i\rangle$. In this formulation, a quantum version of the intrinsic difference measure can be defined as

$$\mathrm{QID}(\rho||\sigma) = \max_i p_i \left(\log(p_i) - \sum_j P_{ij} \log(q_j) \right), \tag{28}$$

analogous to the classical measure. As for the relative entropy, $\mathrm{QID}(\rho||\sigma)$ coincides with the classical case if $\rho\sigma = \sigma\rho$, because, in that case, $P_{ij} = \delta_{ij}$. Moreover, $\mathrm{QID}(\rho||\sigma) = S(\rho||\sigma)$ if ρ is pure, as in the classical case for fully selective distributions. Otherwise, the QID is subadditive, as desired [24].

Zanardi et al. [22] proposed the trace distance as a measure of the cause/effect information based on its simplicity and widespread use in quantum-information theory. The trace distance quantifies the maximal difference in probability between two quantum states across all possible POVM measures [52], which is a useful quantity from the perspective of an experimenter. In contrast, QID is a measure of the *intrinsic* information of a quantum mechanism. Its value is maximized over the eigenvectors $\{|i\rangle\}$ of ρ (28). If ρ is pure, there is only one non-zero eigenvalue and the state identified by the QID measure is simply ρ. If ρ is mixed, the eigenvalue p_i that maximizes Equation (28) may be degenerate. In that case, the QID specifies the eigenspace spanned by the set of eigenvectors for which the difference between ρ and σ is maximal. Otherwise, the QID specifies the eigenvector of ρ with the optimal eigenvalue.

2.2.3. Identifying Intrinsic Causes and Effects

Equipped with the quantum intrinsic difference (QID) measure (28), the intrinsic effect information that the quantum mechanism $M = m$ specifies over a purview Z can be quantified as

$$ii_e(m, Z) = \text{QID}(\pi_e(Z|m), \pi_e(Z)), \tag{29}$$

where $\pi_e(Z) = \pi_c(Z) = \rho^Z_{mm}$ is the maximally mixed state in the quantum case.

Following on from Equation (28), with $\rho = \pi_e(Z \mid m) = \sum_i p_i|i\rangle\langle i|$ as the effect repertoire and $\sigma = \pi_e(Z) = \sum_j q_j|j\rangle\langle j| = \rho^Z_{mm}$ as the unconstrained effect repertoire, the intrinsic effect of mechanism m on purview Z is

$$\begin{aligned} z'_e(m, Z) &= \underset{i\in\mathcal{H}_Z}{\text{argmax}}\, p_i \left(\log p_i - \sum_j P_{ij} \log(q_j)\right) \\ &= \underset{i\in\mathcal{H}_Z}{\text{argmax}}\, p_i \left(\log p_i - \log |\mathcal{H}_Z|^{-1}\right), \end{aligned} \tag{30}$$

where $|\mathcal{H}_Z|$ denotes the cardinality of $\mathcal{H}_Z$. The intrinsic effect $z'_e(m, Z)$ is thus simply the eigenvector $|i\rangle$ of $\pi_e(Z|m)$ with the maximal eigenvalue. If the maximal eigenvalue of $\rho = \pi_e(Z \mid m)$ is degenerate, $z^*_e(m)$ corresponds to the subspace of $\mathcal{H}_{Z^*_e}$ spanned by the set of eigenvectors belonging to the maximal eigenvalue (and the same for the intrinsic cause $z'_c(m, Z)$ evaluated over $\pi_c(Z|m)$).

Note that, in the case that $\pi_e(Z|m)$ is a mixed quantum state (corresponding to a probability distribution with multiple possible effect states in the classical case), this means that the *intrinsic* effect $z'_e(m, Z)$ differs from $\rho = \pi_e(Z|m) = \sum_i p_i|i\rangle\langle i|$.

2.2.4. Disintegrating Partitions

As in the classical case, the quantum integrated information $\varphi(m, Z, \theta)$ is evaluated by comparing the effect repertoire $\pi_e(Z|m)$ to a partitioned effect repertoire $\pi^\theta_e(Z|m)$ (and analogously for $\varphi_c(m, Z, \theta)$).

The set of possible partitions $\theta \in \Theta(M, Z)$ is the same as for the classical case (Equation (12)). Likewise, the partitioned effect repertoire is defined as in (13), as a product over the parts in the partition. In the quantum case, $\pi_e(Z^{(i)}|\varnothing)$ corresponds to the maximally mixed state $\rho^{Z^{(i)}}_{mm}$. The partitioned cause repertoire is defined in the same way.

Note that the disintegrating partition $\theta \in \Theta(M, Z)$ (12) here is applied on top of $\mathcal{P}^*$ (Definition 3). Partitioning may thus affect entanglement within the repertoire. Conceptually, any entanglement in $\pi_e(Z \mid m)$ that is destroyed by the partition θ will count toward $\varphi_e(m, Z, \theta)$. Ultimately, however, $\varphi_e(m, Z)$ is again evaluated over θ' (16), the minimum information partition (MIP). This means that everything else being equal, partitions that affect entanglement less are more likely to correspond to the MIP.

2.2.5. Quantum Mechanism Integrated Information

Having identified the specific effect state z'_e as an eigenstate $|i\rangle$ of $\rho = \pi_e(Z|m)$, the integrated effect information $\varphi(m, Z, \theta)$ is evaluated as the QID$(\rho||\sigma)$ over that eigenstate, such that

$$\varphi(m, Z, \theta) = \varphi(m, z'_e, \theta) = p_i \left(\log p_i - \sum_j P_{ij} \log(p^\theta_j)\right), \tag{31}$$

where $\sigma = \pi_e^\theta(Z|m) = \sum_j p_j^\theta |j\rangle\langle j|$ is now the partitioned effect repertoire.

As above, quantifying the integrated effect information $\varphi_e(m)$ of a mechanism m within a quantum system Q requires a search over all possible partitions $\theta \in \Theta(M, Z)$ to identify the MIP, and a search across all possible purviews $Z \subseteq Q$, such that

$$\varphi_e(m) = \max_{Z \subseteq Q} \varphi_e(m, Z) = \max_{Z \subseteq Q} \varphi(m, Z, \theta'), \tag{32}$$

as in (19), with θ' as in (16), and analogously for $\varphi_c(m)$.

