

Mechanism Integrated Information

Leonardo S. Barbosa ¹, William Marshall ^{1,2}, Larissa Albantakis ¹  and Giulio Tononi ^{1,*}

¹ Department of Psychiatry, University of Wisconsin-Madison, Madison, WI 53719, USA; leonardo.barbosa@wisc.edu (L.S.B.); wmarshall3@wisc.edu (W.M.); albantakis@wisc.edu (L.A.)

² Department of Mathematics and Statistics, Brock University, St. Catharines, ON L2S 3A1, Canada

* Correspondence: gtononi@wisc.edu

Abstract: The Integrated Information Theory (IIT) of consciousness starts from essential phenomenological properties, which are then translated into postulates that any physical system must satisfy in order to specify the physical substrate of consciousness. We recently introduced an information measure (Barbosa et al., 2020) that captures three postulates of IIT—existence, intrinsicity and information—and is unique. Here we show that the new measure also satisfies the remaining postulates of IIT—integration and exclusion—and create the framework that identifies maximally irreducible mechanisms. These mechanisms can then form maximally irreducible systems, which in turn will specify the physical substrate of conscious experience.

Keywords: causation; consciousness; intrinsic; existence



Citation: Barbosa, L.S.; Marshall, W.; Albantakis, L.; Tononi, G. Mechanism Integrated Information. *Entropy* **2021**, *23*, 362. <https://doi.org/10.3390/e23030362>

Academic Editors: Raúl Alcaraz and Kyumin Moon

Received: 12 January 2021

Accepted: 12 March 2021

Published: 18 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Integrated information theory (IIT; [1–3]) identifies the essential properties of consciousness and postulates that a physical system accounting for it—the physical substrate of consciousness (PSC)—must exhibit these same properties in physical terms. Briefly, IIT starts from the existence of one's own consciousness, which is immediate and indubitable. The theory then identifies five essential phenomenal properties that are immediate, indubitable and true of every conceivable experience, namely intrinsicity, composition, information, integration and exclusion. These phenomenal properties, called axioms, are translated into essential physical properties of the PSC, called postulates. The postulates are conceptualized in terms of cause–effect power and given a mathematical formulation in order to make testable predictions and allow for inferences and explanations.

So far, the mathematical formulation employed well-established measures of information, such as Kullback–Leibler divergence (KLD) [4] or earth mover's distance (EMD) [3]. Ultimately, however, IIT requires a measure that is based on the postulates of the theory and is unique, because the quantity and quality of consciousness are what they are and cannot vary with the measure chosen. Recently, we introduced an information measure, called intrinsic difference [5], which captures three postulates of IIT—existence, intrinsicity and information—and is unique. Our primary goal here is to explore the remaining postulates of IIT—composition, integration and exclusion—in light of this unique measure, focusing on the assessment of integrated information φ for the mechanisms of a system. In doing so, we will also revisit the way of performing partitions.

The plan of the paper is as follows. In Section 2, we briefly introduce the axioms and postulates of IIT; in Section 3, we introduce the mathematical framework for measuring φ based on intrinsic difference (ID), which satisfies the postulates of IIT and is unique; in Section 4, we explore the behavior of the measure in several examples; and in Section 5, we discuss the connection between the new framework, previous versions of IIT and future developments.

2. Axioms and Postulates

This section summarizes the axioms of IIT and the corresponding postulates. For a complete description of the axioms and their motivation, the reader should consult [2,3,6].

Briefly, the zeroth axiom, existence, says that experience exists, immediately and indubitably. The zeroth postulate requires that the PSC must exist in physical terms. The PSC is assumed to be a system of interconnected units, such as a network of neurons. Physical existence is taken to mean that the units of the system must be able to be causally affected by or causally affect other units (take and make a difference). To demonstrate that a unit has a potential cause, one can observe whether the unit's state can be caused by manipulating its input, while to demonstrate that a unit has a potential effect one can manipulate the state of the unit and observe if it causes the state of some other unit [7].

The first axiom, intrinsicality, says that experience is subjective, existing from the intrinsic perspective of the subject of experience. The corresponding postulate requires that a PSC has potential causes and effects within itself.

The second axiom, composition, says that experience is structured, being composed of phenomenal distinctions bound by phenomenal relations. The corresponding postulate requires that a PSC, too, must be structured, being composed by causal distinctions specified by subsets of units (mechanisms) over subsets of units (cause and effect purviews) and by causal relations that bind together causes and effects overlapping over the same units. The purviews are then subset of units whose states are constrained by another subset of units, the mechanisms, in its particular state. The set of all causal distinctions and relations within a system compose its cause–effect structure.

The third axiom, information, says that experience is specific, being the particular way it is, rather than generic. The corresponding postulate states that a PSC must specify a cause–effect structure composed of distinctions and relations that specify particular cause and effect states.

The fourth axiom, integration, says that experience is unified, in that it cannot be subdivided into parts that are experienced separately. The corresponding postulate states that a PSC must specify a cause–effect structure that is unified, being irreducible to the cause–effect structures specified by causally independent subsystems. Integrated information (Φ) is a measure of the irreducibility of the cause–effect structure specified by a system [8]. The degree Φ to which a system is irreducible can be interpreted as a measure of its existence. Mechanism integrated information (φ) is an analogous measure that quantifies the existence of a mechanism within a system. Only mechanisms that exist within a system ($\varphi > 0$) contribute to its cause–effect structure.

Finally, the exclusion axiom says that experience is definite, in that it contains what it contains, neither less nor more. The corresponding postulate states that the cause–effect structure specified by a PSC should be definite: it must specify a definite set of distinctions and relations over a definite set of units, neither less nor more. The PSC and associated cause–effect structure is given by the set of units for which the value of Φ is maximal, and its distinctions and relations corresponding to maxima of φ . According to IIT, then, a system is a PSC if it is a maximum of integrated information, meaning that it has higher integrated information than any overlapping systems [3,9]. Moreover, the cause–effect structure specified by the PSC is identical to the subjective quality of the experience [10].

3. Theory

We first describe the process for measuring the integrated information (φ) of a mechanism based on the postulates of IIT. In order to contribute to experience, a mechanism must satisfy the postulates described in Section 2 (note that mechanisms cannot be compositional because, as components of the cause–effect structure, they cannot have components themselves). We then present some theoretical developments related to partitioning a mechanism in order to assess integration and to measuring the difference between probability distributions for quantifying intrinsic information. The subsequent process of measuring the integrated information of the system (Φ) will be discussed elsewhere.

3.1. Mechanism Integrated Information

Our starting point is a stochastic system $S = \{S_1, S_2, \dots, S_n\}$ with state space Ω_S and current state $s_t \in \Omega_S$ (Figure 1a). The system is constituted of n random variables that represent the units of a physical system and has a transition probability function

$$p(s_{t+1} | s_t) = \mathcal{P}(S_{t+1} = s_{t+1} | S_t = s_t), \quad s_t, s_{t+1} \in \Omega_S, \quad (1)$$

which describes how the system updates its state (see Appendix A.1 for details). The goal is to define the integrated information of a mechanism $M \subseteq S$ in a state $m_t \in \Omega_M$ based on the postulates of IIT. To this end, we will develop a difference measure $\varphi(m_t, Z_{t\pm 1}, \psi)$ which quantifies how much a mechanism M in state m_t constrains the state of a purview, a set of units $Z_{t\pm 1} \subseteq S$, compared to a partition

$$\psi = \{(M_1, Z_1), (M_2, Z_2), \dots, (M_k, Z_k)\}, \quad (2)$$

of the mechanism and purview into k independent parts (Figure 1b). As we evaluate the IIT postulates step by step, we will provide mathematical definitions for the required quantities, introduce constraints on φ and eventually arrive at a unique measure. Since potential causes of $M = m_t$ are always inputs to M , and potential effects of $M = m_t$ are always outputs of M , we will omit the corresponding update indices ($t - 1, t, t + 1$) unless necessary.

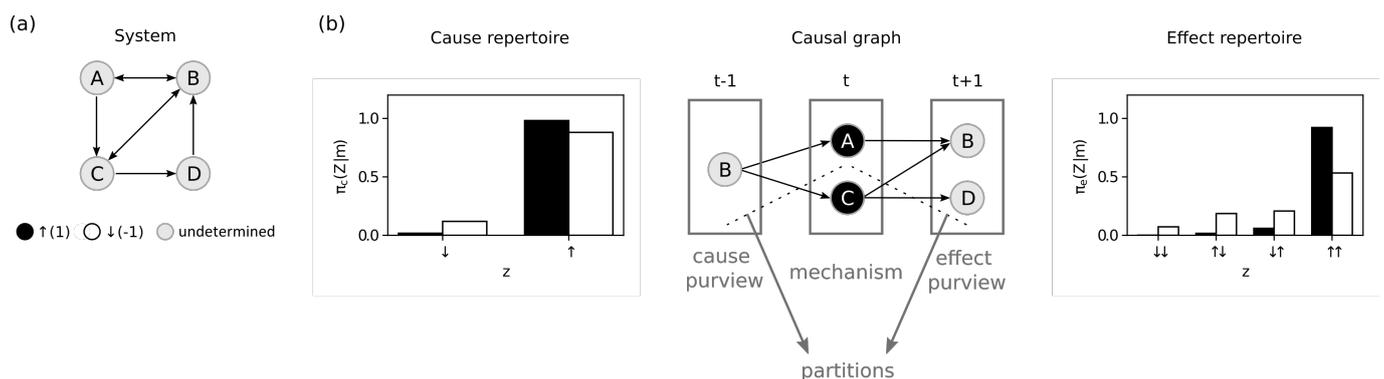


Figure 1. Theory. (a) System S with four random variables. (b) Example of a mechanism $M = \{A, C\}$ in state $m = \{\uparrow, \uparrow\}$ constraining a cause purview $Z = \{B\}$ and an effect purview $Z = \{B, D\}$. Dashed lines show the partitions. The bar plots show the probability distributions, that is the cause repertoire (left) and effect repertoire (right). The black bars show the probabilities when the mechanism is constraining the purview, and the white bars show the probabilities after partitioning the mechanism.

3.1.1. Existence

For a mechanism to exist in a physical sense, it must be possible for something to change its state, and it must be able to change the state of something (it has potential causes and effects). To evaluate these potential causes and effects, we define the cause repertoire $\pi_c(Z | m)$ (see Equation (A2)) and the effect repertoire $\pi_e(Z | m)$ (see Equation (A1)), which describe how m constrains the potential input or output states of $Z \subseteq S$ respectively (Figure 1b) [3,11–13].

The cause and effect repertoires are probability distributions derived from the system’s transition probability function (Equation (1)) by conditioning on the state of the mechanism and *causally marginalizing* the variables outside the purview ($S \setminus Z$). Causal marginalization is also used to remove any contributions to the repertoire from units outside the mechanism ($S \setminus M$). In this way, we capture the constraints due to the mechanism in its state and nothing else. Note that the cause and effect repertoires generally differ from the corresponding conditional probability distributions.

Having introduced cause and effect repertoires, we can write the difference

$$\varphi_e(m, Z, \psi) = D(\pi_e(Z | m), \pi_e^\psi(Z | m)),$$

where $\pi_e^\psi(Z | m)$ corresponds to the partitioned effect repertoire (see Equation (A3)) in which certain connections from M to Z are severed (causally marginalized). When there is no change after the partition, we require that

$$\varphi_e(m, Z, \psi) = 0.$$

The same analysis holds for causes, replacing π_e with π_c in the definition of $\varphi_c(m, Z, \psi)$. Unless otherwise specified, in what follows we focus on effects.

3.1.2. Intrinsicity

The intrinsicity postulate states that, from the intrinsic perspective of the mechanism $M = m$ over a purview Z , the effect repertoire $\pi_e(Z|m)$ is set and has to be taken *as is*. This means that, given the purview units and their connections to the mechanism, the constraints due to the mechanism are defined by how all its units at a particular state m at t constrain all units in the effect purview at $t + 1$ and cause purview at $t - 1$. For example, if the mechanism fully constrains all of its purview units except for one unit which remains fully unconstrained, the mechanism cannot just ignore the unconstrained unit or optimize its overall constraints by giving more weight to some states than others in the effect repertoire. For this reason, the intrinsicity postulate should make the difference measure D between the partitioned and unpartitioned repertoire sensitive to a tradeoff between “expansion” and “dilution”: the measure should increase if the purview includes more units that are highly constrained by the mechanism but decrease if the purview includes units that are weakly constrained. The mathematical formulation of this requirement is given in Section 3.3.

3.1.3. Information

The information postulate states that a mechanism M , by being in its particular state m , must have a specific effect, which means that it must specify a particular effect state z over the purview Z . The effect state should be the one for which m makes the most difference. To that end, we require a difference measure of the form

$$\varphi_e(m, Z, \psi) = D(\pi_e(Z | m), \pi_e^\psi(Z | m)) = \max_{z \in \Omega_Z} \left| f\left(\pi_e(z | m), \pi_e^\psi(z | m)\right) \right|,$$

such that the difference D between effect repertoires is evaluated as the maximum of the absolute value of some function f that is assessed for particular states. The function f is one of the main developments of the current work and is discussed in Section 3.3.

