Mind and Consciousness II
Integrated Information Theory (IIT) and Global Neural Workspace (GNW) with Stochastic Hopfield Networks

J N Tavares *

February 5, 2026

Abstract

This study follows on from a previous one [JNT2025a] where a new computational model was developed for the so-called Global Neuronal Workspace (GNW) by Dehane & Changeux.

This model integrates specialized modules and a central Workspace, implemented as stochastic Hopfield networks, interacting dynamically through feedforward and feedback connectivity. The main focus was the dynamics of neuronal ignition in the Workspace. We investigated how crucial factors, such as the stochasticity inherent in neuronal processing, the storage capacity ($\mathcal{C} = p/N$) of the subsystems, and the coupling strength between them, influence the phase transition that underlies this ignition.

*jntavar@fc.up.pt; Homepage: jntavar; Homepage: Casa das Ciências.

We demonstrate how external fields, orchestrated by mechanisms of modular competition and top-down feedback, allow for pattern-driven ignition, where patterns memorized in the modules and Workspace act as active guides for the dynamics.

In this article, we will use Giulio Tononi's Integrated Information Theory (IIT) to quantify the global coherence of the GNW system, defining the measure of integrated information, Φ, through entropy transfers (TE and RTE).

We hypothesize that consciousness emerges as a globally ordered state of the Workspace, actively orchestrated by the dynamics of the GNW, corresponding to a phase transition to the ordered phase. The results provide insights into the mechanisms underlying the emergence of consciousness, linking microscopic behavior and statistical mechanics to higher-order phenomena and the potential for an integrated experience.

Contents

1 Introduction. Integrated Information Theory (IIT): A new approach

The Integrated Information Theory (IIT), proposed by Giulio Tononi [T2004] [T2014] [T2015] [K2019] [ET2000] [TK2015], aims to explain what consciousness is, why certain systems are conscious and others are not, and how conscious experience relates to the physical world.

The central idea of Giulio Tononi's IIT theory and Gerald Edelman (Dynamic Core Hypothesis) to quantify consciousness through Φ, it can be summarized in a simple phrase: "*Consciousness is Integrated Information*".[1]

Both, with different approaches, propose that consciousness is fundamentally linked to the quantity and quality of information that a system can generate and integrate. Quantity refers to the number of differentiated states that a system can assume – the more different states a system can distinguish, the more "information" it has. Quality (or Integration) refers to how interdependent the information generated by the different parts of the system is. Information is integrated if it cannot be decomposed into independent parts without significant loss.

Φ is a measure that attempts to capture both aspects: quantity and integration of information. Φ quantifies how much the internal causal structure of a system restricts the number of possible states in which the system can be. A system with high Φ has a rich and interdependent causal structure.

Φ measures how much a system is "*more than the sum of its parts*". It is the information that is generated by the organization and interaction of the parts, and that cannot be explained by the parts individually. In simpler terms: a conscious system must be able to:

1. Generate a lot of information, that is, distinguish many different states.
2. Integrate this information forming a unified experience, that is, one that cannot be decomposed into independent parts.

Φ attempts to quantify this combination of quantity and integration, providing a measure of consciousness. Systems with high Φ are more conscious than systems with low Φ.

[1] Giulio Tononi is a neuroscientist and psychiatrist who holds the David P. White Chair in Sleep Medicine, as well as a Distinguished Chair in the Science of Consciousness, at the University of Wisconsin-Madison. He is best known for his Integrated Information Theory (IIT), a mathematical theory of consciousness, which he proposed in 2004. Gerald Edelman (1929 - 2014) was a physician, molecular biologist, and physical chemist born in New York, USA. He was awarded, along with Rodney Porter, the 1972 Nobel Prize in Physiology or Medicine for research on the structure and chemical nature of antibodies. He then shifted the focus of his research to developmental biology and neurology. In 1978 he presented the theory of Neuronal Darwinism, based on the idea of ??the plasticity of neural networks in response to the external environment.

In this article, we will take a new approach to the notions of Integrated Information and Φ, using the concept of Entropy Transfer with reduced interaction, which can be seen as a way to operationalize the concepts of Tononi and Edelman. Here's how it fits:

1. Information and Differentiation: Entropy Transfer (TE) measures the information that one subsystem, $\mathbb{A}$, provides about another, $\mathbb{B}$. If the entropy transfer from $\mathbb{A}$ to $\mathbb{B}$, denoted by $\mathsf{TE}(\mathbb{A} \to \mathbb{B})$, is high, this means that the state of $\mathbb{A}$ helps predict the state of $\mathbb{B}$, indicating a causal informational relationship.

2. In the context of IIT and Dynamic Core Hypothesis, a system's ability to generate information is linked to its ability to distinguish between many different states. By using TE, we are capturing how differentiated the states of $\mathbb{A}$ are from $\mathbb{B}$.

3. Integration and Irreducibility: reducing the interactions between $\mathbb{A}$ and $\mathbb{B}$ is the key point. By reducing the causal influence between subsystems, we are simulating what would happen if the system were less integrated. The difference in Entropy Transfer between the system with intact interactions and the system with reduced interactions represents the loss of information due to the lack of integration. This loss of information is analogous to the irreducibility of the system. In other words, how much the system is more than the sum of its parts. item When finding the Minimum Information Partition MIP, we are looking for the way to cut the system that minimizes the loss of information, that is, that preserves as much integration as possible.

4. Φ, calculated based on the reduction of interactions and Entropy Transfer, becomes a measure of causal irreducibility – it tells us how much the subsystems $\mathbb{A}$ and $\mathbb{B}$ are causally interconnected, and how much their joint activity contributes to the overall dynamics of the system. A system with high Φ is a system where the parts exert a strong causal influence on each other, and where the dynamics of the system as a whole cannot be easily explained by the activity of the isolated parts.

In summary: the approach proposed here, using Entropy Transfer and reduction of $\mathbb{A}/\mathbb{B}$ interactions, provides a way to:

- Measure the causal influence between subsystems.
- Quantify the loss of information (or causal irreducibility) when interactions are reduced.
- Define Φ as a measure of the system's integration and causal irreducibility, in line with the ideas of Tononi and Edelman.

This approach is, in our opinion, interesting because it can also be applied to different types of systems: artificial neural networks (as in the GNW model), brain networks (using fMRI or EEG data), and complex systems in general.

After this preamble, we will now detail all the theoretical aspects mentioned above.

Axioms.

The IIT begins by stating phenomenological axioms, which are considered self-evident truths about conscious experience. From these axioms, IIT derives postulates that specify how consciousness should be measured in physical systems.

- Axiom 0 – Intrinsic Existence. Consciousness exists. Every conscious experience is real and defined.

  If we read this text, we are conscious. This experience is real and not an illusion. The very experience of being conscious is proof of this axiom. It's like saying "light exists", even without understanding what light is or how it works.

- Composition – Consciousness is structured. Every conscious experience is composed of distinct elements that combine in specific ways. ... For example, the experience of seeing a red apple involves the combination of several characteristics: the red color, the round shape, the smooth texture, the sweet smell. Each of these characteristics is an element of the experience, and the combination of all creates a unique experience.

  When we look at a self-portrait by Rembrandt, we see that it is composed of different colors, brushstrokes, shapes, and textures. The combination of these elements creates an image with a meaning that is greater than the sum of its parts.

- Information – Consciousness is specific. Each conscious experience is different from other possible experiences. Each experience has a specific content that distinguishes it from all other experiences.

  The experience of seeing a red apple is different from the experience of listening to music or feeling the wind on your face. Each experience has a unique informational content that makes it specific. Comparing a symphony to a rock song, we conclude that both are music, but each has a unique harmonic, melodic, and rhythmic structure. IIT says that each piece of music generates a different conscious experience, because each generates a unique informational content.

- Integration – Consciousness is unified. The system must be integrated, that is, its elements must affect each other in a way that is not simply reducible to the sum of their individual effects.

The elements of an experience cannot be reduced to independent and separate parts. Conscious experience is an indivisible whole.

The experience of seeing a red apple is not simply the sum of the experience of seeing the color red plus the experience of seeing the round shape. There is an intrinsic relationship between these elements that creates a unified experience. If we separated the color from the shape, we would no longer have the experience of seeing a red apple.

- Exclusion – Consciousness is defined and exclusive. The system must be defined, with a specific set of elements and connections that define its boundaries. Each conscious experience exists with a certain granularity, in a defined time and space. Consciousness excludes other experiences that are not part of it.

  At the moment we read this explanation, we are aware of the content of the words, but we are not aware of other things that may be occurring (e.g., the pressure of the chair on the back, the distant sound of a car). Conscious experience is defined and exclusive, excluding other information that is not relevant to the current focus.