The maximally irreducible effect purview $Z_e^*(m)$

$$Z_e^*(m) = \operatorname*{argmax}_{Z \subseteq Q} \varphi_e(m, Z) \tag{33}$$

again corresponds to the subset of Q upon which the mechanism $M = m$ has the maximally irreducible intrinsic effect z_e^*, which corresponds to the eigenstate of $\rho = \pi_e(Z^*|m)$ that maximizes Equation (30), or the eigenspace spanned by a set of eigenvectors corresponding to a degenerate maximal eigenvalue.

As in the classical case, Z_e^* is not necessarily unique, and we again choose the larger purview in the case of a tie between purviews of different sizes (see above). Any remaining ties are reported in the examples below.

2.2.6. The Intrinsic Structure of a Quantum System

Standard approaches for studying the causal or informational properties of a system typically assume either a reductionist perspective (focused on individual units) or a holistic perspective (describing the system as a whole). As the units in a quantum system can be entangled, focusing on individual units is ill-suited at the quantum level. However, a purely holistic description of a quantum system will still miss differences in the internal structure of a quantum state (see the comparison between the maximally entangled GHZ-type and W-type states below [53]).

In IIT, causation is neither reductionist nor holistic but compositional: the IIT analysis considers the intrinsic causes and effects of every subset within a system and quantifies their irreducibility as $\varphi_{c/e}(m)$ [5]. As a result, it can elucidate the internal structure of composite quantum states and operators, as we will show in the next section.

We note that, typically, the IIT analysis assumes a current system state s_t and identifies its compositional causes at $t-1$ and effects at $t+1$. A subset $m \subseteq s$ with an irreducible cause and effect forms a "causal distinction" within the system s, where $\varphi(m) = \min(\varphi_c(m), \varphi_e(m))$ is the integrated (cause-effect) information of m.

According to IIT, the phenomenal experience of a physical system S in state s is identical to its cause-effect structure, composed of a system's causal distinctions and their relations [54]. Unfolding the full cause-effect structure requires assessing the integrated (cause-effect) information $\varphi(m)$ of every subset of units $m \subseteq s$.

For ease of demonstration, in the following, we will instead evaluate examples of system transitions from state t to $t+1$ and identify the intrinsic effects of the system in state s_t and the intrinsic causes of the system in state s_{t+1} (see also [16]).

3. Results

For a direct comparison between classical and quantum systems, we will focus our attention on computational quantum systems (see [55] for an overview and comparison to classical systems), constituted of a finite number of quantum units with a finite-dimensional Hilbert space, evolving in discrete updates according to unitary transformations, expressed in the computational (or "classical") basis unless stated otherwise.

To compute classical IIT quantities, we made use of the openly available PyPhi python toolbox, developed by the Tononi lab [13,14], using the "iit-4.0" feature branch with standard IIT 4.0 settings. To compute quantum IIT results, we implemented a QIIT toolbox (https://github.com/Albantakis/QIIT, accessed on 30 December 2022), applicable to unitary quantum mechanisms of two and three qubits.

3.1. CNOT

3.1.1. Classical Case

As a first example, we will evaluate the "controlled-NOT" (CNOT) gate. Classically, the CNOT gate corresponds to a reversible XOR gate, with a COPY operation performed on the first input bit (A) and an XOR operation comparing the two input bits, A and B, as the second output (Figure 1). For instance, the input state $AB = (1, 0)$ leads to the output $CD = (1, 1)$. In what follows, we will abbreviate the states of system subsets (mechanisms and purviews) by the state plus a set subscript, for example, 10_{AB} for $AB = (1, 0)$.

Given the input state $AB = (1, 0)$, the IIT analysis identifies two irreducible mechanisms, one first-order and one second-order mechanism. The mechanism 1_A specifies the effect purview 1_C with $\varphi = 1$ ibit; the second-order mechanism 10_{AB} specifies the effect purview 11_{CD} also with $\varphi = 1$ ibit (while there is a tie with the effect 1_D, we choose the larger purview, as described above). Notably, 0_B, by itself (with A replaced by independent noise), does not specify any information about the next state of CD (Figure 2). While this conclusion should be straightforward, it relies on the use of product probabilities instead of simple conditional probabilities (6). The latter would mistakenly count the correlation between C and D as an effect of B, although it is actually due to the common input of A.

In contrast to 0_B on the effect side, 1_D on the cause side specifies irreducible cause information about the previous state of AB in addition to 1_C and 11_{CD}, albeit only $\varphi_c(1_D) = 0.5$ ibit due to the remaining uncertainty about the state of AB (note the quantitative difference between the ID measure, (8) and the KLD (9), which would return a value of 1 bit).

3.1.2. Quantum Case

For a CNOT gate with the input state $\rho^{\mathrm{AB}} = |10\rangle\langle 10|$ (or $|10\rangle_{AB}$), we obtain the same results as for a COPY-XOR gate with input state $AB = (1, 0)$ using the formalism outlined above (Figure 3a). With essentially classical inputs, the CNOT gate thus reproduces the intrinsic causal structure of the classical COPY-XOR gate.

To that end, it was necessary to discount the spurious correlation between qubits A and B through product distributions (23). This demonstrates that standard conditional probabilities are insufficient to identify the causes and effects of system subsets also in the quantum case.

Note that for the CNOT gate, the role of the "control" (COPY) and the "target" (XOR) qubit changes depending on the input state, which is not true for the COPY-XOR gate. For an input state in the Hadamard basis, e.g., $|-+\rangle_{AB}$, information seems to flow from B to C, not A to D, as for a classical input. Accordingly, the quantum IIT analysis now identifies $|+\rangle_B$ and $|-+\rangle_{AB}$ as irreducible mechanisms with $\varphi = 1$ ibit, while $|-\rangle_A$ by itself does not specify any effect information (Figure 3b). However, $|+\rangle_C$ does specify irreducible cause information about AB.

In quantum systems, CNOT is often used to produce the maximally entangled Bell state $|B^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. $CD = |B^+\rangle$ results from the input state $AB = |+0\rangle$, a transition for which there is no classical circuit equivalent [56]. In this case, the quantum IIT analysis identifies only the second-order mechanisms (constituted of two qubits) $|+0\rangle_{AB}$ and $|B^+\rangle_{CD}$ with $\varphi = 2$ ibits each. Individual qubits specify no cause or effect information (Figure 3c). An analogous result obtains for the Bell state as the input to the CNOT gate.