3.1.4. Integration

The integration postulate states that a mechanism must be unitary, being irreducible to independent parts. By comparing the effect repertoire $\pi_e(Z | m)$ against the partitioned repertoire $\pi_e^\psi(Z | m)$, we can assess how much of a difference the partition ψ makes to the effect of m . To quantify how irreducible m 's effect is on Z , one must compare all possible partitioned repertoires to the unpartitioned effect repertoire. In other words, one must evaluate each possible partition ψ . Of all partitions, we define the minimum information partition (MIP)

$$\psi^* = \operatorname{argmin}_{\psi} \varphi_e(m, Z, \psi),$$

which is the one that makes the least difference to the effect. The intrinsic integrated effect information (or integrated effect information for short) of the mechanism M in state m about a purview Z is then defined as

$$\varphi_e(m, Z) = \varphi_e(m, Z, \psi^*).$$

If $\varphi_e(m, Z) = 0$, there is a partition of the candidate mechanism that does not make a difference, which means that the candidate mechanism is reducible.

3.1.5. Exclusion

The exclusion postulate states that a mechanism must be definite, it must specify a definite effect over a definite set of units. That is, a mechanism must be about a maximally irreducible purview

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

which maximizes integrated effect information and is in the effect state

$$z_e^* = \operatorname{argmax}_{z \in \Omega_{Z_e^*}} \left| f \left(\pi_e(z | m), \pi_e^{\psi^*}(z | m) \right) \right|.$$

The purview Z_e^* is then used to define the integrated effect information of the mechanism M

$$\varphi_e(m) = \varphi_e(m, Z_e^*).$$

Returning to the existence postulate, a mechanism must have both a cause and an effect. By an analogous process using cause repertoires π_c instead of effect repertoires π_e , we can define the integrated cause information of m

$$\varphi_c(m) = \varphi_c(m, Z_c^*),$$

and the integrated information of the mechanism

$$\varphi(m) = \min \{ \varphi_c(m), \varphi_e(m) \}. \tag{3}$$

Thus, if a candidate mechanism M in state m is reducible over every purview either on the cause *or* effect side, $\varphi(m) = 0$ and M does not contribute to experience. Otherwise, $M = m$ is irreducible and forms a mechanism within the system. As such, it specifies a *distinction*

$$X(m) = \{ (Z_c^* = z_c^*, Z_e^* = z_e^*, \varphi(m)) : Z_c^*, Z_e^* \subseteq S, z_c^* \in \Omega_{Z_c^*}, z_e^* \in \Omega_{Z_e^*} \},$$

which links its maximally irreducible cause with its maximally irreducible effect, for $M \subseteq S$, $m \in \Omega_M$ and $\varphi(m) \in \{x \in \mathbb{R} : x > 0\}$. While a mechanism always specifies a unique $\varphi(m)$ value, due to symmetries in the system it is possible that there are multiple equivalent solutions for $Z_c^* = z_c^*$ or $Z_e^* = z_e^*$. We expect such “ties” to be exceedingly rare in physical systems with variable connection strengths, as well as a certain amount of indeterminism and outline possible solutions to resolves “ties” in the discussion, Section 5.

3.2. Disintegrating Partitions

According to the integration postulate, a mechanism can only exist from the intrinsic perspective of a system if it is irreducible, meaning that any partition of the mechanism would make a difference to its potential cause or effect. Accordingly, computing the integrated information of a mechanism requires partitioning the mechanism and assessing the difference between partitioned and unpartitioned repertoires. In this section we give additional mathematical details and theoretical considerations for how to partition a mechanism together with its purview Z .

Generally, a partition ψ of a mechanism M and a purview Z is a set of parts as defined in Equation (2), with some restrictions on (M_i, Z_i) . The partition “cuts apart” the mechanism, severing any connections from M_i to Z_j ($i \neq j$). We use causal marginalization (see Appendix A) to remove any causal power M_i has over Z_j ($i \neq j$) and compute a partitioned repertoire. Practically, it is as though we do not condition on the state of M_i

when consider Z_j . Before describing the restrictions on (M_i, Z_i) we will look at a few examples to highlight the conceptual issues. First, consider a third-order mechanism $M = \{A, B, C\}$ with the same units (as inputs or outputs) in the corresponding third order purview $Z = \{A, B, C\}$. A standard example of a partition of this mechanism is

$$\psi^1 = \{(\{A, B\}, \{A, B\}), (\{C\}, \{C\})\},$$

which cuts units $\{A, B\}$ away from unit $\{C\}$. Now consider the situation where we would like to additionally cut $\{B\}$ in the purview away from $\{A, B\}$ in the mechanism. This partition can be represented as

$$\psi^2 = \{(\{A, B\}, \{A\}), (\{\emptyset\}, \{B\}), (\{C\}, \{C\})\}.$$

This example raises the issue of whether to allow the empty set as part of a partition. The question is not only conceptual but also practical, in a situation where $\{A, B\}$ and $\{C\}$ have opposite effects (e.g., excitatory and inhibitory connections), then it may be that the MIP $\psi^* = \psi^2$ (see Section 4.2 for an example). Here, the mechanism is always partitioned together with a purview subset.

While the definition of ψ should include partitions such as ψ^2 above, this raises additional issues. Consider the partition

$$\psi^3 = \{(\{A, B, C\}, \{A, B\}), (\{\emptyset\}, \{C\})\}.$$

In ψ^3 , the set of all mechanism units is contained in one part. Should such a partition count as “cutting apart” the mechanism? The same problem arises for partitions of first-order mechanisms. Consider, for example, $M = \{A\}$ with purview $Z = \{A, B, C\}$ and partition

$$\psi^4 = \{(\{A\}, \{A, B\}), (\{\emptyset\}, \{C\})\}.$$

A first-order mechanism should be considered completely irreducible by definition, yet for the proposed partition only a small fraction of its constraint is considered integrated information: while $M = A$ may constrain A, B , and C , only its constraints over C would be evaluated by ψ^4 . A similar argument applies to ψ^3 , which would only allow us to evaluate the constraint of the mechanism $M = \{A, B, C\}$ on C , not the entire purview $Z = \{A, B, C\}$. In sum, ψ^3 and ψ^4 should not be permissible partitions by the integration postulate. The set of mechanism units may not remain integrated over a purview subset once a partition is applied.

Based on the above argument, we propose a set of *disintegrating partitions*

$$\Psi(M, Z) = \left\{ \{(M_i, Z_i)\}_{i=1}^k \mid k \in \{2, 3, 4, \dots\}, M_i \in \mathbb{P}(M), Z_i \in \mathbb{P}(Z), \bigcup M_i = M, \bigcup Z_i = Z, Z_i \cap Z_j = M_i \cap M_j = \emptyset \text{ for all } i \neq j, M_i = M \implies Z_i = \emptyset \right\}, \quad (4)$$

such that for each $\psi \in \Psi(M, Z)$: $\{M_i\}$ is a partition of M and $\{Z_i\}$ is a partition of Z but allows the empty set to be used as a part. Moreover, if the mechanism is not partitioned into at least two parts, then the mechanism must be cut away from the entire purview.

In summary, the above definition of possible partitions ensures that the mechanism set must be divided into at least two parts, except for the special case where one part contains the whole mechanism but no units in the purview (complete partition, ψ^0). This special partition can be interpreted as “destroying” the whole mechanism at once and observing the impact its absence has on the purview.

3.3. Intrinsic Difference (ID)

In this section we define the measure D , which quantifies the difference between the unpartitioned and partitioned repertoires specified by a mechanism and thus plays an important role in measuring integrated information. We propose a set of properties that D should satisfy based on the postulates of IIT described above, and then identify the unique measure that satisfies them.

Our desired properties are described in terms of discrete probability distributions $P^n = [p_1, p_2, \dots, p_n]$ and $Q^n = [q_1, q_2, \dots, q_n]$. Generally, P^n represents the cause or effect repertoire of a mechanism $\pi(Z|m)$, while Q^n represents the partitioned repertoire $\pi^\psi(Z|m)$.

The first property, *causality*, captures the requirement for physical existence (Section 3.1.1) that a mechanism has a potential cause and effect,

$$D(P^n, Q^n) = 0 \iff P^n \equiv Q^n. \quad (5)$$

The interpretation is that the integrated information m specifies about Z is only zero if the unpartitioned and partitioned repertoires are identical. In other words, by being in state m , the mechanism M does not constrain the potential state of Z above its partition into independent parts.

The second property, *intrinsicity*, captures the requirement that physical existence must be assessed from the perspective of the mechanism itself (Section 3.1.2). The idea is that information should be measured from the intrinsic perspective of the candidate mechanism M in state m , which determines the potential state of the purview Z by itself, independent of external observers. In other words, the constraint m has over Z must depend only on their units and connections. In contrast, traditional information measures were conceived to quantify the amount of signal transmitted across a channel between a sender and a receiver from an extrinsic perspective, typically that of a channel designer who has the ability to optimize the channel's capacity. This can be done by adjusting the mapping between the states of M and Z through encoders and decoders to reduce indeterminism in the signal transmission. However, such a remapping would require more than just the units and connections present in M and Z , thus violating intrinsicity [5].

The intrinsicity property is defined based on the behavior of the difference measure when distributions are extended by adding units to the purview or increasing the number of possible states of a unit [14]. A distribution P_1^n is extended by a distribution P_2^n to create a new distribution $P_1^n \otimes P_2^n$, where \otimes is the Kronecker product. When a fully selective distribution (one where an outcome occurs with probability one) is extended by another fully selective distribution, the measure should increase additively (expansion). However, if a distribution is extended by a fully undetermined distribution (one where all n outcomes are equally likely), then the measure should decrease by a factor of n (dilution). For expansion, suppose P_1^n and P_2^n are fully selective distributions, then for any Q_1^n and Q_2^n we have

$$D(P_1^n \otimes P_2^n, Q_1^n \otimes Q_2^n) = D(P_1^n, Q_1^n) + D(P_2^n, Q_2^n). \quad (6)$$

For dilution, suppose P_2^n and Q_2^n are fully undetermined distributions, then for any P_1^n, Q_1^n we have

$$D(P_1^n \otimes P_2^n, Q_1^n \otimes Q_2^n) = \frac{1}{n} D(P_1^n, Q_1^n). \quad (7)$$

Together, Equations (6) and (7) define the intrinsicity property.

The final property, *specificity*, requires that physical existence must be about a specific purview state (Section 3.1.3),

$$D(P^n, Q^n) = \max_{\alpha} |f(p_{\alpha}, q_{\alpha})|. \quad (8)$$

The function $f(p, q)$ defines the difference between two probability distributions at a specific state of the purview. The mechanism is defined based on the state that maximizes its difference within the system.

Previous work employed similar properties to quantify intrinsic information but used a version of the specificity property that did not include the absolute value [5]. In that work, the goal was to compute the intrinsic information of a communication channel, with an implicit assumption that the source is sending a specific message. In that context, a signal is only informative if it increases the probability of receiving the correct message. Here we are interested in integrated information within the context of the postulates of IIT as a means to quantify existence, which requires causes and effects. A mechanism can be seen as having an effect (or cause) whether it increases or decreases the probability of a specific state.

Together, the three properties (causality, specificity, and intrinsicity) characterize a unique measure, the intrinsic difference, for measuring the integrated information of a mechanism. Note that while causality (Equation (5)) and expansion (Equation (6)) properties are traditionally required by information measures (see [15]), here we also require dilution (Equation (7)) and specificity (Equation (8)). While the maximum operation present in specificity in order to select one specific purview state seems to us uncontroversial, one may argue that the dilution factor $\frac{1}{n}$ in Equation (7) is somewhat arbitrary. However, note that if specificity requires that information is specific to one state, after adding a fully undetermined distribution of size n to the purview, the amount of causal power measured by the function f in state α will be invariably divided by n . This way, we believe that the dilution factor must be necessarily $\frac{1}{n}$, at least in this particular case.

Theorem 1. *If $D(P^n, Q^n)$ satisfies the causality, intrinsicity, and specificity properties, then*

$$D(P^n, Q^n) = \max_{\alpha} |f(p_{\alpha}, q_{\alpha})|,$$

where

$$f(p, q) = k p \log\left(\frac{p}{q}\right).$$

The full mathematical statement of the theorem and its proof are presented in Appendix B. For the rest of the manuscript we assume $k = 1$ without loss of generality. Here, our main interest is using ID to quantify the difference between unpartitioned and partitioned cause or effect repertoires when assessing the integrated information of a mechanism,

$$\varphi(m, Z) = D\left(\pi(Z | m), \pi^{\psi^*}(Z | m)\right) = \max_{z \in \Omega_Z} \left| \pi(z | m) \log\left(\frac{\pi(z | m)}{\pi^{\psi^*}(z | m)}\right) \right|.$$

One can interpret the integrated information as being composed of two terms. First, the informativeness

$$\left| \log\left(\frac{\pi(z | m)}{\pi^{\psi^*}(z | m)}\right) \right|,$$

which reflects the difference in Hartley information contained in state z before and after the partition. Second, the selectivity

$$\pi(z | m),$$

which reflects the likelihood of the cause or effect. Together, the two terms can be interpreted as the density of information for a particular state [5].