  In Baars' theater metaphor, let's imagine a spotlight illuminating only a part of the stage. Only actors and illuminated objects are conscious, while the rest of the stage remains in darkness (unconscious).

Postulates.

Postulates are translations of axioms into the physical world.

- Intrinsic Existence. A system must exist intrinsically, that is, it must be able to influence itself.

- Composition. The system must be structured into a set of elements that influence each other. The postulate of composition states that conscious experience is structured. That is, it is composed of distinct elements that combine in specific ways.

  To understand consciousness in a system, it is necessary to identify these elements and their causal relationships. In the context of IIT, these elements are usually neurons or groups of neurons, but they can be other types of physical components capable of influencing each other. The postulate states not only that consciousness is composite, but also that it is necessary to decompose the system into smaller parts in order to understand its causal structure. This means that if a group of neurons can be subdivided into smaller subgroups that still maintain a significant causal influence, then these subgroups should be considered as separate elements.

- Causal Connections. The most important thing for IIT is the causal relationships between the parts, as it ensures that each part influences the others. Fig. 1.

- Information. The information postulate states that each conscious experience is specific, that is, it has a unique informational content that distinguishes it from other possible experiences.

  For a system to be conscious, it cannot simply exist statically. It must have a dynamic that generates information. The system must be integrated, that is, its elements must affect each other in a way that is not simply reducible to the sum of their individual effects.

- Significant Output. The system's output must have a value, that is, it must be significant to the system in terms of its internal causal relationships. The mere existence of an output is not enough; the output must influence the system to create a complex and integrated causal structure.

- Generation of Differences. Information is created when a system makes a distinction, that is, when it discriminates between different possibilities. The more complex and subtle the distinctions a system can make, the more information it generates and the more conscious it is.

Figure 1: A WEB OF CAUSAL RELATIONS: to every physical system that obeys all phenomenological axioms, the theory of integrated information associates an intrinsic structure of cause and effect, illustrated by a labyrinthine web. The central identity of IIT affirms that the consciousness of being this system, in this state, is defined by the set of causal relations that compose this structure. Figure by [K2019].

An inspiring example. A jazz band.

Let's imagine the system is a jazz band composed of several musicians. Each individual musician (e.g., pianist, bassist, percussionist, saxophonist, etc.) is an element of the system. The status of each musician is defined by the melody, harmony, rhythm, pitch, expression, etc., of what they are playing.

The musicians interact with each other through listening and responding to each other's actions. The drummer establishes the basic rhythm, influencing all the other musicians. The pianist and saxophonist exchange melodies and harmonies, influencing the notes each plays. The bassist provides a baseline that supports the melodies and harmonies.

Let's now analyze this example from the perspective of IIT. What are the causal interactions?

1. INTRINSIC CAUSALITY. For a jazz band to be considered an integrated and

conscious system (in a broad sense), the interaction between the musicians must generate an irreducible causal structure. This means that the music they play together must be more than the sum of each musician's individual parts. Improvisation and real-time response to each other's actions create something new and emergent.

2. Defining the Disorder. Disturbing a part of the band (or a set of parts) is to "fix" the state of the disturbed element. In the example, silencing an instrument corresponds to forcing the element to always be in an inactive state.

3. Levels of Integration.

   Low Integration: If each musician plays their part individually, without listening to or responding to the others, then the band would have a low causal capacity. The music would be just an overlay of individual parts, without real interaction or coherence.

   High Integration: If the musicians listened attentively to each other and responded to their actions creatively and adaptively, the band would have a high causal capacity. The music would be an emergent creation from the interaction between the musicians.

4. Partitions. A bad partition of the band would be to divide it by separating all the musicians by instrument type (e.g., separating all the brass, all the strings, etc.). This would drastically disrupt the music, as it would eliminate the interaction between the different types of instruments. Partitioning the band by separating the rhythmic section (drums and bass) from the melodic instruments (piano and saxophone) could be less destructive, as it would still allow some interaction within each section. However, since rhythm influences melody, and vice versa, this partition would also cause a loss of information.

5. Calculating Φ (Conceptually). Calculating Φ would involve quantifying the difference between the causal capacity of the band as a whole and the sum of the causal capacities of each musician individually.

The IIT suggests that the "mind" of the jazz band resides in its ability to coordinate the activities of its members to create music that is more than the sum of its parts.

The metric Φ attempts to quantify how well this coordination is achieved, with the MIP identifying the parts of the system that are most intrinsically linked and influence the organization of information at any given moment.

■

The preceding axioms and postulates, illustrated with the (superficial, but motivating) metaphor of the jazz band, introduce several concepts that require rigorous formalization.

IIT can be seen as a theory of conscious complexity, that is, a theory that attempts to quantify and explain the specific complexity associated with conscious experience.

It proposes that consciousness does not simply arise from the structural complexity of a system, but rather from its ability to generate and integrate information in an irreducible way.

IIT is not limited to measuring the structural complexity of a system (e.g., the number of connections between neurons). Instead, IIT focuses on the causal complexity of the system, that is, the capacity of each part of the system to influence and be influenced by other parts.

IIT proposes that consciousness is related to the amount of information integrated into the system. Therefore, our goal is to define a measure that captures both the amount of information that the mechanisms generate and how integrated that information is.

IIT proposes that consciousness emerges when a system reaches a certain level of causal complexity, that is, when it is capable of generating a significant amount of integrated information. Consciousness is not simply a property added to individual components, but rather a characteristic that arises from their organization and interaction. In this sense, consciousness is often seen as an emergent property of the brain.

One last aspect: IIT also has several important philosophical implications:

- Panpsychism: IIT implies that consciousness can exist to varying degrees in all systems that possess an integrated causal structure.
- Subjectivity: IIT recognizes the intrinsic and subjective nature of conscious experience. Each system has its own unique perspective on the world.
- Hard problem of consciousness: IIT offers a solution to the philosophical challenge of explaining how and why physical brain processes give rise to subjective and conscious experience, also known as *qualia*, suggesting that *consciousness is a fundamental property of the organization of matter*.

Let's now propose a possible formalism for IIT, based on Information Theory (by Shannon), which is summarized in Appendix **??**.

2 Measuring Integrated Information with Entropy Transfer.

In this section we will present a detailed mathematical formalization of the concept of Integrated Information, using Information Theory (Shannon's), in particular, the notion of Transfer Entropy (TE), for its quantification. Appendix 9 contains a review of these ideas.

The goal is to provide a rigorous basis for the application of these concepts, in the context of Integrated Information Theory (IIT) and, subsequently, in the GNWmodel.

For now, the exposition will be purely abstract and general. Thus, let us consider a system $\mathbb{S}$, composed of N elements, represented by $i = 1, 2, \cdots, N$ (e.g., neurons), to each of which is associated a random variable S_i, which, at each instant t, can only assume two values: $s_i(t) \in \{-1, 1\}$.

The state of the system $\mathbb{S}$, at time t, is represented by the binary vector

$$\mathbf{s}(t) = (s_1(t), s_2(t), \cdots, s_N(t)) \in \{-1, 1\}^N$$

The state space $\Sigma = \Sigma_{\mathbb{S}}$ is the set of all possible configurations (or states) of the system. For binary elements, $|\Sigma| = 2^N$. The probability distribution over the states of the system is denoted by $P(\mathbf{s})$.

A mechanism $\mathbb{A}$ is a subset of va's in $\mathbb{S}$:

$$\mathbb{A} \subset \mathbb{S} \tag{1}$$

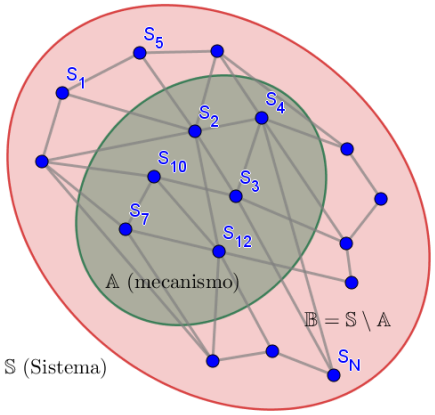


Figure 2: A mechanism $\mathbb{A}$ is a subset of va's in $\mathbb{S}$.