Finally, with $AB = |0+\rangle$ as the input, there appears to be no interaction between qubits, and the quantum IIT analysis only identifies first-order mechanisms on the cause and effect side (Figure 3d).

3.1.3. Mixed States and Extensions to Larger Systems

The purpose of the IIT analysis is to evaluate the cause-effect power of a system in its current state. Evaluating statistical ensembles is conceptually not in line with the theory.

Accordingly, the classical IIT analysis always assumes a particular (fully determined) state for the mechanism m. However, in quantum mechanics, mixed states not only describe statistical ensembles but also subsets of entangled pure states.

If we apply an even mixture $\rho^{AB} = 0.5 * (|00\rangle\langle 00| + |11\rangle\langle 11|)$ to the CNOT gate, we obtain $\rho^{CD} = 0.5 * (|00\rangle\langle 00| + |10\rangle\langle 10|)$ as a result. In this case, only the second-order mechanism $m = \rho^{AB}$ has an irreducible effect with $\varphi_e = 1.0$ ibit over $z^* = |0\rangle_D$. There is no effect on C, as C by itself is undetermined (maximally mixed). In turn, only $|0\rangle_D$ specifies an irreducible cause with $\varphi(|0_D\rangle) = 0.5$ ibits over purview $Z^* = AB$ with z^* corresponding to the subspace spanned by $|00\rangle_{AB}$ and $|11\rangle_{AB}$ (Figure 4a). Note the difference in the causal analysis of the Bell state $|B^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ above, where $|{+0}\rangle_{AB}$ and $|B^+\rangle_{CD}$ both specified second-order mechanisms with $\varphi = 2$ ibit each.

The same transition may also be described as part of a larger system of three qubits. To that end, we can extend the CNOT gate by an identity operator acting on the additional qubit (Figure 4b), which may stand for the environment.

Assuming that the three qubits (ABC) are initially in a maximally entangled GHZ state [57], the state after applying I⊗CNOT leaves the first two qubits (DE) maximally entangled, while the third qubit (F) is in state $|0\rangle_F$. The causal analysis of the three-qubit system reveals additional causes and effects that span all three qubits but also includes the cause and effect identified for the mixed two-qubit subsystem evaluated in Figure 4a.

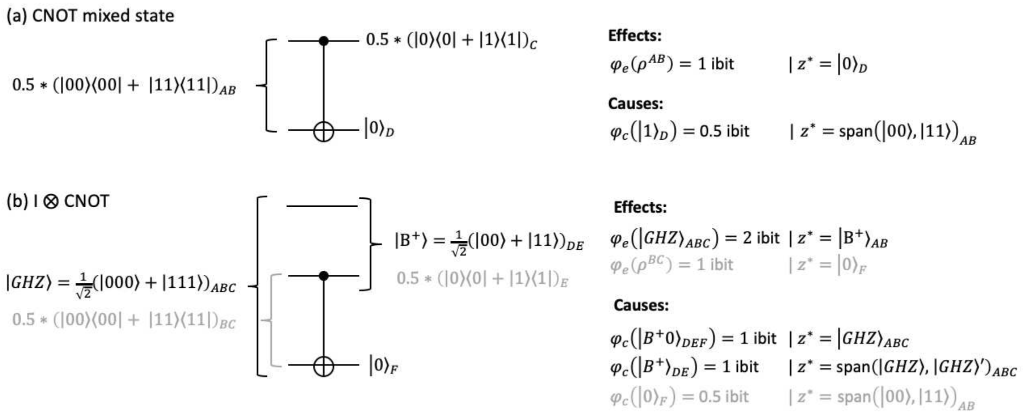
(a) CNOT mixed state
$0.5 * (|00\rangle\langle 00| + |11\rangle\langle 11|)_{AB}$
$0.5 * (|0\rangle\langle 0| + |1\rangle\langle 1|)_C$
$|0\rangle_D$
Effects:
$\varphi_e(\rho^{AB}) = 1$ ibit
$| z^* = |0\rangle_D$
Causes:
$\varphi_c(|1\rangle_D) = 0.5$ ibit
$| z^* = \mathrm{span}(|00\rangle, |11\rangle)_{AB}$
(b) I ⊗ CNOT
$|GHZ\rangle = \frac{1}{\sqrt{2}}(|000\rangle + |111\rangle)_{ABC}$
$0.5 * (|00\rangle\langle 00| + |11\rangle\langle 11|)_{BC}$
$|B^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{DE}$
$0.5 * (|0\rangle\langle 0| + |1\rangle\langle 1|)_E$
$|0\rangle_F$
Effects:
$\varphi_e(|GHZ\rangle_{ABC}) = 2$ ibit
$| z^* = |B^+\rangle_{AB}$
$\varphi_e(\rho^{BC}) = 1$ ibit
$| z^* = |0\rangle_F$
Causes:
$\varphi_c(|B^+0\rangle_{DEF}) = 1$ ibit
$| z^* = |GHZ\rangle_{ABC}$
$\varphi_c(|B^+\rangle_{DE}) = 1$ ibit
$| z^* = \mathrm{span}(|GHZ\rangle, |GHZ\rangle')_{ABC}$
$\varphi_c(|0\rangle_F) = 0.5$ ibit
$| z^* = \mathrm{span}(|00\rangle, |11\rangle)_{AB}$

Figure 4. Mixed states and entanglement with the environment. (**a**) IIT analysis of the CNOT gate with a mixed input state $\rho^{AB} = 0.5 * (|00\rangle\langle 00| + |11\rangle\langle 11|)$. (**b**) It is possible to describe the mixed state as a pure state entangled with the environment. Analyzing such an extended system for the case in (**a**), the cause and effect of the subsystem are preserved in the larger system (gray), but we obtain additional causes and effects that span all three qubits (black). $|GHZ\rangle'$ denotes a maximally entangled superposition of states $|001\rangle$ and $|110\rangle$.