4. Methods and Results

Throughout this section we investigate each step necessary to compute $\varphi(m)$, the integrated information of a mechanism M in state m . To this end, we construct systems S formed by units A, B, C, \dots that are either \uparrow (1) or \downarrow (−1) at time t with probability of being \uparrow defined by (Figure 2a)

$$\mathcal{P}(Y_t = 1 \mid A_{t-1} = a_{t-1}, B_{t-1} = b_{t-1}, \dots) = \frac{1}{1 + \exp\left\{-\frac{2(a_{t-1} + b_{t-1} + \dots + h)}{\tau}\right\}}, \quad (9)$$

for all $Y \in S$, where A, B, \dots are the units that input to Y . Besides the sum of the input states, the function depends on two parameters: $h \in \mathbb{R}$ defines a bias towards being \uparrow ($h > 0$) or \downarrow ($h < 0$), while $\tau \in \{x \in \mathbb{R} : x \geq 0\}$ defines how deterministic unit A is. For $\tau \rightarrow \infty$, the unit turns \uparrow or \downarrow with equal probability (fully undetermined), while for $\tau = 0$ it turns \uparrow whenever the sum of the inputs is greater than the threshold η , and turns \downarrow otherwise (fully selective; Figure 2a). This way, $\tau = 0$ means that the unit is fully constrained by the inputs (deterministic), $\tau = 1$ means the unit is partially constrained, and $\tau = 10$ means the unit is only weakly constrained, etc. Unless otherwise specified, in the following we focus on investigating effect purviews.

4.1. Intrinsic Information

We start by investigating the role of intrinsicity in computing the integrated information of a mechanism. To this end, we will compare $\varphi_e(m, Z, \psi^0)$ for various mechanism-purview pairs, which evaluates the ID over a complete partition

$$\psi^0 = \{(\{M\}, \{\emptyset\}), (\{\emptyset\}, \{Z\})\}$$

of mechanism M and purview units Z , leaving the purview fully unconstrained after the partition (in this case, the partitioned repertoires are equivalent to the unconstrained repertoires defined in Equation (A5) and Equation (A4)). Intrinsicity requires that the ID must increase additively when fully constrained units are added to the purview (expansion, Equation (6)) and decrease exponentially when fully unconstrained units are added to the purview (dilution, Equation (7)). We define the system S depicted in Figure 2b to investigate the expansion and dilution of a mechanism $M = \{A\}$ over different purviews $Z \subseteq S$. Next, we fix the mechanism M in state $m = 1$ and measure the ID of this mechanism over effect purviews with varying levels of indeterminism τ but a fixed threshold $h = 0$ (partially deterministic majority gates).

First consider the purview $Z = \{B\}$ with a fully constrained unit ($\tau_B = 0$), such that (Figure 2B)

$$\varphi_e(m, Z, \psi^0) = \text{ID}(\pi_e(B \mid A = \uparrow), \pi_e^{\psi^0}(B \mid A = \uparrow)) = 0.69.$$

Now consider the same mechanism over a larger purview $Z = \{B, C\}$, which has an additional, partially constrained unit C ($\tau_C = 1$). This purview has a larger repertoire of possible states, resulting in a larger difference between partitioned and unpartitioned probabilities of one state (high informativeness). At the same time, the probability of this state is still very high in absolute terms (high selectivity). Thus, the ID of m over $\{B, C\}$ is higher than over $\{B\}$ alone (Figure 2c):

$$\varphi_e(m, Z, \psi^0) = \text{ID}(\pi_e(BC \mid A = \uparrow), \pi_e^{\psi^0}(BC \mid A = \uparrow)) = 1.11.$$

The higher value for $Z = \{B, C\}$ reflects the *expansion* that occurs whenever informativeness increases while selectivity is still high. Notice that the expansion here is subadditive since the new unit is constrained but not *fully* constrained (or fully selective).

Finally, consider another purview $Z = \{B, D\}$, where D is only weakly constrained ($\tau_D = 10$). While the new purview has a state where informativeness is marginally higher than before, selectivity is much lower (the state has much lower probability). For this reason, $\varphi_e(m, Z, \psi^0)$ is lower for $Z = \{B, D\}$ than for the smaller purview $Z = \{B\}$, reflecting *dilution* (Figure 2c):

$$\varphi_e(m, Z, \psi^0) = \text{ID}(\pi_e(BD \mid A = \uparrow), \pi_e^{\psi^0}(BD \mid A = \uparrow)) = 0.43.$$

Notice that dilution here is not exactly a factor of 2 since the new unit is weakly constrained by the mechanism but not *fully* unconstrained.

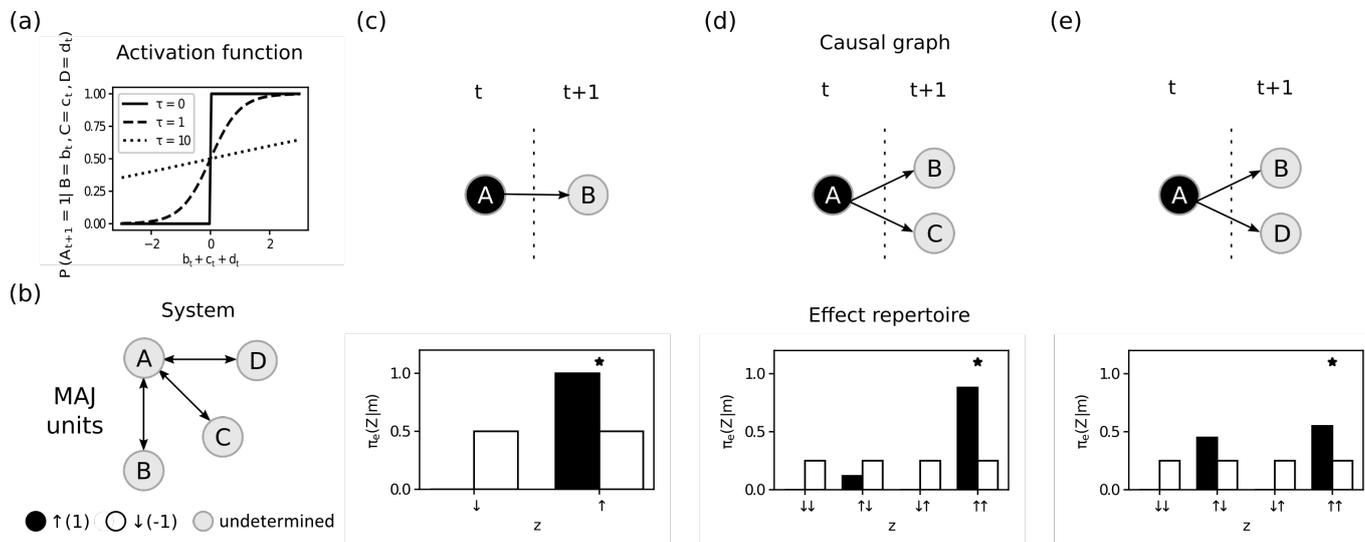


Figure 2. Intrinsicity. (a) Activation functions without bias ($\eta = 0$) and different levels of constraint ($\tau = 0$, $\tau = 1$ and $\tau = 10$). (b) System S analyzed in this figure. The remaining panels show on top the causal graph of the mechanism $M = \{A\}$ at state $m = \{1\}$ constraining different output purviews and on the bottom the probability distributions of the purviews (effect repertoires). The black bars show the probabilities when the mechanism is constraining the purview, and the white bars show the unconstrained probabilities after the complete partition ψ^0 . The “*” indicates the state selected by the maximum operation in the intrinsic difference (ID) function. (c) The mechanism fully constrains the unit B in the purview $Z = \{B\}$ ($\tau_B = 0$), resulting in state $z = \{\uparrow\}$ defining the amount of intrinsic information in the mechanism as $\varphi(m, Z, \psi^0) = ID(\pi_e(B|M = \uparrow) | \pi_e^{\psi^0}(B|M = \uparrow)) = \pi_e(B = \uparrow | A = \uparrow) \cdot \log(\pi_e(B = \uparrow | A = \uparrow) / \pi_e^{\psi^0}(B = \uparrow | M = \uparrow)) = 1 \cdot 0.69 = 0.69$. (d) After adding a slightly undetermined unit ($\tau_C = 1$) to the purview ($Z = \{B, C\}$), the intrinsic information increases to 1.11. The new maximum state ($z = \{\uparrow, \uparrow\}$) has now much higher informativeness ($|\log(\pi_e(BC = \uparrow\uparrow | A = \uparrow) / \pi_e^{\psi^0}(BC = \uparrow\uparrow | A = \uparrow))| = 1.26$) but only slightly lower selectivity ($\pi(BC = \uparrow\uparrow | A = \uparrow) = 0.89$), resulting in expansion. (e) When instead of C , we add the very undetermined unit D to the purview ($\tau_D = 10$), the new purview ($Z = \{B, D\}$) has a new maximum state ($z = \{\uparrow, \uparrow\}$) with marginally higher informativeness ($|\log(\pi_e(BC = \uparrow\uparrow | A = \uparrow) / \pi_e^{\psi^0}(BC = \uparrow\uparrow | A = \uparrow))| = 0.79$) and very low selectivity ($\pi_e(BC = \uparrow\uparrow | A = \uparrow) = 0.55$), resulting in dilution.

Next we investigate the role of the information postulate, which requires that the mechanism must be *specific*, meaning that a mechanism must both be in a specific state and specify an effect state (or a cause state) of a specific purview. Consider the system in Figure 3a where we focus on a high-order mechanism with four units $M = \{A, B, C, D\}$ over a purview with three units $Z = \{A, B, C\}$. The threshold and amount of indeterminism of the purview units are fixed: $h = -3$ and $\tau = 1$, which makes the purview units function like partially deterministic AND gates. We show not only that the mechanism can be more or less informative depending on its state but also that the specific purview state selected by the ID measure depends both on the probability of the state and on how much the state is constrained by the mechanism.

When the state of the mechanism is $m = \{\downarrow, \downarrow, \downarrow, \downarrow\}$ (Figure 3b), the most informative state in the purview is $z = \{\downarrow, \downarrow, \downarrow\}$ since all units are *more* likely to be turned \downarrow than they are after partitioning (high informativeness), and at the same time this state still has high probability (high selectivity). Out of all states, $z = \{\downarrow, \downarrow, \downarrow\}$ maximizes informativeness and selectivity in combination, resulting in

$$\begin{aligned} \varphi_e(m, Z, \psi^0) &= ID(\pi_e(ABC | ABCD = \downarrow\downarrow\downarrow), \pi_e^{\psi^0}(ABC | ABCD = \downarrow\downarrow\downarrow)) \\ &= 0.27. \end{aligned}$$

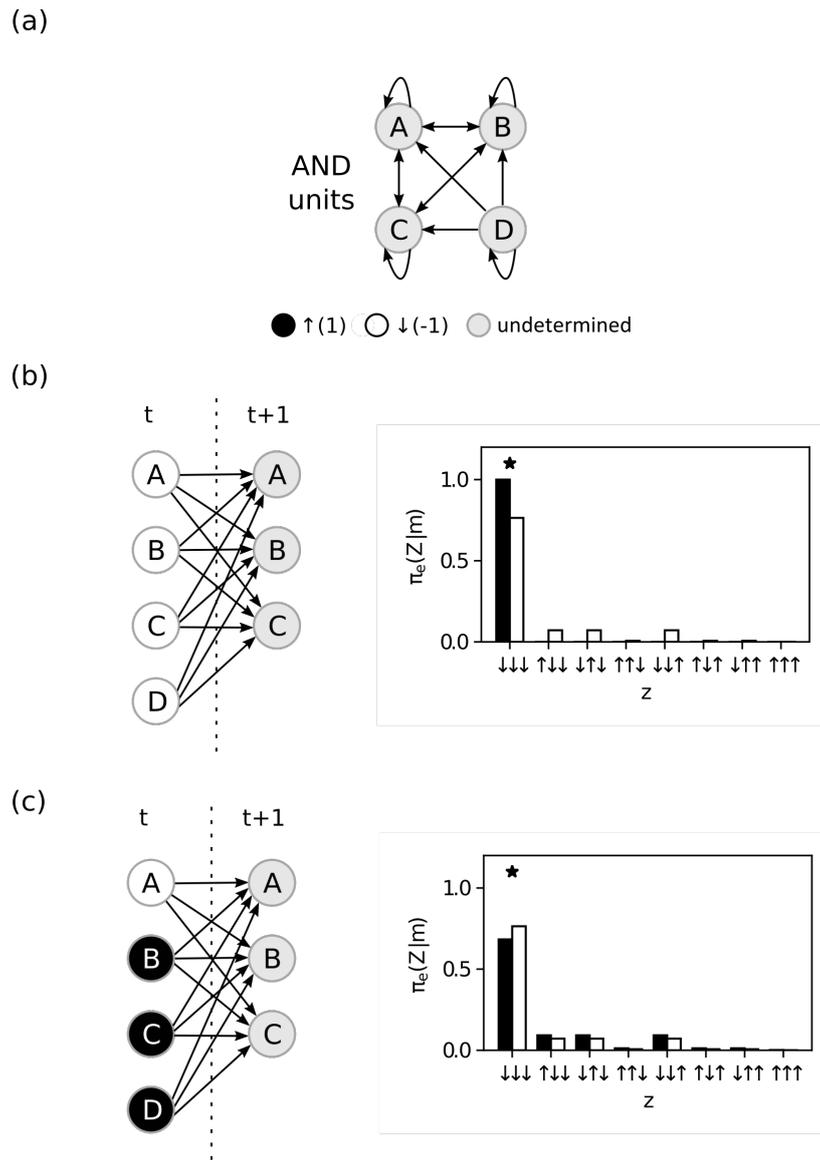


Figure 3. Information. (a) System S analyzed in this figure. All units have $\tau = 1$ and $\eta = -3$ (partially deterministic AND gates). The remaining panels show on the left the time unfolded graph of the mechanism $M = \{A, B, C, D\}$ constraining different output purviews and on the right the probability distribution of the purview $Z = \{A, B, C\}$ (effect repertoires). The black bars show the probabilities when the mechanism is constraining the purview, and the white bars show the unconstrained probabilities after the complete partition. The “*” indicates the state selected by the maximum operation in the ID function. (b) The mechanism at state $m = \{\downarrow, \downarrow, \downarrow, \downarrow\}$. The purview state $z = \{\downarrow, \downarrow, \downarrow\}$ is not only the most constrained by the mechanism (high informativeness) but also very dense (high selectivity). As a result, it has intrinsic information higher than all other states in the purview and defines the intrinsic information of the mechanism as 0.27. (c) If we change the mechanism state to $m = \{\downarrow, \uparrow, \uparrow, \uparrow\}$, the probability of observing the purview state $z = \{\downarrow, \downarrow, \downarrow\}$ is now smaller than chance. However, this probability is still very different from chance and therefore very constrained by the mechanism (high informativeness). At the same time, the state is still very dense, meaning it has a probability of happening much higher than all other states (high selectivity). Together, they define the intrinsic information of the state, which is higher than the intrinsic information of all other states in the purview, defining the intrinsic information of the mechanism as 0.08.