Let us consider a mechanism $\mathbb{A} \subset \mathbb{S}$, and its complement $\mathbb{B} = \mathbb{S} \setminus \mathbb{A}$. We want to:

1. quantify how much the knowledge of the current state of $\mathbb{A}$, $\mathbb{A}_t$ (in the role of cause) influences the future state of $\mathbb{B}$, $\mathbb{B}_{t+1}$ (in the role of effects),
2. quantify how much the knowledge of the current state of $\mathbb{A}$, $\mathbb{A}_t$ (in the role of effect) restricts the past possibilities of $\mathbb{B}$, $\mathbb{B}_{t-1}$ (in the role of causes of $\mathbb{A}_t$), and
3. quantify the two previous points with the roles of $\mathbb{A}$ and $\mathbb{B}$ reversed, under the following two hypotheses:
    A. when the system has all the original $\mathbb{A}/\mathbb{B}$ interactions active (Fig 5 on the left), which we represent by $\mathbb{S} = \mathbb{A} \uplus \mathbb{B}$, and

B. when the system has the interactions $\mathbb{A}/\mathbb{B}$ reduced by a factor $\alpha : 0 \leq \alpha \leq 1$:

$$J_{\mathbb{A}\mathbb{B}} \longrightarrow \alpha \cdot J_{\mathbb{A}\mathbb{B}}$$

(Fig 5 on the right), which we represent by $\mathbb{S}_\alpha = \mathbb{A} \uplus_\alpha \mathbb{B}$. The smaller is $\alpha \in [0, 1]$, the greater is the reduction in the intensity of the interactions $\mathbb{A}/\mathbb{B}$.

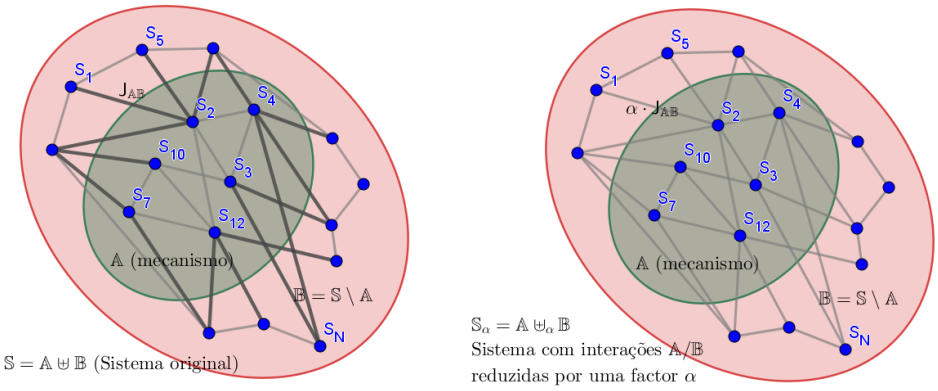


Figure 3: Original system and system with interactions $\mathbb{A}/\mathbb{B}$ reduced by a factor α.

The current state of a mechanism can be seen as a source of information about the rest of the system. If the mechanism has a strong causal influence, then:

(i). Its current state should restrict the set of possible past states of the system. Knowing the current state of the mechanism should reduce our uncertainty about what happened in the system in the past.

(ii). Its current state should restrict the set of possible future states of the system. Knowing the current state of the mechanism should reduce our uncertainty about what will happen in the system in the future.

Here is an illustrative metaphor. Consider a detective investigating a crime. The detective observes a series of evidence in the present (e.g., fingerprints, testimonies). This evidence (the current state of the mechanism) informs the detective about what happened in the past (e.g., who committed the crime, how the crime was committed). The fact that the detective is observing the evidence does not change what happened in the past, but rather restricts the set of possible scenarios and allows him to infer the most probable cause.

Here, the "capacity to influence" will be quantified by the ENTROPY TRANSFER, as a measure of uncertainty or average information – a decrease in entropy

translates into a reduction of uncertainty in the prediction (past and future) or, equivalently, into a gain of information. According to Shannon's information theory, the dynamics of the GNW system tend towards a decrease in entropy, which translates directly into a reduction in the uncertainty of the network state. This reduction in entropy corresponds to a gain of information, allowing the system to increase the reliability of the prediction on the sensory inputs (inference) and stabilize the internal representation (memory).

Causal influence is fundamental to Integrated Information Theory (IIT), because IIT presupposes that consciousness is related to the amount of information a system can integrate. A system can only integrate information if its components have the ability to influence each other, whether in the past or the future.

Note: The common formulation of causality assumes that the cause precedes the effect, which is why the phrase "influencing the past" may seem contradictory. However, in IIT the concept of causal ability is used in a specific way that does not violate the temporal relationship between cause and effect.

In fact, IIT does not propose a violation of temporal causality – it does not say that the current state of the mechanism changes or alters what has already happened in the past. The past is fixed and immutable. Use the term "cause" in a specific sense, which relates to the ability to restrict the set of possible past and future states, rather than a direct action on the past. What we are measuring is the intrinsic causal relevance of a mechanism within the system, and not its ability to alter the past.

We will now use the concept of Entropy Transfer (TE), which is explained in the appendix 9, to quantify the causal influence of $\mathbb{A}$.

If we completely eliminate the interactions between $\mathbb{A}$ and $\mathbb{B}$, separating them completely (that is, if $\alpha = 0$):

$$\mathsf{J}_{\mathbb{AB}} \longrightarrow 0 \cdot \mathsf{J}_{\mathbb{AB}} = 0$$

then $\mathbb{A}$ and $\mathbb{B}$ become, by definition, independent, and the entropy transfers, from $\mathbb{A}$ to $\mathbb{B}$ and from $\mathbb{B}$ to $\mathbb{A}$, would be equal to zero.

We therefore need to reduce the interactions between $\mathbb{A}$ and $\mathbb{B}$, by a factor $0 < \alpha \leq 1$, instead of eliminating them completely. This allows some causal influence to persist, but on a reduced scale.

We now measure the entropy transfer in the system, in the situations described above.

- $\mathsf{TE}_{\alpha}(\mathbb{A}_t \rightarrow \mathbb{B}_{t+1})$, for $0 < \alpha \leq 1$.
- $\mathsf{RTE}_{\alpha}(\mathbb{A}_t \rightarrow \mathbb{B}_{t-1})$, for $0 < \alpha \leq 1$.

where:

$$\begin{aligned} \mathsf{TE}(\mathbb{A}_t \to \mathbb{B}_{t+1}) &= \mathsf{Ent}(\mathbb{B}_{t+1}|\mathbb{B}_t) - \mathsf{Ent}(\mathbb{B}_{t+1}|\mathbb{B}_t, \mathbb{A}_t) \\ &= \mathsf{Inf}(\mathbb{B}_{t+1}; \mathbb{A}_t|\mathbb{B}_t) \\ &= \sum P(\mathsf{b}_{t+1}, \mathsf{a}_t, \mathsf{b}_t) \log_2 \frac{P(\mathsf{b}_{t+1}|\mathsf{a}_t, \mathsf{b}_t)}{P(\mathsf{b}_{t+1}|\mathsf{b}_t)} \end{aligned} \tag{2}$$

Remember that $\mathsf{Ent}(\mathbb{B}_{t+1}|\mathbb{B}_t)$ is the reduction in uncertainty (information gain) about $\mathbb{B}_{t+1}$, caused by the knowledge of $\mathbb{B}_t$ and $\mathsf{Ent}(\mathbb{B}_{t+1}|\mathbb{B}_t, \mathbb{A}_t)$ is the reduction in uncertainty (information gain) in $\mathbb{B}_{t+1}$, caused by the knowledge of $\mathbb{B}_t$ and $\mathbb{A}_t$.

Therefore, $\mathsf{TE}_\alpha(\mathbb{A}_t \to \mathbb{B}_{t+1})$ measures the decrease in the degree of uncertainty about $\mathbb{B}_{t+1}$, caused only by the current value of $\mathbb{B}_t$, in addition to the decrease in the degree of uncertainty caused by both $\mathbb{A}_t$ and $\mathbb{B}_t$. If $\mathsf{TE}(\mathbb{A}_t \to \mathbb{B}_{t+1}) > 0$, then knowledge of $\mathbb{A}_t$ reduces the uncertainty about $\mathbb{B}_{t+1}$, in addition to what is already provided by knowledge of $\mathbb{B}_t$, suggesting that $\mathbb{A}_t$ contains relevant information about the future effects of $\mathbb{B}$.