3.1.4. Intrinsic Structure Due to Entanglement

The IIT analysis evaluates the potential causes and effects of a system in a state before and after an update of the system (20). In the classical case, there is no instantaneous interaction between the units of a system (which corresponds to the conditional independence assumption (3) [16]). In the quantum case, however, entanglement between qubits can lead to additional intrinsic structure (see also [28]). To identify the intrinsic structure of a quantum state that is due to entanglement, we can assume $\mathcal{T} = I$ (the identity operator) in (20). In that case, causes and effects are equivalent and should be viewed as constraints of the quantum state onto itself.

For classical states, causal analysis identifies only first-order constraints for $\mathcal{T} = I$ (Figure 5a). The entanglement of tripartite quantum states is not a trivial extension of the entanglement of bipartite systems [58]. In addition to biseparable states (A-BC, B-AC, C-AB), there exist two classes of genuine tripartite entanglement: GHZ-type and W-type states [53]. For the GHZ-state, $|\mathrm{GHZ}\rangle = \frac{1}{\sqrt{2}}(|000\rangle + |111\rangle)$, all subsystems correspond to unentangled, evenly mixed states of zeros and ones. For the W-state, $|W\rangle = \frac{1}{\sqrt{3}}(|001\rangle + |010\rangle + |100\rangle)$, all bipartite subsystems remain entangled with different probabilities of zeros and ones. The difference between these two states is clearly identified by the IIT analysis. While the GHZ-state only specifies a third-order constraint without any substructure, the W-state has full structure with intrinsic constraints on all subsets.

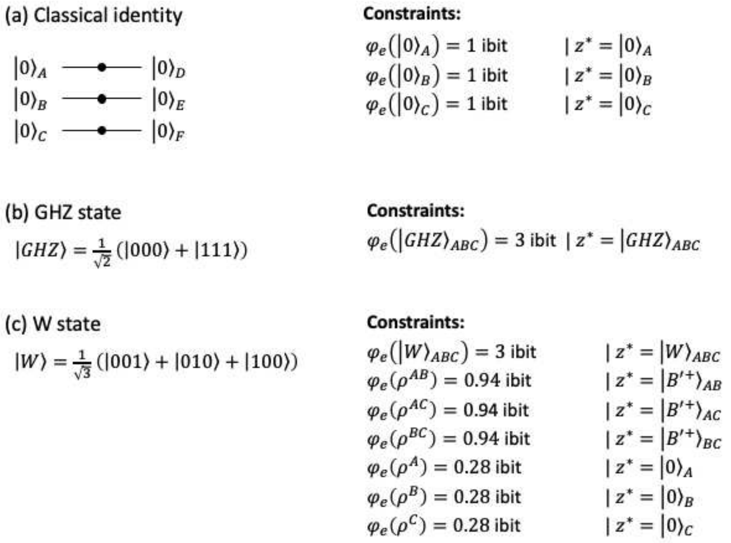
(a) Classical identity
$|0\rangle_A$
$|0\rangle_B$
$|0\rangle_C$
$|0\rangle_D$
$|0\rangle_E$
$|0\rangle_F$
Constraints:
$\varphi_e(|0\rangle_A) = 1$ ibit | $z^* = |0\rangle_A$
$\varphi_e(|0\rangle_B) = 1$ ibit | $z^* = |0\rangle_B$
$\varphi_e(|0\rangle_C) = 1$ ibit | $z^* = |0\rangle_C$
(b) GHZ state
$|GHZ\rangle = \frac{1}{\sqrt{2}}(|000\rangle + |111\rangle)$
Constraints:
$\varphi_e(|GHZ\rangle_{ABC}) = 3$ ibit | $z^* = |GHZ\rangle_{ABC}$
(c) W state
$|W\rangle = \frac{1}{\sqrt{3}}(|001\rangle + |010\rangle + |100\rangle)$
Constraints:
$\varphi_e(|W\rangle_{ABC}) = 3$ ibit | $z^* = |W\rangle_{ABC}$
$\varphi_e(\rho^{AB}) = 0.94$ ibit | $z^* = |B'^{+}\rangle_{AB}$
$\varphi_e(\rho^{AC}) = 0.94$ ibit | $z^* = |B'^{+}\rangle_{AC}$
$\varphi_e(\rho^{BC}) = 0.94$ ibit | $z^* = |B'^{+}\rangle_{BC}$
$\varphi_e(\rho^{A}) = 0.28$ ibit | $z^* = |0\rangle_A$
$\varphi_e(\rho^{B}) = 0.28$ ibit | $z^* = |0\rangle_B$
$\varphi_e(\rho^{C}) = 0.28$ ibit | $z^* = |0\rangle_C$

Figure 5. Intrinsic structure of three-qubit states. (**a**) Classical states specify first-order constraints under an identity function (equivalent to three classical COPY gates). (**b**) The maximally entangled GHZ-state only specifies a third-order constraint. (**c**) By contrast, the W-state, which is also maximally entangled, specifies constraints of all orders. Subsets $m \subseteq s$ of the W-state are indicated by ρ^m. The remaining units $s \setminus m$ are traced out. $|B'^{+}\rangle$ indicates a superposition of $|10\rangle$ and $|01\rangle$.

4. Discussion

Our goal in this study was to extend the mathematical formalism of IIT from discrete, classical dynamical systems to finite-dimensional quantum systems, starting with IIT's mechanism integrated information $\varphi(m)$ [6]. To that end, we translated IIT's intrinsic difference measure [6,24] into a density matrix formalism, and extended the notion of conditional independence and causal marginalization [16] to allow for quantum entanglement. Our results demonstrate that it is possible to extend the applicability of IIT's formal framework to finite-dimensional quantum systems evolving according to unitary transformations, such that the quantum formulation converges to the classical formulation for essentially classical state updates (as demonstrated by the example of the CNOT gate, Figures 1 and 3). In the following, we will compare our work to previous attempts of applying IIT to quantum systems [21–23], discuss several difficulties in applying IIT's causal analysis to measurement dynamics and highlight several limitations and implications of our QIIT formalism.