A different scenario is depicted if we change the state of the mechanism to $ABCD = \{\downarrow, \uparrow, \uparrow, \uparrow\}$ (Figure 3c). In this mechanism state the constrained probability of $ABC = \{\downarrow, \downarrow, \downarrow\}$

is *lower* than than the probability after partitioning. However, the mechanism is informative because the probabilities are different. At the same time, the state $ABC = \{\downarrow, \downarrow, \downarrow\}$ still has high probability while being constrained by the mechanism $ABCD = \{\downarrow, \uparrow, \uparrow, \uparrow\}$. Together the product of the informativeness and selectivity is higher for the purview state $\{\downarrow, \downarrow, \downarrow\}$ than any other state, resulting in

$$\begin{aligned}\varphi_e(m, Z, \psi^0) &= \text{ID}(\pi_e(ABC \mid ABCD = \downarrow\uparrow\uparrow\uparrow), \pi_e^{\psi^0}(ABC \mid ABCD = \downarrow\uparrow\uparrow\uparrow)) \\ &= 0.08.\end{aligned}$$

Although it may be counterintuitive to identify an effect state whose probability is decreased by the mechanism, it highlights an important feature of intrinsic information: it balances informativeness and selectivity. Informativeness is about *constraint*, meaning how much the probability of observing a given state in the purview *changes* due to being constrained by the mechanism. At the same time, selectivity is about probability *density* at a given state, meaning that this constraint is only relevant if the state is realized by the purview. If the mechanism is informative while increasing selectivity, then there is no tension between the two. However, whenever the mechanism decreases the probability of a state, there is a tension between how informative and how selective that state is. As long as *together* the product of informativeness and selectivity of a state (in this case $ABC = \{\downarrow, \downarrow, \downarrow\}$) is higher than all other states, it is selected by the maximum operation in the ID function and thus determines the intrinsic information of the mechanism.

4.2. Integrated Information

The integration postulate of IIT requires that mechanisms be *integrated* or irreducible to parts. In this section we use the system defined in Figure 4a, with $\eta = 0$ and $\tau = 1$ for all units, to investigate how mechanisms are impacted by different partitions. We compute the ID between the intact and all possible partitioned effect repertoires to measure the impact of each partition $\psi \in \Psi(M, Z)$. We identify the partition with *lowest* ID as the MIP of the candidate mechanism over a purview.

First, when considering the mechanism $M = \{A, E\} = \{\uparrow, \downarrow\}$ over the purview $Z = \{A, E\}$, the complete partition ψ^0 (partitioning the entire mechanism away from the entire purview) assigns a positive value

$$\varphi_e(m, Z, \psi^0) = \text{ID}(\pi_e(AE \mid AE = \uparrow\downarrow), \pi_e^{\psi^0}(AE \mid AE = \uparrow\downarrow)) = 0.36.$$

However, if we try the partition

$$\psi^1 = \{(\{A\}, \{A\}), (\{E\}, \{E\})\},$$

we find that the candidate mechanism is not integrated (as is obvious after inspecting Figure 4b):

$$\varphi_e(m, Z, \psi^1) = \text{ID}(\pi_e(AE \mid AE = \uparrow\downarrow), \pi_e^{\psi^1}(AE \mid AE = \uparrow\downarrow)) = 0.$$

We conclude that this candidate mechanism does not exist within the system over this purview.

Next, we consider the candidate mechanism $M = \{A, B\} = \{\uparrow, \uparrow\}$ over the purview $Z = \{A, B\}$. We observe that the partition

$$\psi^2 = \{(\{A\}, \{A, B\}), (\{B\}, \{\emptyset\})\},$$

defines an intrinsic difference

$$\varphi_e(m, Z, \psi^2) = \text{ID}(\pi_e(AB \mid AB = \uparrow\uparrow), \pi_e^{\psi^2}(AB \mid AB = \uparrow\uparrow)) = 0.36,$$

which is *smaller* than the one assigned by the complete partition ψ^0 ,

$$\varphi_e(m, Z, \psi^0) = \text{ID}(\pi_e(AB \mid AB = \uparrow\uparrow), \pi_e^{\psi^0}(AB \mid AB = \uparrow\uparrow)) = 0.51.$$

Although the ID over ψ^2 is smaller than that over the complete partition, this information is not *zero*. Moreover, the partition ψ^2 yields an ID value that is smaller than *any* other partition $\psi \in \Psi(AB, AB)$. In this case, we say that ψ^2 is the MIP ($\psi^* = \psi^2$), and that the candidate mechanism $M = \{A, B\}$ has *integrated* effect information (Figure 4c):

$$\varphi_e(m, Z) = \text{ID}(\pi(AB \mid AB = \uparrow\uparrow) \mid \pi^{\psi^*}(AB \mid AB = \uparrow\uparrow)) = 0.36.$$

Finally, for the candidate mechanism $M = \{A, B, D\} = \{\uparrow, \uparrow, \downarrow\}$ over the purview $Z = \{E, F\}$, any partition that does not include the empty set as a part in $\{M_i\}$ leads to nonzero ID. However, if we allow the empty set for M_i (as discussed in Section 3.2), the candidate mechanism is reducible because disintegrating it with the partition

$$\psi^* = \{(\{A\}, \{\emptyset\}), (\{\emptyset\}, \{F\}), (\{B, D\}, \{E\})\}$$

makes no difference to the purview states, resulting in

$$\varphi_e(m, Z) = \text{ID}(\pi_e(EF \mid ABD = \uparrow\uparrow\downarrow), \pi_e^{\psi^*}(EF \mid ABD = \uparrow\uparrow\downarrow)) = 0.$$

This occurs since B and D have opposite effects over the purview unit E , and by cutting both inputs to E we avoid changing the repertoire. Therefore, $M = \{A, B, D\}$ does not exist as a mechanism over the purview $Z = \{E, F\}$.

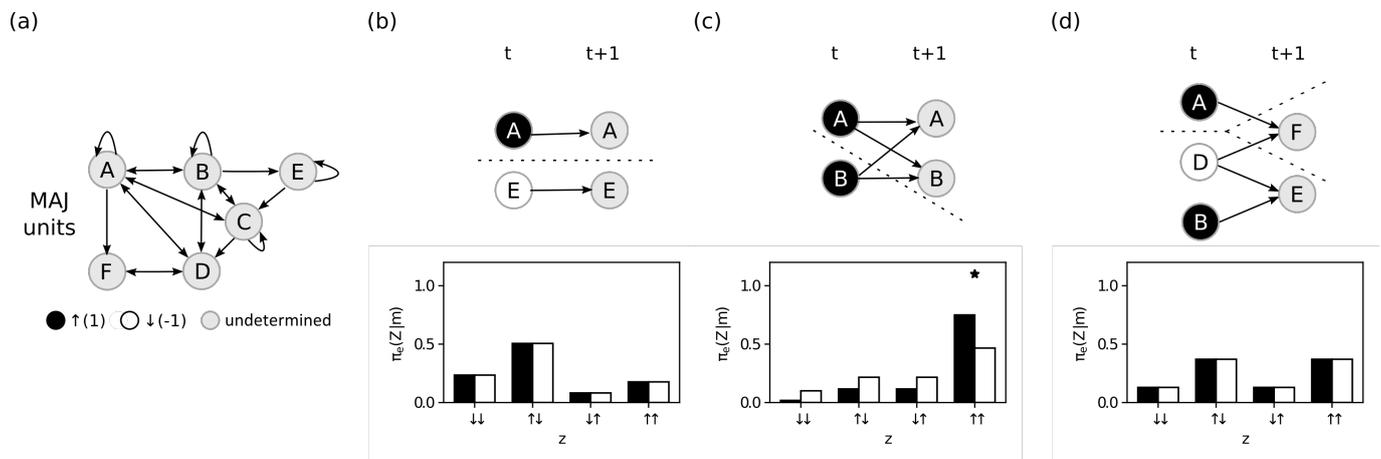


Figure 4. Integration. (a) System S analysed in this figure and in Figure 5. All units have $\tau = 1$ and $\eta = 0$ (partially deterministic MAJORITY gates). The remaining panels show on the top the time unfolded graph of different mechanisms constraining different output purviews and on the bottom the probability distributions (effect repertoires). The black bars show the probabilities when the mechanism is constraining the purview, and the white bars show the partitioned probabilities. The “*” indicates the state selected by the maximum operation in the ID function. (b) The mechanism $M = \{A, E\}$ in state $m = \{\uparrow, \downarrow\}$ constraining the purview $Z = \{A, E\}$. While the complete partition has nonzero intrinsic information, the mechanism is clearly not integrated, as revealed by the MIP partition $\psi^* = \{(\{A\}, \{A\}), (\{E\}, \{E\})\}$, resulting in zero *integrated* information. (c) The mechanism $M = \{A, B\}$ in state $m = \{\uparrow, \uparrow\}$ constraining the purview $Z = \{A, B\}$. The partition $\psi^* = \{(\{A\}, \{A, B\}), (\{B\}, \{\emptyset\})\}$ has less intrinsic information than any other partition, i.e., it is the MIP of this mechanism, and it defines the integrated information as 0.36. (d) The mechanism $M = \{A, B, D\}$ in state $m = \{\uparrow, \uparrow, \downarrow\}$ constraining the purview $Z = \{E, F\}$. The tri-partition $\psi^* = \{(\{A\}, \{\emptyset\}), (\{\emptyset\}, \{F\}), (\{B, D\}, \{E\})\}$ is the MIP and it shows that the mechanism is not integrated, i.e., the mechanism has zero integrated information.

4.3. Maximal Integrated Information

The last postulate we investigate is exclusion, which dictates that mechanisms are defined over a *definite* purview, the one over which the mechanism is maximally irreducible (has maximal integrated effect information). Using the system defined in Figure 4a, we investigate two candidate mechanisms. First, we study the candidate mechanism $M = \{A\} = \uparrow$, similar to the one in Figure 2. Since $M = \{A\}$ is first order (constituted of one unit), there is only one possible partition (the complete partition)

$$\psi^* = \psi^0 = \{(\{A\}, \{\emptyset\}), (\{\emptyset\}, \{Z\})\}.$$

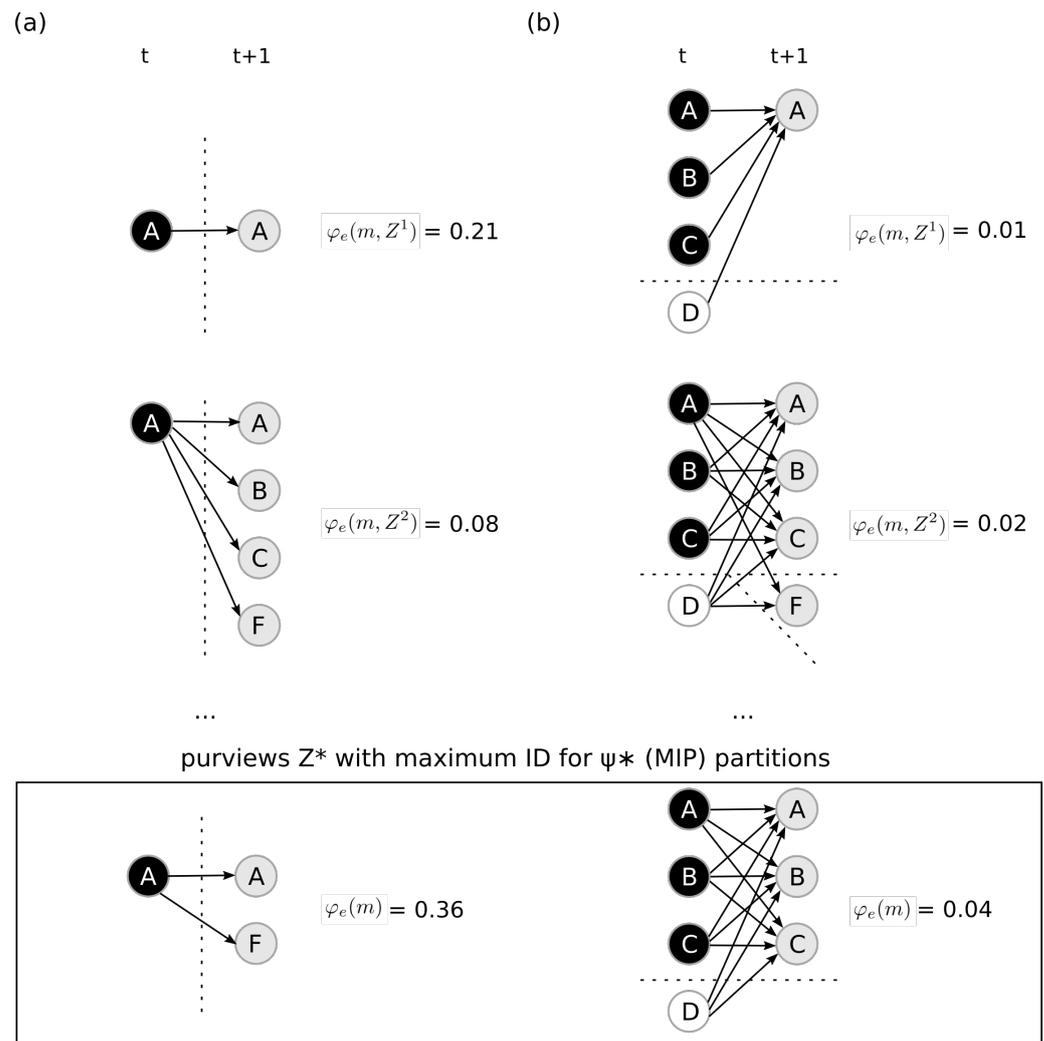


Figure 5. Exclusion. Causal graphs of different mechanisms constraining different purviews. The system S used in these examples is the same as in Figure 4a. Each line shows the mechanism M constraining different purviews Z . (a) The mechanism $M = \{A\}$ at state $m = \{\uparrow\}$. The bottom line shows the purview $Z \in S$ with maximum integrated effect information and the MIP is the complete partition. (b) The mechanism $M = \{A, B, C, D\}$ at state $m = \{\uparrow, \uparrow, \uparrow, \downarrow\}$. The bottom line is the purview $Z \in S$ with maximum integrated effect information and the MIP is $\psi^* = \{(\{A, B, C\}, \{A, B, C\}), (\{D\}, \{\emptyset\})\}$.