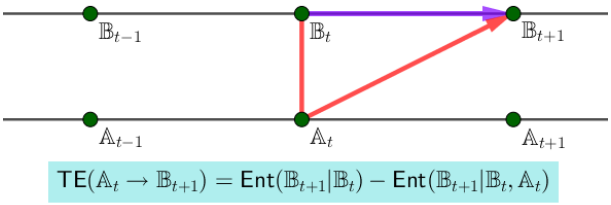


Figure 4: Entropy transfer of $\mathbb{A}$ (as cause) for $\mathbb{B}$: $\mathsf{TE}(\mathbb{A}_t \to \mathbb{B}_{t+1})$

RTE stands for Reversed Transfer Entropy. It measures how much knowledge of the current state $\mathbb{B}_t$ helps to restrict (or retro-predict) the set of possible past causes $\mathbb{B}_{t-1}$, in addition to the retroprediction induced by both $\mathbb{A}_t$ and $\mathbb{B}_t$.

$$\begin{aligned} \mathsf{RTE}(\mathbb{A}_t \to \mathbb{B}_{t-1}) &= \mathsf{Ent}(\mathbb{B}_{t-1}|\mathbb{B}_t) - \mathsf{Ent}(\mathbb{B}_{t-1}|\mathbb{B}_t, \mathbb{A}_t) \\ &= \mathsf{Inf}(\mathbb{A}_t; \mathbb{B}_{t-1}|\mathbb{B}_t) \\ &= -\sum_{\mathsf{b}_t} P(\mathsf{b}_t) \sum_{\mathsf{b}_{t-1}} P(\mathsf{b}_{t-1}|\mathsf{b}_t) \log_2[P(\mathsf{b}_{t-1}|\mathsf{b}_t)] \\ &+ \sum_{\mathsf{a}_t, \mathsf{b}_t} P(\mathsf{a}_t, \mathsf{b}_t) \sum_{\mathsf{b}_{t-1}} P(\mathsf{b}_{t-1}|\mathsf{a}_t, \mathsf{b}_t) \log_2[\mathsf{b}_{t-1}|\mathsf{a}_t, \mathsf{b}_t)] \end{aligned} \tag{3}$$

Reversed Transfer Entropy can be expressed as a difference of conditional entropies:

$$\mathsf{RTE}(\mathbb{A}_t \to \mathbb{B}_{t-1}) = \mathsf{Ent}(\mathbb{B}_{t-1}|\mathbb{B}_t) - \mathsf{Ent}(\mathbb{B}_{t-1}|\mathbb{A}_t, \mathbb{B}_t) \tag{4}$$

where:

- $\mathsf{Ent}(\mathbb{B}_{t-1}|\mathbb{B}_t)$ is the conditional entropy of $\mathbb{B}$, at time $t-1$, given the knowledge of the state of $\mathbb{B}$ at time t.

  It is given by:

$$\mathsf{Ent}(\mathbb{B}_{t-1}|\mathbb{B}_t) = -\sum_{\mathsf{b}_t} P(\mathsf{b}_t) \sum_{\mathsf{b}_{t-1}} P(\mathsf{b}_{t-1}|\mathsf{b}_t) \log_2[P(\mathsf{b}_{t-1}|\mathsf{b}_t)] \tag{5}$$

  It represents the uncertainty we still have about the past state of $\mathbb{B}$ (at $t-1$) after observing its present state (at t). It quantifies how much the present state of $\mathbb{B}$ tells us about its past.

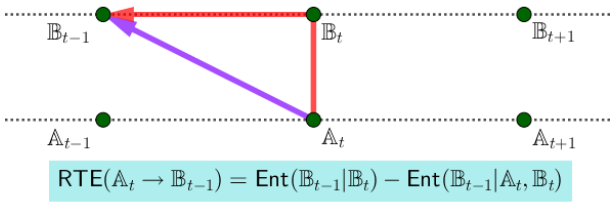
- $\mathsf{Ent}(\mathbb{B}_{t-1}|\mathbb{A}_t, \mathbb{B}_t)$ is the conditional entropy of $\mathbb{B}$, at time $t-1$, given the joint knowledge of the state of $\mathbb{A}$ at time t and the state of $\mathbb{B}$ at time t. It is given by:

$$\mathsf{Ent}(\mathbb{B}_{t-1}|\mathbb{A}_t, \mathbb{B}_t) = -\sum_{\mathsf{a}_t,\mathsf{b}_t} P(\mathsf{a}_t, \mathsf{b}_t) \sum_{\mathsf{b}_{t-1}} P(\mathsf{b}_{t-1}|\mathsf{a}_t, \mathsf{b}_t) \log_2[P(\mathsf{b}_{t-1}|\mathsf{a}_t, \mathsf{b}_t)] \tag{6}$$

Represents the remaining uncertainty about the past state of $\mathbb{B}$ after observing both the present state of $\mathbb{B}$ and the present state of $\mathbb{A}$. It quantifies how much the knowledge of $\mathbb{A}_t$ and $\mathbb{B}_t$ together tells us about the past of $\mathbb{B}$.A RTE is equal to the Mutual Information between $\mathbb{A}_t$ and $\mathbb{B}_{t-1}$ conditioned on $\mathbb{B}_t$:

$$\mathsf{RTE}(\mathbb{A}_t \to \mathbb{B}_{t-1}) = \mathsf{Inf}(\mathbb{A}_t; \mathbb{B}_{t-1}|\mathbb{B}_t)$$

This means that RTE quantifies how much information $\mathbb{A}_t$ shares with $\mathbb{B}_{t-1}$ that is not shared with $\mathbb{B}_t$. In other words, it measures how much $\mathbb{A}_t$ helps predict $\mathbb{B}_{t-1}$ given that we already know $\mathbb{B}_t$.



to $\mathbb{B}$: $\mathsf{RTE}(\mathbb{A}_t \to \mathbb{B}_{t-1})$

Figure 5: Reverse entropy transfer from $\mathbb{A}$ (as an effect) to $\mathbb{B}$: $\mathsf{RTE}(\mathbb{A}_t \to \mathbb{B}_{t-1})$

If $\mathsf{RTE}(\mathbb{A}_t \to \mathbb{B}_{t-1}) > 0$, then knowledge of $\mathbb{A}_t$ reduces the uncertainty about $\mathbb{B}_{t-1}$, in addition to what is already provided by knowledge of $\mathbb{B}_t$. This suggests that $\mathbb{A}_t$ contains relevant information about the past causes of B.

Knowledge of the present of $\mathbb{A}$ significantly restricts the possibilities for the past of $\mathbb{B}$, indicating a strong dependency or reverse causal constraint. RTE

quantifies the ability to "retro-predict" or infer about the past of a system based on its current state.

If $\mathsf{RTE}(\mathbb{A}_t \to \mathbb{B}_{t-1}) = 0$, then knowledge of $\mathbb{A}_t$ provides no additional information about $\mathbb{B}_{t-1}$ beyond what is already provided by knowledge of $\mathbb{B}_t$. This suggests that $\mathbb{A}_t$ is not causally related to the past of $\mathbb{B}$ (or that the relationship is too weak to be detected).

Again, the detective analogy: think of $\mathbb{B}_{t-1}$ as the crime to be solved, $\mathbb{A}_t$ as new evidence discovered in the present, and $\mathbb{B}_t$ as the prior knowledge the detective already has about the case. $\mathsf{Ent}(\mathbb{B}_{t-1}|\mathbb{B}_t)$ represents the uncertainty the detective still has about the crime before analyzing the new evidence. $\mathsf{Ent}(\mathbb{B}_{t-1}|\mathbb{A}_t, \mathbb{B}_t)$ represents the uncertainty the detective has about the crime after analyzing the new evidence. $\mathsf{RTE}(\mathbb{A}_t \to \mathbb{B}_{t-1})$ represents how much the new evidence helped the detective narrow down possible explanations for the crime. If RTEis high, it means the evidence was very useful in solving the case. With these definitions, we now construct the integrated information of $\mathbb{A}$:

$$\mathbb{II}_\alpha(\mathbb{A}_t) = \mathsf{TE}_\alpha(\mathbb{A}_t \to \mathbb{B}_{t+1}) + \mathsf{RTE}_\alpha(\mathbb{A}_t \to \mathbb{B}_{t-1}) \tag{7}$$

This formula quantifies how much $\mathbb{A}_t$ restricts its causes, $\mathbb{B}_{t-1}$, and influences its effects, $\mathbb{B}_{t+1}$. It depends on α, because it depends on how much the interactions $\mathbb{A}/\mathbb{B}$ are attenuated. When $\alpha = 0$, $\mathbb{II}_\alpha(\mathbb{A}_t) = 0$, as expected.