4.1. Comparison with Previous Approaches

Potential extensions of IIT to quantum systems have been explored in [21–23,28,29]. Of these, only Zanardi et al. [22] aimed for a direct translation of the IIT formalism (specifically, "IIT 3.0" [1]) from a classical into a quantum-mechanical framework. As demonstrated by Kleiner and Tull [23], the quantum IIT formalism proposed in [22] captures the higher-level mathematical structure of the canonical framework (IIT 3.0). However, it does not converge to the classical IIT framework and thus does not allow for quantitative comparison across quantum and classical systems. Among other differences, Zanardi et al. omitted the causal marginalization of variables outside the cause or effect repertoires and across partitions. As we have shown above (Figures 2 and 3), causal marginalization is necessary to identify the causal constraints specific to a subset of variables within the system in the quantum case. Paired with the conditional independence assumption, this also implies that the IIT formalism does not obey time-reversal symmetry, even when applied to unitary transformations (see also [5] for classical reversible systems).

Compared to [22], we have, moreover, incorporated several updates of the IIT formalism from "IIT 3.0" [1] to "IIT 4.0" [4,6]. These include an updated partitioning scheme [6,16], as well as a novel measure of intrinsic information based on the intrinsic difference (ID) introduced in [24]. While Zanardi et al. [22] used the trace distance to quantify φ, we have developed a quantum version of the novel intrinsic information measure, starting from the quantum relative entropy between two density matrices. In combination with the implementation of causal marginalization in quantum systems, the QIIT formalism proposed above thus converges to the classical version for essentially classical state updates. This means that the quantum and classical formalism yield the same quantitative results for classical, reversible logic operations applied to classical basis states.

While [22,23] are mainly concerned with the mathematical framework of IIT, refs. [28,29] apply the notion of integrated information within the context of a consciousness-induced collapse model of quantum mechanics. To that end, Chalmers and McQueen [29] utilize the QIIT framework proposed in [22]. Kremnitzer and Ranchin [28] present an independent quantum-integrated information measure based on quantum relative entropy. However, their measure applies to the quantum state itself and does not take the dynamics of the quantum system into account. Our work has a different focus. IIT does not require a role for consciousness in the collapse of the wave function (but see Section 4.2 below). Conversely, our work also does not assign special explanatory power to quantum effects over classical cause-effect power when it comes to consciousness and its contents (which stands in contrast to quantum theories of consciousness, such as Orch OR [30,31]). As shown in Figure 5, entangled subsystems may contribute to the integrated information and cause-effect structure of a quantum system even in the absence of causal interactions (which unfold over the state update). However, according to IIT, entanglement or non-separability, more generally [59], are not required for integration (see IIT's integration postulate [4]). Moreover, entanglement should not affect the overall bounds on a system's integrated information (φ_s) [40] or structured information (Φ) [4] as derived in [41]. Based on empirical work investigating the spatio-temporal scale of human consciousness and its contents, IIT would predict a maximum of integrated information at the level of neural interactions in certain parts of the cortico-thalamic system [2] (see Section 4.4).

Finally, Tegmark [21] leans on the general principles of IIT's approach to understanding and explaining consciousness in physical systems and addresses the so-called "quantum factorization problem" [60] using generalized measures of information integration. While we regard the quantum factorization problem as a serious issue, it is beyond the scope of this work. Our assumed starting point is a particular density matrix that undergoes a particular unitary transformation (21). While the QID measure (28) is basis independent, a system's cause-effect structure and the mechanism integrated information values $\varphi(m)$ of its subsets $m \subseteq Q$ will typically change under an additional unitary transformation and also depend on the specific factorization of the Hilbert space ($\mathcal{H}_Q = \bigotimes_{i=1}^{n} \mathcal{H}_i$) [61].

4.2. Measurement Dynamics

The dynamics of a quantum measurement can be described by a quantum operator $\mathcal{F} = \{F_\alpha\}$ with $\sum_\alpha F_\alpha^\dagger F_\alpha = I$. While the output of a unitary transformation is a density matrix corresponding to a pure or mixed quantum state, the outcome of a measurement is probabilistic with $\Pr(\alpha) = tr(F_\alpha^\dagger F_\alpha \rho_t)$ for measurement outcome α [52].

The IIT analysis evaluates the potential effects and potential causes of a mechanism in a state. From the perspective of the quantum state ρ_t being measured, the measurement outcome is still unknown. The effect repertoire of the quantum state ρ_t directly before the measurement (23) could thus be computed from

$$\rho_{t+1} = \sum_\alpha F_\alpha \rho_t F_\alpha^\dagger, \tag{34}$$

following Equation (20). The density matrix ρ_{t+1} then corresponds to a mixed state, that is, a probability distribution of possible measurement outcomes.

Measurement dynamics become problematic if we want to evaluate the quantum state directly after the measurement (and the same considerations apply in the case of a spontaneous collapse). Here, the cause repertoire has to be computed from the perspective of the quantum state post measurement ρ^{α}_{t+1}, corresponding to a particular measurement outcome α

$$\rho^{\alpha}_{t+1} = \frac{F_\alpha \rho_t F^{\dagger}_{\alpha}}{tr(F^{\dagger}_{\alpha} F_\alpha \rho)}. \tag{35}$$

Since measurements are not unitary transformations, the adjoint operator $\mathcal{T}^{\dagger}$ is not the same as the inverse $\mathcal{T}^{-1}$. For this reason, we cannot use Equation (24) to compute the cause repertoire of ρ^{α}_{t+1} (note that the same holds for prior proposals [22,23]).

In the classical case, the cause repertoire of an irreversible mechanism can be computed using Bayes' Rule [4,6]. However, in the quantum case, all information about the basis of the original quantum state before the measurement is lost, which means that there are infinitely many possible past states. While different past states should still be more or less likely, we do not know of any available method for obtaining a probability distribution of possible causes in this case.

That said, the amount of cause information specified by a post-measurement state ρ^{α}_{t+1} depends on the way the measurement dynamics are conceptualized, and thus on the specific interpretation of quantum theory applied. While ρ^{α}_{t+1} specifies (almost) no cause information under spontaneous collapse theories, the case may be quite different for deterministic hidden variable theories. No, or very low, cause information at the quantum level would imply that quantum systems are poor substrates for consciousness and may offer room for macro level descriptions to reach maximal values of integrated information, as predicted by IIT.