After computing $\varphi_e(m, Z)$ for all possible purviews $Z \in S$, we find that the mechanism has maximum integrated effect information over the purview $Z_e^* = \{A, F\}$, thus according to Equation (3) we have

$$\varphi_e(m) = \text{ID}(\pi_e(AF | A = \uparrow), \pi_e^{\psi^*}(AF | A = \uparrow)) = 0.36.$$

Next, similarly to Figure 3, we investigate the candidate mechanism $M = \{A, B, C, D\} = \{\uparrow, \uparrow, \uparrow, \downarrow\}$. After computing $\varphi_e(m, Z, \psi)$ over all possible purviews in the system ($Z \subseteq S$) and over all possible partitions for each purview ($\psi \in \Psi(ABCD, Z)$), we find that the mechanism has maximum integrated effect information over the purview $Z_e^* = \{A, B, C\}$, with partition

$$\psi^* = \{(\{A, B, C\}, \{A, B, C\}), (\{D\}, \{\emptyset\})\},$$

and that

$$\begin{aligned} \varphi_e(m) &= \text{ID}(\pi_e(ABC | ABCD = \uparrow\uparrow\uparrow\downarrow), \pi_e^{\psi^*}(ABC | ABCD = \uparrow\uparrow\uparrow\downarrow)) \\ &= 0.04. \end{aligned}$$

Finally, IIT requires that mechanisms have *both* causes and effects within the system. We perform an analogous process using the cause repertoire $\pi_c(Z | ABCD = \uparrow\uparrow\uparrow\downarrow)$ (see Equation (A2) and Figure 6) to identify the maximally irreducible cause purview Z_c^* . We find that $Z_c^* = \{A, B, F\}$ with MIP

$$\psi^* = \{(\{A\}, \{A\}), (\{B, C, D\}, \{F\})\},$$

and integrated cause information

$$\begin{aligned} \varphi_c(m) &= \text{ID}(\pi_c(ABF | ABCD = \uparrow\uparrow\uparrow\downarrow), \pi_c^{\psi^*}(ABF | ABCD = \uparrow\uparrow\uparrow\downarrow)) \\ &= 0.09. \end{aligned}$$

Since $M = \{A, B, C, D\} = \{\uparrow, \uparrow, \uparrow, \downarrow\}$ has an irreducible cause ($\varphi_c > 0$) and effect ($\varphi_e > 0$), we say that the mechanism *ABCD exists* within the system with integrated information

$$\varphi(m) = \min\{\varphi_e(m), \varphi_c(m)\} = 0.04.$$

This means that, given the system S , the mechanism $M = \{A, B, C, D\} \subseteq S$ in state $m = \{\uparrow, \uparrow, \uparrow, \downarrow\} \in \Omega_M$ specifies the *distinction*

$$\begin{aligned} X(\{A, B, C, D\} = \{\uparrow, \uparrow, \uparrow, \downarrow\}) \\ = \{(Z_c^* = \{A, B, F\} = \{\uparrow, \downarrow, \downarrow\}, Z_e^* = \{A, B, C\} = \{\uparrow, \uparrow, \uparrow\}, \varphi(\{A, B, C, D\}) = 0.04)\}. \end{aligned}$$

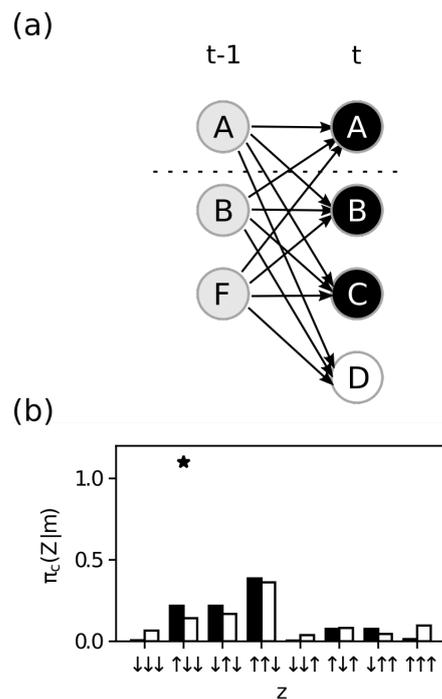


Figure 6. Integrated cause information. (a) Causal graph of mechanism $M = \{A, B, C, D\}$ at state $m = \{\uparrow, \uparrow, \downarrow, \downarrow\}$ constraining the purview $Z = \{A, B, F\}$, which has the maximum integrated information of all $Z \subseteq S$ and defines the mechanism integrated information. (b) The black bars show the probabilities when the mechanism is constraining the purview (cause repertoire), and the white bars show the probabilities after the partition (partitioned cause repertoire). The “*” indicates the state selected by the maximum operation in the ID function and defines Z_C^* .

5. Discussion

Mechanism integrated information $\varphi(m)$ is a measure of the intrinsic cause–effect power of a mechanism $M = m$ within a system. It reflects how much a mechanism as a whole (above and beyond its parts) constrains the units in its cause and effect purview. We characterize three properties of information based on the postulates of IIT: *causality*, *intrinsicity*, and *specificity*, and demonstrate that there is a unique measure (ID) that satisfies these properties. Notably, intrinsicity requires that information increases when expanding a purview with a fully constrained unit (expansion) but decreases when expanding a purview with a fully unconstrained unit (dilution). In situations with partial constraint, finding a unique measure gives us a principled way to balance expansion and dilution.

Early versions of IIT used the KLD to measure the difference between probability distributions [4,16]. The KLD was a practical solution given its unique mathematical properties and ubiquity in information theory; however, there was no principled reason to select it over any other measure. In [3], the KLD was replaced by the EMD, which was an initial attempt to capture the idea of relations among distinctions. The more two distinctions overlap in their purview units and states, the smaller the EMD distance between them; this distance was used as the ground distance to compute the system integrated information (Φ). This aspect of the EMD is now encompassed by including relations as an explicit part of the cause–effect structure, defined in a way that is consistent with the postulates of IIT [10]. The new intrinsic difference measure is the first principled measure based on properties derived from the postulates of IIT. Importantly, ID is shown to be the unique measure that satisfies the three properties—causality, intrinsicity and specificity—the KLD and EMD measures do not satisfy intrinsicity or specificity. See Appendix C for an example of how the different measures change the purview with maximum integrated information.

Furthermore, we define a set of possible partitions of a mechanism and its purview ($\Psi(M, Z)$), which ensures that the mechanism is destroyed (“distinguished”) after the

partition operation is applied. Previous formulations of mechanism integrated information restricted the set of all possible partitions to bipartitions of a mechanism and its purview but allowed for partitions that do not qualify as “disintegrating” the mechanism (for example, cutting away a single purview unit) [3]. For most mechanisms the minimum information partition ψ^* still partitions the mechanism in two parts; exceptions tend to occur if multiple inputs to the same unit counteract each other. The requirement for disintegrating partitions is more consequential, especially for first-order mechanisms (those composed of a single unit). Without this restriction, the ψ^* of a first-order mechanism would always be to cut away its weakest purview unit, and the integrated information of the mechanism would then be equal to the information the mechanism specifies about its least constrained purview unit. With the disintegrating partitions, a first-order mechanism must be cut away from its entire purview, reflecting the notion that everything that a first-order mechanism does is irreducible (since it is unified).

The particular partition $\psi^* \in \Psi(M, Z)$ that yields the minimum ID between partitioned and unpartitioned repertoires defines the integrated information of a mechanism over a purview. The balance between expansion and dilution, together with the set of possible partitions, allows us to find the purviews Z_c^* and Z_e^* with maximum integrated cause and effect information. Moreover, the ID measure identifies the specific cause state z_c^* and effect state z_e^* that maximize the mechanism’s integrated cause and effect information. Finally, the overall integrated information of a mechanism M in state m is the minimum between its integrated cause and effect information: $\varphi(m) = \min\{\varphi_c(m), \varphi_e(m)\}$.

Mechanisms that exist within a system ($\varphi(m) > 0$) specify a distinction (a cause and effect) for the system, and the set of all distinctions and the relations among them define the cause–effect structure of the system [10]. As mentioned above (Section 3.1.5), it is in principle possible that there are multiple solutions for $Z_c^* = z_c^*$ or $Z_e^* = z_e^*$ for a given mechanism m in degenerate systems with symmetries in connectivity and functionality (but note that $\varphi(m)$ is uniquely defined). However, by the exclusion postulate, distinctions within the cause–effect structure of a conscious system should specify a definite cause and effect, which means that they should specify a definite cause and effect purview in a specific state. As also argued in [17], distinctions that are underdetermined should thus not be included in the cause–effect structure until the tie between purviews or states can be resolved. In physical systems that evolve in time with a certain amount of variability and indeterminism, ties are likely short lived and may typically resolve on a faster scale than the temporal scale of experience.

The principles and arguments applied to mechanism information will need to be extended to relation integrated information and system integrated information, laying the ground work for an updated 4.0 version of the theory. Relations describe how causes and effects overlap in the cause–effect structure, by being over the same units and specifying the same state. Like distinctions, relations exist within the cause–effect structure, and their existence is quantified by an analogous notion of relation integrated information (φ_r). Similarly, the intrinsic existence of a candidate system and its cause–effect structure as a PSC with an experience is quantified by system integrated information (Φ). Both φ_r and Φ measure the difference made by “cutting apart” the object (relation or system) according to its ψ^* . As a measure of existence, the difference measures used for φ_r and Φ must also satisfy the causality, intrinsicity and specificity properties. In the case of Φ , the expansion and dilution properties will need to be adapted to the combinatorial nature of the measure, since adding a single unit to a PSC doubles the number of potential distinctions.

According to IIT, a system is a PSC if its cause–effect structure is maximally irreducible (it is a maximum of system integrated information, Φ). Moreover, if a system is a PSC, then its subjective experience is identical to its cause–effect structure [3]. Since the quantity and quality of consciousness are what they are, the cause–effect structure cannot vary arbitrarily with the chosen measure of intrinsic information. For this reason, a measure of intrinsic information that is based on the postulates and is unique is a critical requirement of the theory.

Author Contributions: Conceptualization, L.S.B., W.M., L.A. and G.T.; software, L.S.B.; investigation, L.S.B., W.M. and L.A.; writing—original draft preparation, L.S.B. and W.M.; writing—review and editing, L.S.B., L.A., W.M. and G.T.; visualization, L.S.B.; supervision, G.T.; project administration, G.T.; funding acquisition, G.T. All authors have read and agreed to the published version of the manuscript.

Funding: This project was made possible through the support of a grant from Templeton World Charity Foundation, Inc. (#TWCF0216) and by the Tiny Blue Dot Foundation (UW 133AAG3451). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation, Inc. and Tiny Blue Dot Foundation.

Institutional Review Board Statement: Not applicable.

Informed Consent 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: Would like to thank Shuntaro Sasai, Andrew Haun, William Mayner, Graham Findlay and Matteo Grasso for their helpful comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Cause and Effect Repertoires

The cause and the effect repertoire can be derived from the system defined in Equation (1). The random variables S_i define the system state space $\Omega_S = \times_{i=1}^n \Omega_{S_i}$, where \times is the cross product of each individual state space. We also require that the random variables are conditional independent

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

that the transitions are time invariant

$$p(s_{t+1}|s_t) = p(s_t|s_{t-1}),$$

and that the probabilities are well-defined for all possible states

$$\exists p(s_{t+1}|s_t) \text{ for all } s_t, s_{t+1} \in \Omega_S.$$

Given the former definitions, the stochastic system S corresponds to a causal network where $p(s_{t+1}|s_t) = p(s_{t+1}|do(s_t))$ [12,18,19].

We use uppercase letters as parameters of the probability function to define probability distributions, e.g., $p(S_{t+1}|s_t) = \{p(s_{t+1}|s_t) : s_{t+1} \in \Omega_S\}$, and the operators \sum and \prod are applied to each state independently.