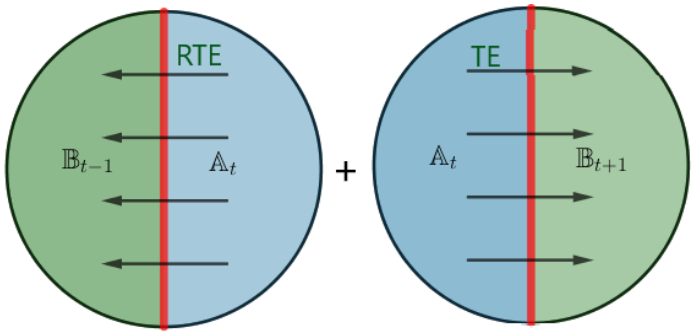


Figure 6: $\mathbb{II}_\alpha(\mathbb{A}_t) = \mathsf{TE}_\alpha(\mathbb{A}_t \to \mathbb{B}_{t+1}) + \mathsf{RTE}_\alpha(\mathbb{A}_t \to \mathbb{B}_{t-1})$

The formalization presented so far focuses only on the in $\mathbb{A}_t$. To complete the analysis, we also need to quantify the CAUSAL INFLUENCE of $\mathbb{B}$ on $\mathbb{A}$, when the interactions $\mathbb{A}/\mathbb{B}$ are attenuated by a factor α. The integrated information of $\mathbb{B}$:

$$\mathbb{II}_\alpha(\mathbb{B}_t) = \mathsf{TE}_\alpha(\mathbb{B}_t \to \mathbb{A}_{t+1}) + \mathsf{RTE}_\alpha(\mathbb{B}_t \to \mathbb{A}_{t-1}) \tag{8}$$

which quantifies how much $\mathbb{B}_t$ restricts its causes, $\mathbb{A}_{t-1}$, and influences its effects, $\mathbb{A}_{t+1}$.

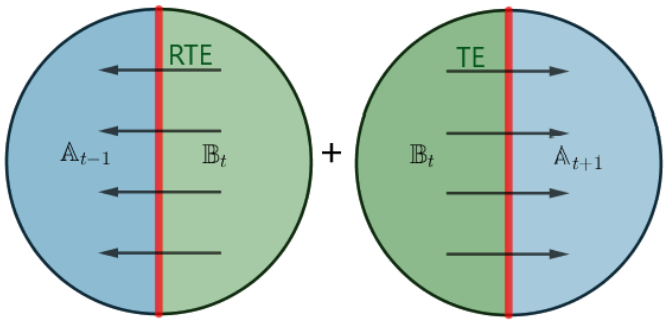


Figure 7: $\mathbb{III}_\alpha(\mathbb{B}_t) = \mathsf{TE}_\alpha(\mathbb{B}_t \to \mathbb{A}_{t+1}) + \mathsf{RTE}_\alpha(\mathbb{B}_t \to \mathbb{A}_{t-1})$

We still need one last notion that will be useful in defining Φ. The notion of INTEGRATED INFORMATION OF A PARTITION OF THE SYSTEM $\mathscr{P} = \mathbb{A} \uplus_\alpha \mathbb{B}$, with reduction of interactions $\mathbb{A}/\mathbb{B}$ reduced by a factor α, which is defined by

$$\mathbb{III}_\alpha(\mathscr{P})(t) = \mathbb{III}_\alpha(\mathbb{A}_t) + \mathbb{III}_\alpha(\mathbb{B}_t) \tag{9}$$

Notes and Interpretations:

- The way interactions are reduced in the attenuated system is crucial. It is important to use a method that disturbs the internal dynamics of $\mathbb{A}$ and $\mathbb{B}$ as little as possible.
- The previous approach assumes that the partition between $\mathbb{A}$ and $\mathbb{B}$ is significant. If the choice of $\mathbb{A}$ is arbitrary, the results may not be informative.
- Calculating TE's requires estimates of the probability distributions from simulated or empirical data, which can be computationally demanding.

  This mathematical formalization provides a solid basis for quantifying the causal capacity of a mechanism in a complex system, combining the tool of Entropy Transfer with the key idea of ??the causal irreducibility of IIT.
- Em (2), $P(\mathsf{b}_{t+1}, \mathsf{a}_t, \mathsf{b}_t)$ is the joint probability function of the states b_{t+1} of mechanism $\mathbb{B}$ at time $t+1$, and of the states b_t and a_t at time t, etc. The sum $\sum$ is calculated over all possible combinations of states for a_t, b_t, and b_{t+1}. Each term in the sum corresponds to a specific combination of states, and the probability $P(\mathsf{b}_{t+1}, \mathsf{a}_t, \mathsf{b}_t)$ represents the probability that this combination will occur. The number of terms in the sum is equal to the total number of possible combinations of states. Analogous considerations are valid for the joint and conditional probabilities that appear in (2).

  In practice, we estimate the probabilities from simulated data:

  1. We count how many times each combination of states occurs in the simulated data.

2. We divide the count of each combination by the total number of samples to obtain an estimate of the probability.

3. We substitute the probability estimates into the formulas for TE and RTE, and calculate the sum.

Now that we have a way to quantify the integrated information of the partition $\mathscr{P} = \mathbb{A} \uplus_\alpha \mathbb{B}$, at a given instant t, we can use this information to define a measure of Φ for the system as a whole, following the principles of Integrated Information Theory (IIT).

IITproposes that consciousness is related to the amount of integrated information generated by a system. Therefore, our goal is to define Φ to capture the amount of information that the system generates and how integrated that information is.

Minimal Information Partitioning (MIP)

In Integrated Information Theory (IIT), the goal is to quantify how much a system is "*more than the sum of its parts*". This involves measuring its causal irreducibility, that is, how much its internal causal structure contributes to its dynamics, in a way that cannot be explained solely by the individual components. Let's see how:

1. Partitions as "cuts". To measure irreducibility, we imagine "cutting" the system $\mathbb{S}$ into separate α parts, $\mathscr{P} = \mathbb{A} \uplus_\alpha \mathbb{B}$, that is, the "cuts" include the reduction of the interactions between the resulting parts by a factor α, with $0 \leq \alpha \leq 1$:

$$\mathsf{J}_{\mathbb{AB}} \longrightarrow \alpha \cdot \mathsf{J}_{\mathbb{AB}}$$

   When $\alpha = 0$ the separation is total, the interactions $\mathbb{A}/\mathbb{B}$ are eliminated and the parts are independent – there is no information flow between them. On the other hand, when $\alpha = 1$ the system is fully integrated. The interactions $\mathbb{A}/\mathbb{B}$ are re-established, fully active and identical to the originals.

2. If, when cutting $\mathbb{S}$, $\mathscr{P} = \mathbb{A} \uplus_\alpha \mathbb{B}$, and making α vary, there is no loss of relevant information, then the system is reducible to its parts. If there is a loss of information, then the system is irreducible.

3. MIP – the Least Destructive Cut. The Minimum Information Partition (MIP) is the cut, or partition of the system, that causes the least possible loss of information. In causal terms, it is the way to divide the system into parts, in order to minimize the impact on its causal structure.

In general, we consider "cuts" of the system into only two independent parts. In fact, partitioning into two subsystems is a fundamental way to assess irreducibility. The idea is this: if a system is truly integrated, then it shouldn't be

possible to divide it into two parts that function independently without losing something essential about the system as a whole.

In mathematical and computational terms, it's easier to analyze partitions into two than partitions into many subsystems. The complexity of calculating all possible partitions increases exponentially with the number of subsystems.

Although IIT can, in principle, consider partitions into more than two subsystems, generally the binary partition (into two) is sufficient to capture most of the relevant information about irreducibility. If necessary, the MIP process can be applied recursively to the resulting subsystems to obtain a more detailed description of the causal structure of the system.

Let's now see how to mathematically formalize these ideas. For this, let us consider again a system $\mathbb{S}$, with N va's, and $\mathscr{P} = \mathbb{A} \uplus_\alpha \mathbb{B}$, a partition of $\mathbb{S}$ into two subsystems α-separate of va's: $\mathbb{A}$ and $\mathbb{B}$.

When $\mathbb{A}$ and $\mathbb{B}$ are part of a single system ($\alpha = 1$), they can freely exchange information. Actions in $\mathbb{A}$ can affect $\mathbb{B}$ (and vice versa) through direct and indirect interactions. Entropy Transfer captures this mutual causal influence. The formula (9), for $\alpha = 1$:

$$\mathbb{II}_1(\mathbb{A} \uplus \mathbb{B})(t) = \mathbb{II}_1(\mathbb{A}_t) + \mathbb{II}_1(\mathbb{B}_t) \tag{10}$$

represents the integrated information of the partition $\mathbb{A} \uplus \mathbb{B}$ of the original system, with all interactions $\mathbb{A}/\mathbb{B}$ active.

By attenuating (by a factor of $0 \leq \alpha < 1$) the interactions between $\mathbb{A}$ and $\mathbb{B}$, the information that previously flowed freely from $\mathbb{A}$ to $\mathbb{B}$ (and from $\mathbb{B}$ to $\mathbb{A}$) is now reduced to a value equal to:

$$\mathbb{II}_\alpha(\mathscr{P})(t) = \mathbb{II}_\alpha(\mathbb{A}_t) + \mathbb{II}_\alpha(\mathbb{B}_t), \quad \alpha < 1 \tag{11}$$

the integrated information of the partition, after the reduction of the interactions between $\mathbb{A}$ and $\mathbb{B}$, by a factor of $0 \leq \alpha < 1$.