Finally, the technical difficulties introduced by probabilistic measurement dynamics would naturally be avoided by so-called "no-collapse" models of quantum mechanics, such as the Many-Worlds Interpretation [62]. However, theories that rely only on a density matrix encoding the state of the universe and a unitary transformation determining its time-evolution [21,61] face a different issue when it comes to identifying conscious entities through causal, informational, or computational means. If applied at the fundamental level, any entities obtained would correspond to subsets of the universal density matrix, never subsets within individual "branches" only (see for example Figure 3c). While the QIIT measures (and other quantities) could formally be applied within a branch, there is no principled justification for doing so from the perspective of a fundamental theory of consciousness (note that the notion of decoherence cannot resolve this issue).

4.3. Formal Considerations and Limitations

Formally, the restriction to unitary transformations eliminated differences between the unconstrained cause and effect repertoire that commonly arise in the classical formulation. Nevertheless, due to the assumption of conditional independence on the effect side, but not the cause side, cause repertoires are formally distinct from effect repertoires even under unitary transformations.

The quantum formulation also provides justification for treating all variables outside the candidate system under consideration as fixed background conditions, which is motivated by IIT's intrinsicality postulate [1,16]: by the no-communication theorem [52], any unitary transformation on a system will leave the density matrix of its environment unchanged. However, not all subsets of unitary transforms are unitary. Future work should explore the implications of assuming fixed background conditions in such cases.

The IIT formalism for classical systems starts from a transition probability matrix (TPM), which corresponds to a complete set of transition probabilities (from every possible system state to every possible system state) (1). This has led some to criticize IIT on conceptual grounds, as it seems to imply that subjective experience would depend not only on the actual states a system inhabits in the course of its dynamical evolution, but also on hypothetical counterfactuals that may never happen [63]. In the QIIT formalism, the role of the classical transition probability matrix (TPM) is assumed by the unitary transform (21) applied to the quantum state. Just as evolution operators in quantum mechanics essentially are TPMs, in IIT, the TPM simply serves as a complete description of the system's dynamics.

In this work, we have focused on mechanism integrated information φ [6]. In principle, it should be possible to formally extend our QIIT formalism to incorporate the full "IIT 4.0" framework, including the system-integrated information (φ_s) [40], a full characterization of the system's cause-effect structure comprised of causal distinctions and causal relations [54], and the amount of structured information (Φ) specified by a system.

Nevertheless, there are several conceptual issues that need to be resolved before the QIIT formalism can be applied to identify conscious systems, which have to comply with all of IIT's requirements for being a substrate of consciousness (IIT's "postulates") [1]. For example, it is unclear whether mixed states should count as permissible states for evaluating the system's integrated information. While only specific sets of units, not ensembles, qualify as substrates, a particular set of units may still be in a mixed state if it is entangled with the environment (Figure 4). However, IIT's information postulate requires systems and mechanisms to have specific cause-effect power. It thus remains to be determined whether mixed states can comply with IIT's information postulate.

Recurrent quantum systems are another issue. In the classical formulation, recurrent connections between system units are required for positive system integrated information [1,40]. Physical units (e.g., neurons, transistors) are thus assumed to be dynamically persistent variables with at least two possible states. However, it is less obvious whether qubits, or qudits, more generally, may indeed be treated as variables that maintain a causal identity across their state updates.

4.4. From Micro to Macro?

Current empirical evidence suggests that consciousness and its contents are correlated with the dynamics and activity of neurons in some parts of the cerebral cortex [64]. While our experiences seem to unfold over macroscopic spatial and temporal scales, the brain can, in principle, be described at a multitude of levels, for example, as a network comprised of a few interacting brain regions, or a microphysical quantum system. Why is it then that the contents of our experiences correlate with neural activity in certain regions of the cortex rather than their underlying microphysical processes [9,21]?

IIT offers a single, general principle for identifying conscious systems: a substrate of consciousness must correspond to a set of units that forms a maximum of intrinsic cause-effect power over grains of units, updates, and states [2,9,10]. However, it remains to be determined whether IIT's propositions are compatible with our current best knowledge about microphysics [17,18].

The QIIT formalism presented above allows for a quantitative comparison between (macroscopic) classical and (microscopic) quantum systems. Squaring IIT (as well as any other causal, computational, or information-based theory of consciousness) with our current knowledge of microphysics, moreover, requires a method for obtaining macroscopic causal models from microscopic dynamics. This could be achieved by a "black-boxing" of quantum circuits into suitable macro-units [10] or a quantitative framework that formalizes the emergence of well-defined probability distributions [65].

To identify the maximally irreducible description of a system across a hierarchy of spatio-temporal scales, we have to compare micro- and macro-level descriptions of the *same* system. While it is always possible to implement the global function performed by a classical system with a quantum circuit [52], these systems will typically not have the same causal structure (the CNOT gate described in Figure 3 is exceptional in that way). One reason is that quantum gates have to be reversible, and thus require so-called "ancilla qubits" to implement convergent logic gates, such as AND-gates or NOR-gates. These ancilla qubits cannot simply be ignored in the IIT analysis, as this would introduce an observer-dependent, extrinsic perspective. They also cannot typically be treated as fixed background conditions. Understanding whether and how irreversible logic functions might emerge from reversible quantum circuits is thus an important subject for future investigations.

As is, QIIT and its classical counterpart are only partially overlapping in their domains of applicability. While QIIT is, in principle, more fundamental as an extension of IIT's classical, macroscopic causal framework to quantum systems, it is currently limited to reversible, unitary transformations, and thus cannot directly be applied to irreversible processes, commonly assumed in classical computational/cognitive systems.

Overall, we see it as a positive development that the updated IIT 4.0 formalism for computing the mechanism integrated information [6] is readily applicable within a quantum mechanical framework. Our work revealed several conceptual issues regarding theories of consciousness as they relate to fundamental physics. Regardless, the theoretical framework for identifying causes and effects of subsets of units within a quantum system should be of interest within the field of quantum information theory and quantum causal models more generally.

Author Contributions: Conceptualization, L.A., R.P. and I.D.; methodology, L.A. and I.D.; software, L.A.; validation, L.A., R.P. and I.D.; formal analysis, L.A.; investigation, L.A.; writing—original draft preparation, L.A.; writing—review and editing, R.P. and I.D.; visualization, L.A. All authors have read and agreed to the published version of the manuscript.