Appendix A.1. Causal Marginalization

Given a mechanism $M \subseteq S$ in a state $m_t \in \Omega_M$ and a purview $Z \subseteq S$, *causal marginalization* serves to remove any contributions to the repertoire of states Ω_Z that are outside the mechanism M and purview Z . Explicitly, given the set $W = S \setminus M$, we define the effect repertoire of a single unit $Z_i \in Z$ as

$$\pi_e(Z_i | m) = \sum_{w_t \in \Omega_W} p(Z_{i,t+1} | m_t, w_t) |\Omega_W|^{-1}.$$

Note that, for causal marginalization, we impose a uniform distribution as $p(W_t)$. This ensures that the repertoire captures the constraints due to the mechanism alone and not to whatever external factors might bias the variables in W to one state or another.

Given the set $V = S \setminus Z$, the cause repertoire for a single unit $M_i \in M$, using Bayes' rule, is

$$\pi_c(Z | m_i) = \frac{\sum_{v_{t-1} \in \Omega_V} p(m_{i,t} | Z_{t-1}, v_{t-1})}{\sum_{s_{t-1} \in \Omega_S} p(m_{i,t} | s_{t-1})},$$

where again we impose the uniform distributions as $p(V_{t-1})$, $p(Z_{t-1})$, and $p(S_{t-1})$.

Note that the transition probability function $p(Z_{t+1} | m_t)$ not only contains dependencies of Z_{t+1} on m_t but also correlations between the variables in Z due to common inputs from units in W , which should not be counted as constraints due to m_t . To discount such correlations, we define the effect repertoire over a set Z of r units Z_i as the product of the effect repertoires over individual units

$$\pi_e(Z | m) = \bigotimes_{i=1}^r \pi_e(Z_i | m), \tag{A1}$$

where \otimes is the Kronecker product of the probability distributions. In the same manner, given that the mechanism M has q units M_i , we define the cause repertoire of Z as

$$\pi_c(Z | m) = \frac{\prod_{i=1}^q \pi_c(Z | m_i)}{\sum_{z \in \Omega_Z} \prod_{i=1}^q \pi_c(z | m_i)}. \tag{A2}$$

Appendix A.2. Partitioned Repertoires

Given a partition $\psi \in \Psi(M, Z)$ constituted of k parts (see Equation (4)), we can define the partitioned repertoire

$$\pi^\psi(Z | m) = \bigotimes_{j=1}^k \pi(Z_j | m_j), \tag{A3}$$

with $\pi(\emptyset | m_j) = \pi(\emptyset) = 1$. In the case of $m_j = \emptyset$, $\pi(Z_j | \emptyset) = \pi(Z_j)$ corresponds to an unconstrained effect repertoire

$$\pi_e(Z) = \bigotimes_{i=1}^r \pi_e(Z_i) = \bigotimes_{i=1}^r \sum_{s_t \in \Omega_S} p(Z_{i,t+1} | s_t) |\Omega_S|^{-1}, \tag{A4}$$

which follows from Equation (A1) and cause repertoire

$$\pi_c(Z) = \frac{1}{|\Omega_Z|}, \tag{A5}$$

which follows from Equation (A2).

Appendix B. Full Statement and Proof of Theorem 1

We now give the full statement and proof of the Theorem 1, demonstrating the uniqueness of the function f (see also [15,20]). We start with some preliminary definitions:

$$\begin{aligned} \mathbb{R}^+ &= \{x \in \mathbb{R} : x > 0\}, \quad \mathbb{N}_2 = \{2, 3, 4, \dots\}, \\ J &= (0, 1), \quad \hat{J} = (0, 1], \quad \bar{J} = [0, 1], \quad K = (\bar{J} \times \bar{J}) \setminus (\hat{J} \times \{0\}), \\ I_\delta(x) &= \{y \in J : |x - y| < \delta\}. \end{aligned}$$

For each $n \in \mathbb{N}_2$, we further define

$$\Gamma^n = \left\{ X^n = (x_1, \dots, x_n) : x_1, \dots, x_n \in \bar{J}, \sum_{\alpha=1}^n x_\alpha = 1 \right\},$$

$$V^n = (1, 0, \dots, 0) \in \Gamma^n, \quad U^n = \left(\frac{1}{n}, \dots, \frac{1}{n} \right) \in \Gamma^n,$$

$$\Delta^n = \{ (X^n, Y^n) \in \Gamma^n \times \Gamma^n : (x_\alpha, y_\alpha) \in K, \text{ for all } \alpha \in \{1, \dots, n\} \}.$$

Using these definitions, we further define the following properties.

Property I : Causality. Let $(P^n, Q^n) \in \Delta^n$. The difference $D(P^n, Q^n)$ is defined as $D: \Delta^n \rightarrow \mathbb{R}$, such that

$$D(P^n, Q^n) = 0 \iff P^n = Q^n.$$

Property II : Intrinsicity. Let $(P^l, Q^l) \in \Delta^l$ and $(P^m, Q^m) \in \Delta^m$. Then

$$(a) \text{ expansion: } D(V^l \otimes V^m, P^l \otimes Q^m) = D(V^l, P^l) + D(V^m, Q^m),$$

$$(b) \text{ dilution: } D(P^l \otimes U^m, Q^l \otimes U^m) = \frac{D(P^l, Q^l) + D(U^m, U^m)}{m},$$

where $P^l \otimes Q^m = (p_1q_1, \dots, p_1q_m, \dots, p_lq_1, \dots, p_lq_m) \in \Gamma^{lm}$ and from Property I $D(U^m, U^m) = 0$.

Property III : Specificity. The difference must be state-specific, meaning there exists $f: K \rightarrow \mathbb{R}$ such that for all $(P^n, Q^n) \in \Delta^n$ we have $D(P^n, Q^n) = f(p_\alpha, q_\alpha)$, where $\alpha \in \{1, \dots, n\}$, $p_\alpha \in P^n$ and $q_\alpha \in Q^n$. More precisely, we define

$$D(P^n, Q^n) := \max_\alpha \{ |f(p_\alpha, q_\alpha)| \},$$

where f is continuous on K , analytic on $\hat{J} \times J$ and $f(0, q_\alpha)$ is analytic on J .

The following lemma allows the analytic extension of real analytic functions.

Lemma A1 (See Proposition 1.2.3 in [21]). *If f and g are real analytic functions on an open interval $U \in \mathbb{R}$ and if there is a sequence of distinct points $\{x_n\}_n \in U$ with $x_0 = \lim_{n \rightarrow \infty} x_n \in U$ such that*

$$f(x_n) = g(x_n),$$

then

$$f(x) = g(x), \quad \text{for all } x \in U.$$

Corollary A1 (See Corollary 1.2.6 in [21]). *If f and g are analytic functions on an open interval U and if there is an open interval $W \subseteq U$ such that*

$$f(x) = g(x), \quad \text{for all } x \in W,$$

then

$$f(x) = g(x), \quad \text{for all } x \in U.$$

The following lemma shows that a strict maximum over continuous functions, each evaluated at fixed points, must hold for an open interval around such fixed points.

Lemma A2. *Let $g_\alpha: J \rightarrow \mathbb{R}$ be continuous functions, where $\alpha \in \{1, \dots, n\}$, fix $x_1, \dots, x_n \in J$. If there exists α^* such that for $\alpha \neq \alpha^*$,*

$$g_{\alpha^*}(x_{\alpha^*}) > g_\alpha(x_\alpha),$$

then there exists $\delta > 0$ such that

$$\max_{\alpha} \{g_{\alpha}(\hat{x}_{\alpha})\} = g_{\alpha^*}(\hat{x}_{\alpha^*}),$$

for all $\hat{x}_{\alpha} \in I_{\delta}(x_{\alpha})$ and for all $\alpha \in \{1, \dots, n\}$.

We now provide the solution to a functional equation similar to the Pexider logarithmic equation [22].

Lemma A3. Let $f, g, h : J \rightarrow \mathbb{R}$ be analytic functions on J . Suppose the functional equation

$$|f(pq)| = \max\{|g(p)|, |h(p)|\} + \max\{|g(q)|, |h(q)|\},$$

holds for all $pq \in I_{\delta}(p'q')$, where $I_{\delta}(p'q') \subseteq J$. Then there exists $c, d \in \mathbb{R}$ such that

$$f(x) = c \log(x) + d, \quad \text{for all } x \in J.$$

Proof. First, for some $i \in \{g, h\}$ suppose that there exists $(p_i, q_i) \in J \times J$ such that $p_i q_i \in I_{\delta}(p'q')$ and

$$|i(p_i)| = \max\{|g(p_i)|, |h(p_i)|\}$$

is a strict maximum. Then by Lemma A2 there exists $\delta_p > 0$ such that

$$|i(p)| = \max\{|g(p)|, |h(p)|\}, \quad \text{for all } p \in I_{\delta_p}(p_i). \tag{A6}$$

Second, if there does not exist $(p_i, q_i) \in J \times J$ such that $p_i q_i \in I_{\delta}(p'q')$ and $|i(p_i)|$ is a strict maximum, then we set $q_i = q', p_i = p'$ and $\delta_p = \frac{\delta}{q'}$ so that Equation (A6) holds since $|g(p)| = |h(p)|$ for all $(p, q) \in J \times J$ such that $pq \in I_{\delta}(p'q')$. Next, define $\delta' := \min\{\delta - |p_i q_i - p'q'|, \delta_p q'\}$. Suppose that there exists $q_j \in J$ such that $p_i q_j \in I_{\delta'}(p_i q_i)$, and for some $j \in \{g, h\}$,

$$|j(q_j)| = \max\{|g(q_j)|, |h(q_j)|\}$$

is a strict maximum. Then by Lemma A2 there exists $\delta_q > 0$ such that

$$|j(q)| = \max\{|g(q)|, |h(q)|\}, \quad \text{for all } q \in I_{\delta_q}(q_j). \tag{A7}$$

Finally, if there does not exist $q_j \in J$ such that $p_i q_j \in I_{\delta'}(p_i q_i)$ and $|j(q_j)|$ is a strict maximum, then we set $q_j = q_i$ and $\delta_q = \frac{\delta'}{p_i}$ so that Equation (A7) holds since $|g(q)| = |h(q)|$ for all $(p, q) \in J \times J$ such that $pq \in I_{\delta'}(p_i q_i)$. Let $pq = x$ and define $\delta'' := \min\{\delta' - |p_i q_j - p_i q_i|, \delta_q p_i\}$, then

$$|f(x)| = |i(p)| + |j(q)|, \quad \text{for all } x \in I_{\delta''}(p_i q_j).$$

Moreover, it follows that one of the following options must be true

$$f(x) = \pm i(p) \pm j(q), \quad \text{for all } x \in I_{\delta''}(p_i q_j).$$

Since the functions are analytic on J and therefore twice differentiable, then

$$\frac{\partial}{\partial q} \left[\frac{\partial}{\partial p} [f(x)] \right] = x \frac{d}{dx^2} f(x) + \frac{d}{dx} f(x) = \pm \frac{\partial}{\partial q} \left[\frac{\partial}{\partial p} i(p) \right] \pm \frac{\partial}{\partial q} \left[\frac{\partial}{\partial p} j(q) \right] = 0.$$

Integrating with respect to x yields

$$f(x) = c \log(x) + d,$$

for $c, d \in \mathbb{R}$ and for all $x \in I_{\mathcal{G}''}(x')$ where $x' = p'q'$. Since f is analytic on J and since $I_{\mathcal{G}''}(x') \subset J$, by Corollary A1, we can extend $f(x)$ such that

$$f(x) = c \log(x) + d, \quad \text{for all } x \in J.$$

□

Lemma A4. If $D : \Delta^n \rightarrow \mathbb{R}$ satisfies properties I and III for some $f : K \rightarrow \mathbb{R}$, then

$$f(p, p) = 0, \quad \text{for all } p \in \bar{J}.$$

Proof. Let $P^2 = (p, 1 - p) \in \Gamma^2$. By properties I and III

$$0 = D(P^2, P^2) = \max\{|f(p, p)|, |f(1 - p, 1 - p)|\}, \quad \text{for all } p \in \bar{J}. \quad (\text{A8})$$

□

Theorem A1. Let $(P^n, Q^n) \in \Delta^n$ for some $n \in \mathbb{N}_2$ and $D : \Delta^n \rightarrow \mathbb{R}$ where D satisfies properties I, II and III. Then

$$D(P^n, Q^n) = \max_{\alpha} \{|f(p_{\alpha}, q_{\alpha})|\}, \quad (\text{A9})$$

where for some $k \in \mathbb{R} \setminus \{0\}$,

$$f(p, q) = k p \log\left(\frac{p}{q}\right), \quad \text{for all } (p, q) \in K. \quad (\text{A10})$$

Proof of Theorem A1. First we show that the function in Equation (A10) satisfies properties I, II and III. To see that the function satisfies Property I, notice that for each $(P^n, Q^n) \in \Delta^n$ where $P^n \neq Q^n$, since $k \neq 0$, then there exists $\beta \in \{1, \dots, n\}$ such that

$$D(P^n, Q^n) = \max_{\alpha} \left\{ \left| k p_{\alpha} \log\left(\frac{p_{\alpha}}{q_{\alpha}}\right) \right| \right\} \geq \left| k p_{\beta} \log\left(\frac{p_{\beta}}{q_{\beta}}\right) \right| > 0,$$

and for each $P^n = Q^n$,

$$D(P^n, P^n) = \max_{\alpha} \left\{ \left| k p_{\alpha} \log\left(\frac{p_{\alpha}}{p_{\alpha}}\right) \right| \right\} = 0.$$

To see that it satisfies Property II.a, for each $P^l \in \Gamma^l$ and for each $Q^m \in \Gamma^m$

$$\begin{aligned} D(V^l \otimes V^m, P^l \otimes Q^m) &= \max \left\{ \left| k \log\left(\frac{1}{p_1 q_1}\right) \right|, 0 \right\} = \left| k \log\left(\frac{1}{p_1 q_1}\right) \right| \\ &= \left| k \log\left(\frac{1}{p_1}\right) \right| + \left| k \log\left(\frac{1}{q_1}\right) \right| \\ &= \max \left\{ \left| k \log\left(\frac{1}{p_1}\right) \right|, 0 \right\} + \max \left\{ \left| k \log\left(\frac{1}{q_1}\right) \right|, 0 \right\} \\ &= D(V^l, P^l) + D(V^m, Q^m). \end{aligned}$$

Similarly by Property II.b notice that for each $(P^l, Q^l) \in \Delta^l$

$$\begin{aligned}
 D(P^l \otimes U^m, Q^l \otimes U^m) &= \max_{\alpha} \left\{ \left| \frac{kp_{\alpha}}{m} \log \left(\frac{p_{\alpha}}{q_{\alpha}} \right) \right| \right\} \\
 &= \frac{1}{m} \max_{\alpha} \left\{ \left| kp_{\alpha} \log \left(\frac{p_{\alpha}}{q_{\alpha}} \right) \right| \right\} \\
 &= \frac{1}{m} D(P^l, Q^l).
 \end{aligned}$$

It is clear that the function f in Equation (A10) satisfies Property III.