The information loss associated with the partition $\mathscr{P} = \mathbb{A} \uplus_\alpha \mathbb{B}$, of the system $\mathbb{S}$, is defined by

$$\mathsf{Loss}_\alpha(\mathscr{P}) = \frac{\mathbb{II}_1(\mathscr{P})(t) - \mathbb{II}_\alpha(\mathscr{P})(t)}{\mathbb{II}_1(\mathscr{P})} \tag{12}$$

If $\alpha = 1$ then obviously $\mathsf{Loss}_\alpha(\mathscr{P}_1) = 0$. When $\alpha = 0$, $\mathbb{A}$ and B do not influence each other and therefore the information loss is maximum: $\mathsf{Loss}_0(\mathscr{P}_0) = 1$. The function $\mathsf{Loss}_\alpha(\mathscr{P})$ is a decreasing function of α, for a fixed $\mathscr{P}$.

If for a certain small tolerance $\tau > 0$ (in the figure **??**, $\tau = 0.08$) and $\alpha \approx 0$, $\mathsf{Loss}_\alpha(\mathscr{P}) \leq \tau$, then this means that, although the interactions $\mathbb{A}/\mathbb{B}$ have been strongly reduced (because $\alpha \approx 0$), the loss of integrated information remains within the tolerance ($\leq \tau$). This means that a separation of the system into two weakly interconnected parts results in weak information loss – there is almost no information loss in the separated α system and, therefore, the system is reducible to its parts $\mathbb{A}$ and $\mathbb{B}$ (Fig. **??**).

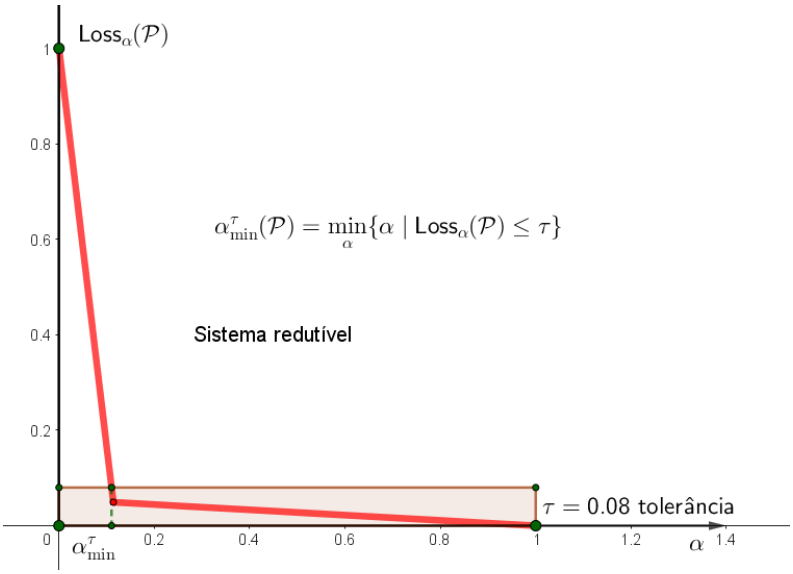


Figure 8: System reducible to its parts $\mathbb{A}$ and $\mathbb{B}$.

If for a certain small tolerance $\tau > 0$ (in the figure 8, $\tau = 0.08$) and $\alpha \approx 1$, $\text{Loss}_\alpha(\mathscr{P}) \geq \tau$, then this means that attenuating the interactions $\mathbb{A}/\mathbb{B}$, however little ($\alpha \approx 1$), causes a rapid loss of integrated information. (Fig. 9), and therefore the system is irreducible.

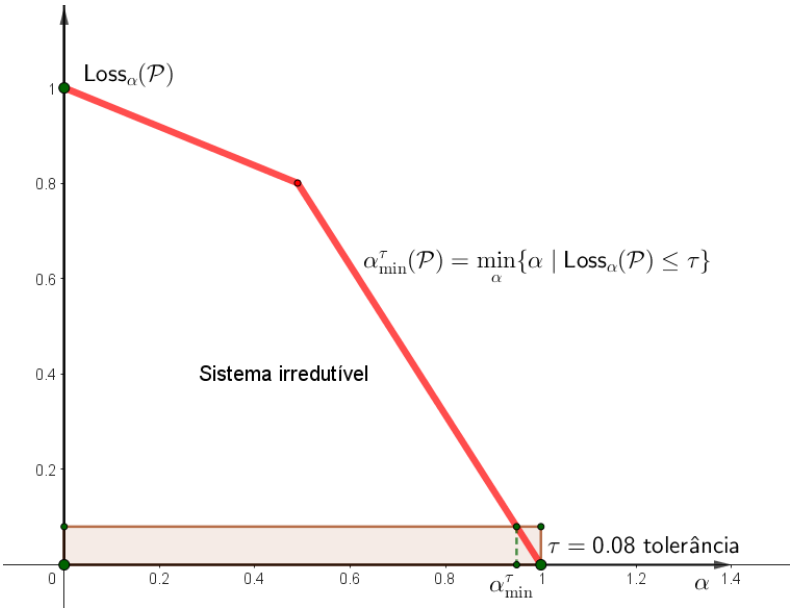


Figure 9: Irreducible system.

What underlies this is therefore comparing the loss of integrated information between two subsystems: first integrated into a single system ($\alpha = 1$),

and then α-separated ($0 < \alpha < 1$), with the interactions $\mathbb{A}/\mathbb{B}$ attenuated. The information loss quantifies this loss.

Let's consider $\tau > 0$ as the maximum tolerance for information loss. We define the *maximum tolerable reduction*, that is, the $\alpha^{\tau}_{\min}$ value, as the smallest value of α for which the information loss due to attenuation between $\mathbb{A}$ and $\mathbb{B}$ is still tolerable – less than or equal to τ.

$$\alpha^{\tau}_{\min}(\mathscr{P}) = \min_{\alpha}\{\alpha \mid \mathsf{Loss}_{\alpha}(\mathscr{P}) \leq \tau\} \tag{13}$$

For $\alpha < \alpha^{\tau}_{\min}(\mathscr{P})$ the reduction of interactions $\mathbb{A}/\mathbb{B}$ is stronger, which causes an unacceptable loss of information (greater than the tolerance τ). $\alpha^{\tau}_{\min}(\mathscr{P})$ therefore measures the greatest reduction of interactions for which the loss of information is acceptable (less than the tolerance τ). It is an approximation for the "weakest connection" between $\mathbb{A}$ and $\mathbb{B}$, for which the loss of information is tolerable.

In simpler terms, $\alpha^{\tau}_{\min}$ tells us what the maximum reduction in interactions $\mathbb{A}/\mathbb{B}$ (= smallest value of α) is that we can allow, keeping the loss of information within the tolerable limit τ, that is, before the system loses more than a tolerated amount of integrated information (τ). The greater the loss of information, the stronger the mutual influence between the subsystems in the original system ($\alpha = 1$) and the more significant the change caused by the separation.

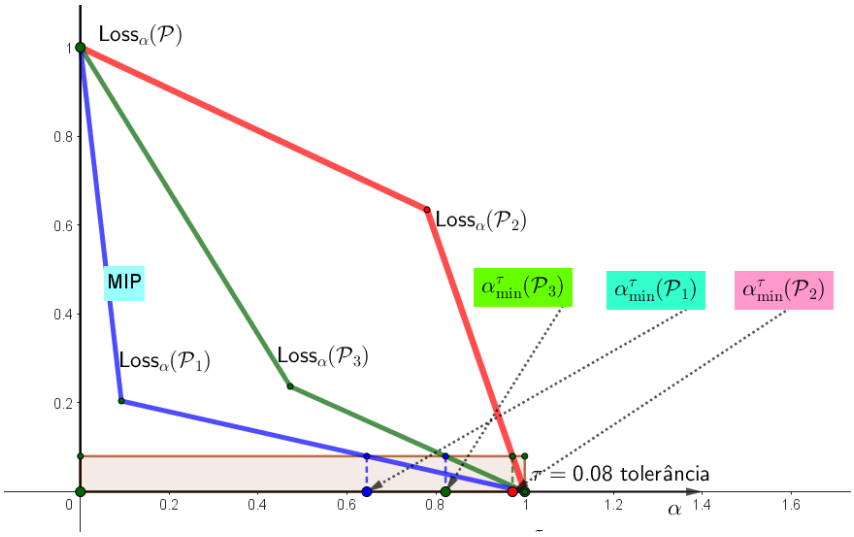


Figure 10: MIP

The MIP (Minimum (loss) Information Partition) is the partition $\mathscr{P}$ for which $\alpha^{\tau}_{\min}(\mathscr{P})$ is minimal, or that is, it is the *partition that corresponds to the maximum reduction of the interactions* $\mathbb{A}/\mathbb{B}$ (the percentage reduction is given by $(1-\alpha)\cdot 100\%$) *that keeps the information loss within the tolerance limit* τ.

$$\mathsf{MIP} = \{\mathscr{P} | \alpha^{\tau}_{\min}(\mathscr{P}) \text{ is minimum}\} \tag{14}$$

EXPLANATION OF THIS DEFINITION: we want the system to resemble the maximum possible to the extreme "reducible". The MIP is the "cut" of the system that causes the smallest change in its ability to integrate information.