Funding: L.A. acknowledges the support of a grant from the Templeton World Charity Foundation (TWCF-2020-20526).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors thank the Association for Mathematical Consciousness Science (AMCS) for its institutional support. L.A. thanks William Marshall, Alireza Zaeemzadeh, and Giulio Tononi for helpful discussions and comments on earlier drafts of this article, Will Mayner for maintaining PyPhi and his advice on implementing the QIIT toolbox, and Matteo Grasso for his help with figures. R.P. acknowledges the support of the Munich Center for Mathematical Philosophy.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Oizumi, M.; Albantakis, L.; Tononi, G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Comput. Biol.* **2014**, *10*, e1003588. [CrossRef] [PubMed]
2. Tononi, G.; Boly, M.; Massimini, M.; Koch, C. Integrated information theory: From consciousness to its physical substrate. *Nat. Rev. Neurosci.* **2016**, *17*, 450–461. [CrossRef] [PubMed]
3. Albantakis, L. Integrated information theory. In *Beyond Neural Correlates of Consciousness*; Overgaard, M., Mogensen, J., Kirkeby-Hinrup, A., Eds.; Routledge: Abingdon-on-Thames, UK , 2020; pp. 87–103. [CrossRef]
4. Albantakis, L.; Barbosa, L.; Findlay, G.; Grasso, M.; Haun, A.M.; Marshall, W.; Mayner, W.G.; Zaeemzadeh, A.; Boly, M.; Juel, B.E.; Sasai, S.; Fujii, K.; David, I.; Hendren, J.; Lang, J.P.; Tononi, G. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *arXiv* **2022**. [CrossRef]
5. Albantakis, L.; Tononi, G. Causal Composition: Structural Differences among Dynamically Equivalent Systems. *Entropy* **2019**, *21*, 989. [CrossRef]
6. Barbosa, L.S.; Marshall, W.; Albantakis, L.; Tononi, G. Mechanism Integrated Information. *Entropy* **2021**, *23*, 362. [CrossRef]
7. Grasso, M.; Albantakis, L.; Lang, J.P.; Tononi, G. Causal reductionism and causal structures. *Nat. Neurosci.* **2021**, *24*, 1348–1355. [CrossRef]
8. Hoel, E.P.; Albantakis, L.; Tononi, G. Quantifying causal emergence shows that macro can beat micro. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 19790–19795. [CrossRef]
9. Hoel, E.P.; Albantakis, L.; Marshall, W.; Tononi, G. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neurosci. Conscious.* **2016**, *2016*, niw012. [CrossRef]
10. Marshall, W.; Albantakis, L.; Tononi, G. Black-boxing and cause-effect power. *PLoS Comput. Biol.* **2018**, *14*, e1006114. [CrossRef]