The remaining part of the proof is divided into two steps:

Step 1. First we show that under four assumptions, f satisfies properties I, II and III iff $f(p, q) = kp \log\left(\frac{p}{q}\right)$, $k \in \mathbb{R} \setminus \{0\}$.

Step 2. Next we show that if any of our assumptions is violated, then no suitable f exists.

Verification of Step 1. We apply Property II.a with $P^2 = (p, 1 - p)$, $Q^2 = (q, 1 - q)$ for some $p, q \in J$ where $p \neq q$. We then have

$$D(V^2 \otimes V^2, P^2 \otimes Q^2) = D(V^2, P^2) + D(V^2, Q^2).$$

By Property III, the following identity holds

$$\begin{aligned}
 \max\{|f(1, pq)|, |f(0, p(1 - q))|, |f(0, (1 - p)q)|, |f(0, (1 - p)(1 - q))|\} \\
 = \max\{|f(1, p)|, |f(0, 1 - p)|\} + \max\{|f(1, q)|, |f(0, 1 - q)|\}.
 \end{aligned} \tag{A11}$$

Our first assumption (AS1) states that there exists some $p', q' \in J$ such that $|f(1, p'q')|$ is a strict maximum. By Lemma A2, there exists a $\delta > 0$ such that

$$|f(1, pq)| = \max\{|f(1, p)|, |f(0, 1 - p)|\} + \max\{|f(1, q)|, |f(0, 1 - q)|\}, \tag{A12}$$

for all $pq \in I_{\delta}(p'q')$. Further, by Lemma A3 there exists $c, d \in \mathbb{R}$ such that

$$f(1, q) = c \log(q) + d, \quad \text{for all } q \in J,$$

and since by Property III f is continuous, the application of Lemma A4 yields

$$\lim_{q \rightarrow 1} f(1, q) = \lim_{q \rightarrow 1} c \log(q) + d = d = 0,$$

i.e., for $k_1 = -c$, we have

$$f(1, q) = k_1 \log\left(\frac{1}{q}\right), \quad \text{for all } q \in J. \tag{A13}$$

Now applying Property II.b for $l = 2$, $P^2 = V^2$, $Q^2 = (r, 1 - r)$ for all $r \in J$ and for each $m \in \mathbb{N}_2$, we have

$$D(V^2 \otimes U^m, Q^2 \otimes U^m) = \frac{1}{m} D(V^2, Q^2),$$

and by Property III

$$\max\left\{ \left| f\left(\frac{1}{m}, \frac{r}{m}\right) \right|, \left| f\left(0, \frac{1-r}{m}\right) \right| \right\} = \frac{1}{m} \max\{|f(1, r)|, |f(0, 1 - r)|\}, \tag{A14}$$

for all $r \in J$. Our second assumption (AS2) states that there exists $q_1 \in J$ such that $|f(1, q_1)| > |f(0, 1 - q_1)|$. For some $a \in \{-1, 1\}$, we have

$$\max\{|f(1, q_1)|, |f(0, 1 - q_1)|\} = af(1, q_1).$$

Further, by Lemma A2 and Equation (A13), there exists $\delta > 0$ such that

$$\max\{|f(1, r)|, |f(0, 1 - r)|\} = af(1, r) = ak_1 \log\left(\frac{1}{r}\right), \quad \text{for all } r \in I_\delta(q_1).$$

Plugging this result back into Equation (A14) yields

$$\max\left\{\left|f\left(\frac{1}{m}, \frac{r}{m}\right)\right|, \left|f\left(0, \frac{1-r}{m}\right)\right|\right\} = \frac{ak_1}{m} \log\left(\frac{1}{r}\right), \quad \text{for all } r \in I_\delta(q_1). \tag{A15}$$

Our third assumption (AS3) states that $\left|f\left(0, \frac{1-r}{m}\right)\right|$ is never a strict maximum in Equation (A15), so that for some $b \in \{-1, 1\}$, we have

$$\max\left\{\left|f\left(\frac{1}{m}, \frac{r}{m}\right)\right|, \left|f\left(0, \frac{1-r}{m}\right)\right|\right\} = bf\left(\frac{1}{m}, \frac{r}{m}\right), \quad \text{for all } r \in I_\delta(q_1).$$

Let $q = \frac{r}{m}$ and let $I = I_\delta\left(\frac{r}{m}\right) \cap I_\delta\left(\frac{q_1}{m}\right)$. Then by Equation (A15)

$$f\left(\frac{1}{m}, q\right) = \frac{k_m}{m} \log\left(\frac{1}{mq}\right), \quad \text{for all } q \in I,$$

where $k_m = \frac{ak_1}{b}$. By Corollary A1, we can extend $f\left(\frac{1}{m}, q\right)$ to J

$$f\left(\frac{1}{m}, q\right) = \frac{k_m}{m} \log\left(\frac{1}{mq}\right), \quad \text{for all } q \in J, \tag{A16}$$

where $k_m \in \{+k_1, -k_1\}$.

Let $n \in \mathbb{N}_2$ and let $0 < q_2 < \frac{n-1}{2n}$, then $q_2 \in J$. By Property II.b for $l = 2$, $P^2 = \left(\frac{n-1}{2n}, \frac{n+1}{2n}\right)$, $Q^2 = (q_2, 1 - q_2)$ and $m = (n - 1)(n + 1)$, we have

$$D(P^2 \otimes U^m, Q^2 \otimes U^m) = \frac{1}{m} D(P^2, Q^2),$$

and by Property III

$$\begin{aligned} \max\left\{\left|f\left(\frac{1}{2n(n+1)}, \frac{q_2}{(n-1)(n+1)}\right)\right|, \left|f\left(\frac{1}{2n(n-1)}, \frac{1-q_2}{(n-1)(n+1)}\right)\right|\right\} \\ = \frac{1}{(n-1)(n+1)} \max\left\{\left|f\left(\frac{n-1}{2n}, q_2\right)\right|, \left|f\left(\frac{n+1}{2n}, 1-q_2\right)\right|\right\}. \end{aligned}$$

By Equation (A16), we have

$$\begin{aligned} \max\left\{\left|\frac{k_{2n(n+1)}}{2n(n+1)} \log\left(\frac{n-1}{2nq_2}\right)\right|, \left|\frac{k_{2n(n-1)}}{2n(n-1)} \log\left(\frac{n+1}{2n(1-q_2)}\right)\right|\right\} \\ = \frac{1}{(n-1)(n+1)} \max\left\{\left|f\left(\frac{n-1}{2n}, q_2\right)\right|, \left|f\left(\frac{n+1}{2n}, 1-q_2\right)\right|\right\}. \end{aligned}$$

Since $q_2 < \frac{n-1}{2n} < \frac{1}{2}$, this yields

$$\frac{ak_{2n(n+1)}(n-1)}{2n} \log\left(\frac{n-1}{2nq_2}\right) = \max\left\{\left|f\left(\frac{n-1}{2n}, q_2\right)\right|, \left|f\left(\frac{n+1}{2n}, 1-q_2\right)\right|\right\},$$

for some $a \in \{+1, -1\}$. Then we have that for the sequence $h_{\frac{1}{2}} := \left\{\frac{n-1}{2n}\right\}_n$

$$k_{p_n} p_n \log\left(\frac{p_n}{q}\right) = \max\{|f(p_n, q)|, |f(1 - p_n, 1 - q)|\}, \tag{A17}$$

for all $(p_n, q) \in h_{\frac{1}{2}} \times (0, p_n)$, where $k_{p_n} = a_{p_n} k_{2n(n+1)}$. Our fourth and last assumption (AS4) is that $|f(1 - p_n, 1 - q)|$ is a strict maximum only for a finite number of $p_n \in h_{\frac{1}{2}}$. More specifically, let

$$\bar{\mathcal{P}} = \{p_n \in h_{\frac{1}{2}} : \exists q \in (0, p_n) \text{ s.t. } |f(1 - p_n, 1 - q)| > |f(p_n, q)|\}.$$

Then (A4) states that $\sup\{\bar{\mathcal{P}}\} < \frac{1}{2}$ where, for convention, $\sup\{\bar{\mathcal{P}}\} := -\infty$ if $\bar{\mathcal{P}} = \emptyset$. Let $n' := 0$ if $\bar{\mathcal{P}} = \emptyset$, else there exists $n' \in \mathbb{N}_2$ such that $\sup\{\bar{\mathcal{P}}\} = \frac{n'-1}{2n'}$. Define $h'_{\frac{1}{2}} = \{\frac{n+n'-1}{2(n+n')}\}_n$. Then for a fixed $p_n \in h'_{\frac{1}{2}}$, there exists $q_n \in (0, p_n)$ such that

$$\max\{bf(p_n, q_n), |f(1 - p_n, 1 - q_n)|\} = bf(p_n, q_n),$$

for some $b \in \{+1, -1\}$. By Lemma A2, there exists $\delta > 0$ such that

$$\max\{bf(p_n, q), |f(1 - p_n, 1 - q)|\} = bf(p_n, q), \quad \text{for all } q \in I_\delta(q_n).$$

By Equation (A17), for $I_n = (0, p_n) \cap I_\delta(q_n)$, we have

$$f(p_n, q) = k_{p_n} p_n \log\left(\frac{p_n}{q}\right), \quad \text{for all } (p_n, q) \in h_{\frac{1}{2}} \times I_n, \tag{A18}$$

where the sign b was absorbed by the constant k_{p_n} . By Corollary A1, for a fixed $p_n^* \in h_{\frac{1}{2}}$, we can extend $f(p_n^*, q)$ to J , i.e.,

$$f(p_n^*, q) = k_{p_n^*} p_n^* \log\left(\frac{p_n^*}{q}\right), \quad \text{for all } q \in J.$$

For a fixed $q^* \in J$, since by Property III f is continuous, we have

$$\lim_{n \rightarrow \infty} k_{p_n} = k \in \{+k_1, -k_1\}.$$

By Lemma A1, we can uniquely extend $f(p_n, q^*)$ to J such that

$$f(p, q^*) = kp \log\left(\frac{p}{q^*}\right), \quad \text{for all } p \in J.$$

Since this is true for all $q^* \in J$, we have that

$$f(p, q) = kp \log\left(\frac{p}{q}\right), \quad \text{for all } (p, q) \in (J \times J).$$

Note that $k = 0$ violates Property I since for some $q \in J$ and $Q^2 = (q, 1 - q) \neq V^2$, we have

$$D(V^2, Q^2) = \max\{|f(1, q)|, |f(0, 1 - q)|\} = 0.$$

By Property III, f is continuous in K and the following limits exist for all $p, q \in J$:

$$\begin{aligned}
 f(1, q) &= \lim_{p \rightarrow 1^-} kp \log\left(\frac{p}{q}\right) = k \log\left(\frac{1}{q}\right), \\
 f(p, 1) &= \lim_{q \rightarrow 1^-} kp \log\left(\frac{p}{q}\right) = kp \log(p), \\
 f(0, q) &= \lim_{p \rightarrow 0^+} kp \log\left(\frac{p}{q}\right) = 0, \\
 f(0, 1) &= \lim_{p \rightarrow 0^+} kp \log(p) = 0, \\
 f(0, 0) &= \lim_{p \rightarrow 0^+} kp \log\left(\frac{p}{p}\right) = 0, \\
 f(1, 1) &= \lim_{p \rightarrow 1^-} kp \log\left(\frac{p}{p}\right) = 0.
 \end{aligned}$$

Consequently,

$$f(p, q) = kp \log\left(\frac{p}{q}\right), \quad \text{for all } (p, q) \in K.$$

Verification of Step 2. Up until here we have showed that Equation (A10) not only defines a function which satisfies properties I, II and III, but it also defines the *only* function which satisfies properties I, II and III for $l = m = 2$ given the following assumptions

AS1: $\exists p', q' \in J$ such that $|f(1, p'q')|$ is a strict maximum in Equation (A11),

AS2: $\exists q_1 \in J$ such that $|f(1, q_1)| > |f(0, 1 - q_1)|$ in Equation (A14),

AS3: $\left|f\left(0, \frac{1-r}{m}\right)\right|$ is never a strict maximum in Equation (A15),

AS4: $\sup\{\tilde{\mathcal{P}}\} < \frac{1}{2}$.