We have already mentioned earlier in the (almost) extreme situation, with $\alpha^{\tau}_{\text{min}}(\mathscr{P}) \approx 0$, a system in which the loss remains below the tolerable level τ: $\text{Loss}(\alpha \approx 0) < \tau$ is considered reducible. We can then understand MIP as the partition $\mathscr{P}$ for which α can reach the lowest value, while the loss of integrated information, $\text{Loss}_{\alpha}(\mathscr{P})$, remains below the tolerance τ (the degree to which we can disrupt the system while maintaining its information). If the largest disturbance we can cause to the system still remains below the tolerance level, it is highly likely that the system is completely reducible.

MIP is the process of finding the most "natural" division of a system into two parts, minimizing the disruption of its causal structure. Although the theory allows for divisions into more parts, the focus on two parts provides a fundamental approach to quantifying irreducibility and facilitates analysis.

3 Φ

Φ is a measure of *"how much a system, as a whole, is more than the sum of its parts"*.

Before formally defining **Φ**, let's recap the definitions from the previous section:

1. We consider a disjoint partition of the system, $\mathscr{P} = \mathbb{A} \uplus_{alpha} \mathbb{B}$, with the interactions between $\mathbb{A}$ and $\mathbb{B}$ reduced by a factor of α, where $0 \leq \alpha \leq 1$:
2. The information loss due to the reduction by a factor of α is given by:

$$\text{Loss}_{\alpha}(\mathscr{P}) = \frac{\mathbb{III}_1(\mathbb{A} \uplus \mathbb{B}) - \mathbb{III}_{\alpha}(A \uplus_{\alpha} \mathbb{B})}{\mathbb{III}_1(\mathbb{A} \uplus \mathbb{B})}$$

3. Tolerance τ: A tolerance $\tau > 0$ is inserted to define an acceptable level of information loss.
4. $\alpha^{\tau}_{\text{min}}$: is the minimum value of α such that the information loss does not exceed the tolerance τ:

$$\alpha^{\tau}_{\text{min}} = \min\{\alpha : \text{Loss}_{\alpha}(\mathscr{P})\tau\}$$

5. MIP:

$$\text{MIP} = \arg\min_{\mathscr{P}} \alpha^{\tau}_{\text{min}}(\mathscr{P})$$

With these definitions, we finally define **Φ** as the integrated information of the partition MIP, with the interactions reduced by the factor $\alpha^* = \alpha^{\tau}_{\text{min}}(\text{MIP})$:

$$\mathbf{\Phi} = \mathbb{III}_{\alpha^*}(\text{MIP}) \tag{15}$$

Φ quantifies the integrated information of the minimum irreducible partition: MIP. It is a measure of the "irreducibility" of the system, or how much the system as a whole is "more than the sum of its parts."

If the system $\mathbb{S}$ were simply the sum of its parts (i.e., reducible), dividing it would cause a small loss of information, and Φ would be low. If, on the other hand, the system has a high level of integration and its parts interact in a complex and interdependent way, dividing it would cause a large loss of information, and Φ would be high.

In short:

- MIP finds the best way to divide the system.
- Φ quantifies how much the system is "*more than the sum of its parts*", based on this division.

By attenuating the interactions between subsystems, we could be considering interactions that are irrelevant to the behavior of the system. The introduction of tolerance τ and the definition of $\alpha^{\tau}_{\text{min}}$ help to solve this problem in the following way:

1. Relevance Threshold: The tolerance τ defines a minimum threshold for the relevance of interactions. We consider an interaction relevant if its reduction causes a loss of information greater than τ.
2. Selection of $\alpha^{\tau}_{\text{min}}$: By choosing $\alpha^{\tau}_{\text{min}}$ as the maximum reduction value, which causes an acceptable loss of information (up to τ), we are ensuring that we are not attenuating the interactions too much. We maintain the interactions that are important to keep the system integrated.
3. MIP as Optimal Cut: By finding the MIP, we are looking for the partition that causes the least loss of information while preserving relevant interactions. This means that we are not only minimizing information loss, but also ensuring that we are only considering the interactions that are truly important for the functioning of the system.

In summary: by inserting the tolerance τ and defining $\alpha^{\tau}_{\text{min}}$ and α^{*}, the formalism captures the idea that only relevant interactions should be considered in the calculation of Φ. The tolerance acts as a filter, removing interactions that are too weak to have a significant impact on the causal structure of the system. The MIP, in turn, ensures that the cut is made to minimize the loss of relevant information, that is, minimizing the impact on the system's ability to generate its own dynamics.

More notes and observations

1. Traditionally, MIP is seen as the way to divide the system that minimizes the loss of integrated information. In the context of this study, this "in-

formation" is causal information, that is, the ability of one subsystem to influence another forward and backward in time.

Thus, MIP becomes the way to cut the system that least disrupts the causal relationships between the subsystems. It identifies the division that results in the least loss of the ability of the subsystems to influence each other, both in the future and in the past. MIP represents the weakest joint (or the easiest connection to break) in the causal structure of the system. It is the division that has the least impact on the overall dynamics of the system.

2. A highly integrated system is a system where the parts depend heavily on each other. If we divide this system randomly, the loss of (causal) information will be significant.

   MIP attempts to find the division that preserves this causal interdependence as much as possible. By choosing MIP, we are trying to keep the system as "irreducible" as possible, even after we divide it.

   In practical terms: when analyzing a system using MIP, we are identifying which subsystems have the weakest causal relationships with each other. This information can be valuable for understanding the system's organization and for designing interventions or modifications that have the least possible impact.

   For example, if we have a complex control system and want to modify one of its components, we can use MIP to identify which component can be modified with the least impact on the overall system performance.

3. The inclusion of RTE in the definition of Φ makes MIP even more significant: MIP not only minimizes the loss of influence in the future (TE), but also minimizes the loss of the ability to constrain the past (RTE). It identifies a division that preserves both "forward" causality and "retrocausality." In short, MIP is a tool for identifying the causal structure of the system. Finding the "weakest joint" (least disruptive division). Preserving the irreducibility of the system as much as possible. Considering both forward causality and retrocausality.

More notes and observations on Φ

IIT emphasizes the importance of connectivity and interdependence among the components of a conscious system.

A system with high Φ is one where the components are strongly interconnected and where the activity of each component significantly influences the activity of the other components. A system with high Φ is highly integrated and generates a lot of information, while a system with low Φ is poorly integrated or generates little information.

A system with a high Φ value is a highly integrated system. Its components interact causally, and the information generated by the system as a whole is greater than the sum of the information generated by its individual parts – "*the whole is more than the sum of its parts*". Its current state provides a lot of information, not only about what will happen, but also about what has already happened. The system has a strong "causal memory" and its past and future are strongly constrained by its present.

In the context of theories of consciousness, Φ is a candidate for a measure of the amount of conscious experience generated by a system. A system with high Φ has a rich, complex, and integrated experience, characterized by a strong interdependence between its parts. Systems with low Φ would have simpler, fragmented, and less integrated experiences.

The human brain, with billions of interconnected neurons. The causal relationships between neurons are extremely complex and vary depending on the brain region and the task performed.

Finding the MIP for the brain is a computationally intractable problem. However, in theory, MIP would be the partition that minimizes the perturbation of neuronal activity and consciousness.

The human brain is believed to have a very high Φ, which reflects its high complexity and integration. This high integration is considered a necessary condition for consciousness.

IIT offers an ambitious framework for quantifying and understanding consciousness in terms of physical systems. While the exact calculation of Φ can be challenging, the concepts and insights of IIT provide a valuable guide for exploring the nature of subjective experience.

Example: The Ecosystem of a Forest

Let's consider a forest ecosystem as the system $\mathbb{S}$.