11. Albantakis, L.; Massari, F.; Beheler-Amass, M.; Tononi, G. A Macro Agent and Its Actions. *Synth. Libr.* **2021**, *439*, 135–155. [CrossRef]
12. Albantakis, L.; Tononi, G. The Intrinsic Cause-Effect Power of Discrete Dynamical Systems—From Elementary Cellular Automata to Adapting Animats. *Entropy* **2015**, *17*, 5472–5502. [CrossRef]
13. Mayner, W.G.; Marshall, W.; Albantakis, L.; Findlay, G.; Marchman, R.; Tononi, G. PyPhi: A toolbox for integrated information theory. *PLoS Comput. Biol.* **2018**, *14*, e1006343.
14. Gomez, J.D.; Mayner, W.G.P.; Beheler-Amass, M.; Tononi, G.; Albantakis, L. Computing Integrated Information (Φ) in Discrete Dynamical Systems with Multi-Valued Elements. *Entropy* **2021**, *23*, 6. [CrossRef] [PubMed]
15. Pearl, J. *Causality: Models, Reasoning and Inference*; Cambridge Univ Press: Cambridge, UK, 2000; Volume 29.
16. Albantakis, L.; Marshall, W.; Hoel, E.; Tononi, G. What caused what? A quantitative account of actual causation using dynamical causal networks. *Entropy* **2019**, *21*, 459.
17. Barrett, A.B.; Mediano, P.A. The phi measure of integrated information is not well-defined for general physical systems. *J. Conscious. Stud.* **2019**, *26*, 11–20.
18. Carroll, S. Consciousness and the Laws of Physics. *J. Conscious. Stud.* **2021**, *28*, 16–31. [CrossRef]
19. Brukner, **C**. Quantum causality. *Nat. Phys.* **2014**, *10*, 259–263. [CrossRef]
20. D'Ariano, G.M. Causality re-established. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2018**, *376*, 20170313. [CrossRef]
21. Tegmark, M. Consciousness as a state of matter. *Chaos Solitons Fractals* **2015**, *76*, 238–270. . [CrossRef]
22. Zanardi, P.; Tomka, M.; Venuti, L.C. Towards Quantum Integrated Information Theory. *arXiv* **2018**, arXiv:1806.01421.
23. Kleiner, J.; Tull, S. The Mathematical Structure of Integrated Information Theory. *Front. Appl. Math. Stat.* **2021**, *6*, 74. [CrossRef]
24. Barbosa, L.S.; Marshall, W.; Streipert, S.; Albantakis, L.; Tononi, G. A measure for intrinsic information. *Sci. Rep.* **2020**, *10*, 18803. [CrossRef]
25. Atmanspacher, H. Quantum Approaches to Consciousness. In *The Stanford Encyclopedia of Philosophy*, Summer 2020 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2020.
26. Prakash, C. On Invention of Structure in the World: Interfaces and Conscious Agents. *Found. Sci.* **2020**, *25*, 121–134. [CrossRef]
27. Wigner, E.P. Remarks on the mind-body question. In *Philosophical Reflections and Syntheses*; Springer: Berlin/Heidelberg, Germany, 1995; pp. 247–260.
28. Kremnizer, K.; Ranchin, A. Integrated Information-Induced Quantum Collapse. *Found. Phys.* **2015**, *45*, 889–899. [CrossRef]
29. Chalmers, D.J.; McQueen, K.J. Consciousness and the Collapse of the Wave Function. *arXiv* **2021**. [CrossRef]
30. Penrose, R. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*; Oxford University Press: Oxford, UK, 1999.
31. Hameroff, S.; Penrose, R. Consciousness in the universe: A review of the 'Orch OR' theory. *Phys. Life Rev.* **2014**, *11*, 39–78. [CrossRef]
32. Tegmark, M. Importance of quantum decoherence in brain processes. *Phys. Rev. E* **2000**, *61*, 4194–4206. [CrossRef]
33. Koch, C.; Hepp, K. Quantum mechanics in the brain. *Nature* **2006**, *440*, 611–611. [CrossRef]
34. Hagan, S.; Hameroff, S.R.; Tuszyński, J.A. Quantum computation in brain microtubules: Decoherence and biological feasibility. *Phys. Rev. E* **2002**, *65*, 061901.
35. Vaziri, A.; Plenio, M.B. Quantum coherence in ion channels: Resonances, transport and verification. *New J. Phys.* **2010**, *12*, 085001.
36. Rourk, C.J. Indication of quantum mechanical electron transport in human substantia nigra tissue from conductive atomic force microscopy analysis. *Biosystems* **2019**, *179*, 30–38. [CrossRef]
37. Atmanspacher, H.; Prentner, R. Desiderata for a Viable Account of Psychophysical Correlations. *Mind Matter* **2022**, *20*, 63–86.
38. Janzing, D.; Balduzzi, D.; Grosse-Wentrup, M.; Schölkopf, B. Quantifying causal influences. *Ann. Stat.* **2013**, *41*, 2324–2358. [CrossRef]
39. Ay, N.; Polani, D. Information Flows in Causal Networks. *Adv. Complex Syst.* **2008**, *11*, 17–41. [CrossRef]
40. Marshall, W.; Grasso, M.; Mayner, W.G.P.; Zaeemzadeh, A.; Barbosa, L.S.; Chastain, E.; Findlay, G.; Sasai, S.; Albantakis, L.; Tononi, G. System Integrated Information. *Entropy* **2023**, *25*, 334. [CrossRef] [PubMed]
41. Zaeemzadeh, A.; Tononi, G. Upper Bounds for Integrated Information. *In preparation*.
42. Krohn, S.; Ostwald, D. Computing integrated information. *Neurosci. Conscious.* **2017**, *2017*. [CrossRef]
43. Moon, K. Exclusion and Underdetermined Qualia. *Entropy* **2019**, *21*, 405. [CrossRef]
44. Vedral, V. The role of relative entropy in quantum information theory. *Rev. Mod. Phys.* **2002**, *74*, 197. [CrossRef]
45. Zhou, Y.; Zhao, Q.; Yuan, X.; Ma, X. Detecting multipartite entanglement structure with minimal resources. *Npj Quantum Inf.* **2019**, *5*, 83.
46. Gühne, O.; Tóth, G. Entanglement detection. *Phys. Rep.* **2009**, *474*, 1–75.
47. Li, J.L.; Qiao, C.F. A Necessary and Sufficient Criterion for the Separability of Quantum State. *Sci. Rep.* **2018**, *8*, 1442.
48. Skorobagatko, G.A. Universal separability criterion for arbitrary density matrices from causal properties of separable and entangled quantum states. *Sci. Rep.* **2021**, *11*, 15866. [CrossRef] [PubMed]
49. Peres, A. Separability Criterion for Density Matrices. *Phys. Rev. Lett.* **1996**, *77*, 1413.
50. Horodecki, M.; Horodecki, P.; Horodecki, R. Separability of mixed states: Necessary and sufficient conditions. *Phys. Lett. A* **1996**, *223*, 1–8.

51. Bennett, C.H.; Di Vincenzo, D.P.; Mor, T.; Shor, P.W.; Smolin, J.A.; Terhal, B.M. Unextendible Product Bases and Bound Entanglement. *Phys. Rev. Lett.* **1999**, *82*, 5385.
52. Nielsen, M.A.; Chuang, I.L. *Quantum Computation and Quantum Information: 10th Anniversary Edition*, 10th ed.; Cambridge University Press: New York, NY, USA, 2011.
53. Dur, W.; Vidal, G.; Cirac, J.I. Three qubits can be entangled in two inequivalent ways. *Phys. Rev. A* **2000**, *62*, 062314.
54. Haun, A.; Tononi, G. Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy* **2019**, *21*, 1160. [CrossRef]
55. Mermin, N.D. *Quantum Computer Science. An Introduction*; Cambridge University Press: Cambridge, UK, 2007.
56. Schumacher, B.; Westmoreland, M.D. Isolation and Information Flow in Quantum Dynamics. *Found. Phys.* **2012**, *42*, 926–931. [CrossRef]
57. Greenberger, D.M.; Horne, M.A.; Zeilinger, A. Going Beyond Bell's Theorem. In *Bell's Theorem, Quantum Theory and Conceptions of the Universe*; Springer: Dordrecht, The Netherlands, 1989; pp. 69–72. [CrossRef]
58. Acín, A.; Bruß, D.; Lewenstein, M.; Sanpera, A. Classification of Mixed Three-Qubit States. *Phys. Rev. Lett.* **2001**, *87*, 040401.
59. Arkhipov, A. Non-Separability of Physical Systems as a Foundation of Consciousness. *Entropy* **2022**, *24*, 1539. [CrossRef]
60. Schwindt, J.M. Nothing happens in the Universe of the Everett Interpretation. *arXiv* **2012**. [CrossRef]
61. Carroll, S.M.; Singh, A. Quantum mereology: Factorizing Hilbert space into subsystems with quasiclassical dynamics. *Phys. Rev. A* **2021**, *103*, 022213.
62. Everett, H, III. " Relative state" formulation of quantum mechanics. *Rev. Mod. Phys.* **1957**, *29*, 454. [CrossRef]
63. Seth, A. *Being You: A New Science of Consciousness*; Penguin: London, UK, 2021.
64. Koch, C.; Massimini, M.; Boly, M.; Tononi, G. Neural correlates of consciousness: Progress and problems. *Nat. Rev. Neurosci.* **2016**, *17*, 307–321. [CrossRef]
65. Durham, I. A Formal Model for Adaptive Free Choice in Complex Systems. *Entropy* **2020**, *22*, 568. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.