All that is left to prove the theorem is to show that violating any of these assumptions also violates some property. First assume that (A1), (A2) and (A3) are true but (A4) is violated, i.e., $\sup\{\tilde{\mathcal{P}}\} = \lim_{n \rightarrow \infty} p_n = \frac{1}{2}$ for $p_n \in h_{\frac{1}{2}}$. Let $p'_n = 1 - p_n, q' = 1 - q$ and let $g = \{1 - p_i\}_i$ be the sequence of ordered elements in $\tilde{\mathcal{P}}$ such that $p_i < p_{i+1}$. Then by Equation (A17), for all $p'_n \in g$, there exists $q' \in (p'_n, 1)$ such that

$$f(p'_n, q') = k_{p'_n} (1 - p'_n) \log\left(\frac{1 - p'_n}{1 - q'}\right).$$

By Lemma A2, for a fixed $p_n^* \in g$, there exists $\delta' > 0$ such that

$$f(p_n^*, q) = k_{p_n^*} (1 - p_n^*) \log\left(\frac{1 - p_n^*}{1 - q}\right), \quad \text{for all } q \in I_{\delta'}(q').$$

Since this holds for all $p_n^* \in g$, we have

$$f(p'_n, q) = k_{p'_n} (1 - p'_n) \log\left(\frac{1 - p'_n}{1 - q}\right), \quad \text{for all } (p'_n, q) \in g \times I_{\delta'}(q').$$

Similarly to Equation (A18), using the sequence g instead of the sequence $h_{\frac{1}{2}}$, this result can be extended to $J \times J$, i.e.,

$$f(p, q) = k(1 - p) \log\left(\frac{1 - p}{1 - q}\right), \quad \text{for all } (p, q) \in (J \times J). \tag{A19}$$

Applying Property II.b to $l = m = 2, P^2 = (p, 1 - p), Q^2 = (q, 1 - q)$ with $q \neq p$ yields

$$D(P^2 * U^2, Q^2 * U^2) = \frac{1}{2} D(P^2, Q^2).$$

Further, by Property III

$$\max\left\{\left|f\left(\frac{p}{2}, \frac{q}{2}\right)\right|, \left|f\left(\frac{1-p}{2}, \frac{1-q}{2}\right)\right|\right\} = \frac{1}{2} \max\{|f(p, q)|, |f(1-p, 1-q)|\}.$$

However this contradicts Equation (A19) since

$$\begin{aligned} \max\left\{\left|k\left(1-\frac{p}{2}\right)\log\left(\frac{1-\frac{p}{2}}{1-\frac{q}{2}}\right)\right|, \left|k\left(1-\frac{1-p}{2}\right)\log\left(\frac{1-\frac{1-p}{2}}{1-\frac{1-q}{2}}\right)\right|\right\} \\ \neq \frac{1}{2} \max\left\{\left|k(1-p)\log\left(\frac{1-p}{1-q}\right)\right|, \left|kp\log\left(\frac{p}{q}\right)\right|\right\}. \end{aligned}$$

Next we assume that (AS1) and (AS2) are true but (AS3) is violated, meaning that there exists $(m', r') \in \mathbb{N}_2 \times I_\delta(q_1)$ such that $|f(0, \frac{1-r'}{m'})|$ is a strict maximum in Equation (A15), i.e.,

$$f\left(0, \frac{1-r'}{m'}\right) = \frac{ak_1}{m'} \log\left(\frac{1}{r'}\right),$$

for some $a \in \{-1, 1\}$. For some $q' = \frac{1-r'}{m'} \in I_{\frac{\delta}{m'}}\left(\frac{1-q_1}{m'}\right)$, we have

$$f(0, q') = \frac{ak_1}{m'} \log\left(\frac{1}{1-m'q'}\right).$$

By Lemma A2, there exists $\delta' > 0$ such that

$$f(0, q) = \frac{ak_1}{m'} \log\left(\frac{1}{1-m'q}\right), \quad \text{for all } q \in I_{\delta'}(q').$$

By Corollary A1, we can extend $f(0, q)$ to $(0, \frac{1}{m'})$, i.e.,

$$f(0, q) = \frac{ak_1}{m'} \log\left(\frac{1}{1-m'q}\right), \quad \text{for all } q \in \left(0, \frac{1}{m'}\right).$$

However, this implies that f is discontinuous and violates Property III since

$$\lim_{q \rightarrow \frac{1}{m'}} f(0, q) = \lim_{q \rightarrow \frac{1}{m'}} \frac{ak_1}{m'} \log\left(\frac{1}{1-m'q}\right) = \pm\infty.$$

We now assume that AS1 is true but AS2 is violated, i.e.,

$$|f(1, q)| \leq |f(0, 1-q)|, \quad \text{for all } q \in J.$$

For some $p', q' \in J$ and $\delta > 0$, by Equation (A11) and by Equation (A13)

$$\left|k_1 \log\left(\frac{1}{pq}\right)\right| = |f(0, 1-p)| + |f(0, 1-q)|, \quad \text{for all } pq \in I_\delta(p'q'). \tag{A20}$$

Let $q_1, q_2 \in J$ and let $p'q_1, p'q_2 \in I_{\delta'}(p'q')$, then

$$\begin{aligned} F(q_1) &= \left|k_1 \log\left(\frac{1}{q_1}\right)\right| - |f(0, 1-q_1)| = |f(0, 1-p')| - \left|k_1 \log\left(\frac{1}{p'}\right)\right| \\ &= \left|k_1 \log\left(\frac{1}{q_2}\right)\right| - |f(0, 1-q_2)| = F(q_2). \end{aligned}$$

Therefore, $F(q) = d$ is constant and

$$|f(0, 1 - q)| = \left| k_1 \log\left(\frac{1}{q}\right) \right| + d, \quad \text{for all } q \in I_{\frac{\delta}{p'}}(q').$$

Plugging this back into Equation (A20) we see that $d = 0$, and for some $a \in \{-1, 1\}$, there exists $q^* \in I_{\frac{\delta}{p'}}(q')$ such that

$$f(0, q^*) = ak_1 \log\left(\frac{1}{1 - q^*}\right).$$

By Lemma A2, there exists $\delta' > 0$ such that

$$f(0, q) = ak_1 \log\left(\frac{1}{1 - q}\right), \quad \text{for all } q \in I_{\delta'}(q^*),$$

and by Corollary A1

$$f(0, q) = ak_1 \log\left(\frac{1}{1 - q}\right), \quad \text{for all } q \in J.$$

However this is a contradiction since by Equation (A13), for $k_1 \neq 0$ and $x < \frac{1}{2}$, we have

$$|f(1, x)| = \left| k_1 \log\left(\frac{1}{x}\right) \right| > \left| k_1 \log\left(\frac{1}{1 - x}\right) \right| = |f(0, 1 - x)|.$$

Note that $k_1 = 0$ violates Property I since for any $q \in J$ and $Q^2 = (q, 1 - q) \neq V^2$, we have

$$D(V^2, Q^2) = \max\left\{ \left| k_1 \log\left(\frac{1}{q}\right) \right|, \left| k_1 \log\left(\frac{1}{1 - q}\right) \right| \right\} = 0.$$

Finally, if AS1 is violated then there does not exist $p', q' \in J$ such that $|f(1, p'q')|$ is a strict maximum in Equation (A11). Hence, for all $p, q \in J$, there exists $x' \in \{p(1 - q), q(1 - p), (1 - p)(1 - q)\}$ such that

$$|f(0, x')| = \max\{|f(1, p)|, |f(0, 1 - p)|\} + \max\{|f(1, q)|, |f(0, 1 - q)|\}.$$

If $|f(0, x')|$ is a strict maximum, then by Lemma A2 there exists $\delta > 0$ such that

$$|f(0, x)| = \max\{|f(1, p)|, |f(0, 1 - p)|\} + \max\{|f(1, q)|, |f(0, 1 - q)|\}, \quad (A21)$$

for all $x \in I_{\delta}(x')$. If there does not exist $x' \in \{p(1 - q), q(1 - p), (1 - p)(1 - q)\}$ such that $|f(0, x')|$ is a strict maximum, then there must exist $x \in \{p(1 - q), q(1 - p), (1 - p)(1 - q)\}$ such that Equation (A21) holds for all $x \in J$. In both cases, by Lemma A3 there exists $c, d \in \mathbb{R}$ such that

$$f(0, x) = c \log(x) + d, \quad \text{for all } x \in J.$$

However, this implies that f is discontinuous since

$$\lim_{x \rightarrow 0} f(0, x) = \lim_{x \rightarrow 0} c \log(x) + d = \pm\infty,$$

even though $f(0, 0) = 0$ by Lemma A4. \square

Appendix C. Comparison between ID and EMD

Since the different information measures satisfy different properties, the distinctions that exist in a given system may be different depending on the information measure used to compute integrated information. Here, using the same system used in Figure 5, we

provide an example where the cause purview with maximum integrated information is larger when using the EMD measure (Figure A1a) when compared to the same mechanism when using the ID measure (Figure A1b).

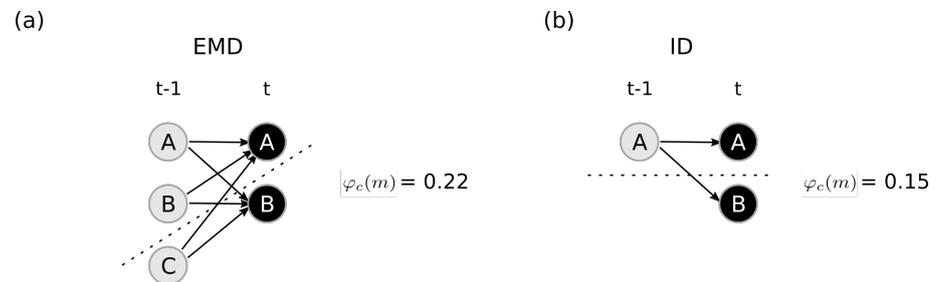


Figure A1. Comparison between earth mover's distance (EMD) and ID. Using the same system S used in Figure 4a, we find the cause purview with maximum integrated information for the mechanism $M = \{A, B\}$ in state $m = \{\uparrow, \uparrow\}$, which is larger when using the EMD measure (a) when compared to the ID measure (b). The integrated information when using the EMD measure is also larger than the ID measure.

References

- Albantakis, L. Integrated information theory. In *Beyond Neural Correlates of Consciousness*; Overgaard, M., Mogensen, J., Kirkeby-Hinrup, A., Eds.; Routledge: London, UK, 2020; pp. 87–103.
- 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](#)]
- 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](#)]
- Balduzzi, D.; Tononi, G. Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLoS Comput. Biol.* **2008**, *4*, e1000091. [[CrossRef](#)] [[PubMed](#)]
- Barbosa, L.; Marshall, W.; Streipert, S.; Albantakis, L.; Tononi, G. A measure for intrinsic information. *Sci. Rep.* **2020**, *10*, 1–9. [[CrossRef](#)] [[PubMed](#)]
- Tononi, G. The Integrated Information Theory of Consciousness. In *The Blackwell Companion to Consciousness*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2017; pp. 243–256.
- 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. [[CrossRef](#)] [[PubMed](#)]
- Tononi, G. Consciousness as integrated information: A provisional manifesto. *Biol. Bull.* **2008**, *215*, 216–242. [[CrossRef](#)] [[PubMed](#)]
- Marshall, W.; Albantakis, L.; Tononi, G. Black-boxing and cause-effect power. *PLoS Comput. Biol.* **2018**, *14*, e1006114. [[CrossRef](#)] [[PubMed](#)]
- Haun, A.; Tononi, G. Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy* **2019**, *21*, 1160. [[CrossRef](#)]
- 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](#)]
- Albantakis, L.; Tononi, G. Causal Composition: Structural Differences among Dynamically Equivalent Systems. *Entropy* **2019**, *21*, 989. [[CrossRef](#)]
- Marshall, W.; Gomez-Ramirez, J.; Tononi, G. Integrated Information and State Differentiation. *Conscious. Res.* **2016**, *7*, 926. [[CrossRef](#)] [[PubMed](#)]
- 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](#)]
- Csiszár, I. Axiomatic Characterizations of Information Measures. *Entropy* **2008**, *10*, 261–273. [[CrossRef](#)]
- Tononi, G. An information integration theory of consciousness. *BMC Neurosci.* **2004**, *5*, 42. [[CrossRef](#)] [[PubMed](#)]
- Kyumin, M. Exclusion and Underdetermined Qualia. *Entropy* **2019**, *21*, 405.
- Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.
- Janzing, D.; Balduzzi, D.; Grosse-Wentrup, M.; Schölkopf, B. Quantifying causal influences. *Ann. Stat.* **2013**, *41*, 2324–2358. [[CrossRef](#)]
- Ebanks, B.; Sahoo, P.; Sander, W. *Characterizations of Information Measures*; World Scientific: Singapore, 1998.
- Krantz, S.G.; Parks, H.R. *A Primer of Real Analytic Functions*, 2nd ed.; Birkhäuser Advanced Texts Basler Lehrbücher; Birkhäuser: Basel, Switzerland, 2002.
- Aczél, J. *Lectures on Functional Equations and Their Applications*; Dover Publications: Mineola, NY, USA, 2006.