- The components (neurons) are: trees, plants, animals, fungi, soil, water, sunlight.
- The interactions are food chains, symbioses, competition for resources, nutrient cycling,

1. The goal of Minimum Information Partitioning (MIP) is to find the least destructive way to divide the forest into two subsystems, say $\mathbb{A}$ and $\mathbb{B}$. Some possible divisions:
    - $\mathscr{P}_1$: $\mathbb{A}$ = Plants, $\mathbb{B}$ = Animals,
    - $\mathscr{P}_2$: $\mathbb{A}$ = Producers (plants), $\mathbb{B}$ = Decomposers (fungi, bacteria) + Consumers (animals),

- $\mathscr{P}_3$: $\mathbb{A}$ = Northern half of the forest, $\mathbb{B}$ = Southern half of the forest

Now, we need to define the reduction factor α. For example:

- $\mathscr{P}_1$ ($\mathbb{A}$ = Plants, $\mathbb{B}$ = Animals): α can represent the reduction in the availability of plant food for animals (e.g., partial removal of certain plants).
- $\mathscr{P}_2$ ($\mathbb{A}$ = Producers, $\mathbb{B}$ = Decomposers + Consumers): α can represent the reduction in the nutrient cycle (for example, removal of some decomposers).
- $\mathscr{P}_3$ ($\mathbb{A}$ = Northern half, $\mathbb{B}$ = Southern half): α can represent the reduction in water flow between the two halves.

For each partition $\mathscr{P}$ and for each value of α, we measure the loss of information ($\mathsf{Loss}_\alpha(\mathscr{P})$). This loss can be measured, for example, by the reduction in biodiversity or by the decrease in forest productivity. We then calculate $\alpha^\tau_{\min}(\mathscr{P})$, for each partition. Let's say:

- $\alpha^\tau_{\min}(\mathscr{P}_1) = 0.6$ (a small change in consumers affects the rest of the system),
- $\alpha^\tau_{\min}(\mathscr{P}_2) = 0.97$, (decomposers can be slightly disturbed, affecting producers)
- $\alpha^\tau_{\min}(\mathscr{P}_3) = 0.8$, (A change in water flow results in more impact on the rest of the system)

The MIP would be the partition $\mathscr{P}_1$ ($\mathbb{A}$ = Plants, $\mathbb{B}$ = Animals), because it has the smallest $\alpha^\tau_{\min}$ ($= 0.60$).

Figure 11 shows that we can reduce the $\mathbb{A}/\mathbb{B}$ interactions to at most 60%, while maintaining a tolerable information loss ($\leq \tau$).

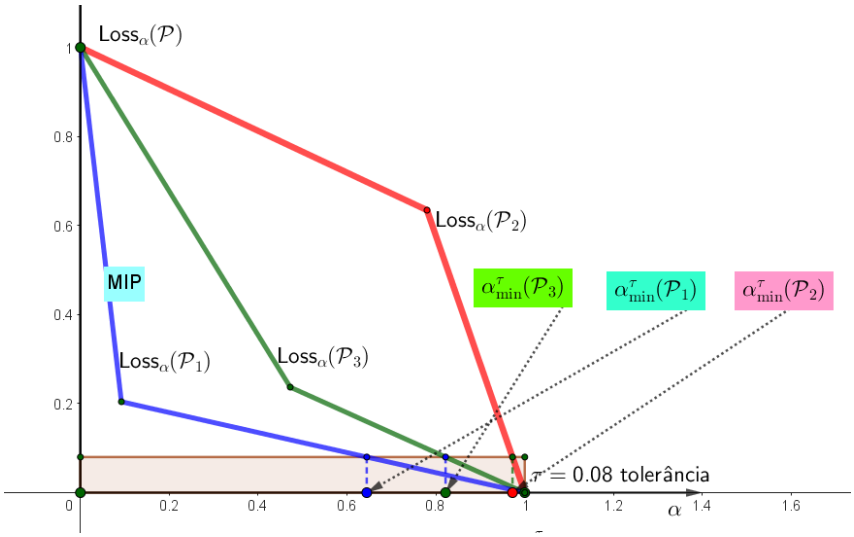


Figure 11: MIP

Φ is the information loss that effectively occurs when we divide the forest according to MIP and attenuate the connections by the factor α^* corresponding to $\alpha^\tau_{\min}(\mathscr{P}_1)$. If Φ is high, this means that the forest is highly integrated and the interactions between producers and consumers are crucial to its function. A small disturbance in these interactions causes a large loss.

If Φ is low, the forest is more reducible. The parts can function reasonably well independently. This illustration of the forest shows how MIP looks for the "most natural" way to divide a system, and how Φ quantifies how crucial the interactions between the parts are to the functioning of the system as a whole.

4 Integrated Information Theory (IIT) and GNW Model

The Global Neuronal Workspace (GNW) model, analyzed in detail in [JNT2025a], and the Integrated Information Theory (IIT) address the question of consciousness from different perspectives, but share some fundamental concepts and ideas. Here is a look at how the concepts of the GNW model can be related to the concepts of the IIT model, developed in the previous section.

Hopfield Modules (Specialized Brain Areas).

- GNW. Each Hopfield module, $\mathcal{M}_m$, represents a brain area specialized in processing a specific type of information (e.g., vision, hearing, language).
- IIT. Hopfield modules can be seen as *mechanisms* that make up the conscious system. Each module is not conscious in itself, but contributes to the system's ability to generate and integrate information (i.e., to the value of Φ), as a whole.

Module Activation.

- GNW. The activation of a Hopfield module represents the degree to which the information it processes becomes relevant, based on the patterns it stores (these induced by external inputs). We call this *pattern-driven dynamics* (memories).
- IIT. The activation of a module can be seen as the *strength* or the *specificity* of that module's contribution to conscious experience. The more active a module is, the greater its impact on the causal structure of the system and, therefore, the greater its contribution to Φ.

Lateral Competition.

- GNW. Lateral competition between modules, guided by patterns, ensures that only the most relevant and salient information accesses the workspace and that, once the ignition state is reached, it disseminates information via feedback to the modules (see [JNT2025a]).
- IIT. Lateral competition can be seen as a mechanism for selecting the information that is integrated. By suppressing irrelevant information, lateral competition increases processing efficiency and the coherence of conscious experience.

  Lateral competition influences the causal structure of the system, determining which connections are more important and which are suppressed. This affects how information is integrated and, therefore, the value of Φ.

Global Workspace.

- GNW. The global workspace is the location where information from different modules is integrated and made available to other cognitive processes.
- IIT. Workspace can be seen as the substrate of consciousness, that is, the set of elements that form the conscious system. The structure and dynamics of the workspace determine the quantity and quality of integrated information that the system can generate, directly influencing the value of Φ.

Ignition.

- GNW. The ignition of the workspace represents the moment when information becomes "conscious" or accessible to other cognitive processes.
- IIT. Ignition can be seen as the transition of the system to a state of high complexity and integration, characterized by an abrupt increase in Φ. Ignition can be an event that marks a threshold of consciousness, where the system reaches a critical level of integrated information.

Synaptic Plasticity.

- GNW. Synaptic plasticity allows the model to learn and adapt to new environments and tasks.
- IIT. Plasticity can be seen as a mechanism to optimize the causal structure of the system, increasing its ability to generate and integrate information. Plasticity influences the connections between neurons and modules, altering the structure of the MICS (Maximum Irreducible Conceptual Structure) and, therefore, the value of Φ.

In conclusion: by analyzing the concepts of the GNW model through the lens of IIT, we can gain a deeper understanding of the relationship between architecture, dynamics, and the ability to generate conscious experience. IIT can provide a framework to quantify and compare different configurations of the GNW model, allowing us to explore how different factors influence consciousness.

We will now further systematize the role of IIT, and, in particular, of Φ, in the GNW model developed in [JNT2025a].

GNW **Model and Integrated Information Theory (**IIT**).**

In the context of the Global Workspace Neuronal (GNW) model based on Hopfield stochastic networks, Integrated Information Theory (IIT) and, in particular, the metric Φ, play a fundamental role in quantifying information integration and evaluating the system's consciousness potential. Here are the main points about the role of IIT and Φ in the GNWmodel:

1. IIT provides a theoretical basis for understanding how consciousness can emerge from complex physical systems like GNW. It establishes principles that link the physical properties of the system (connections between neurons, dynamics) to its ability to generate conscious experience.
2. IIT offers a framework for interpreting the dynamics of GNW in terms of information integration. It suggests that global ignition (the coordinated activation of the workspace) is a phenomenon where information from different modules integrates to create a unified experience.
3. IIT generates testable predictions about how different parameters of the model (coupling, temperature, noise) affect consciousness, which can be explored through simulations.

Role of Φ.

1. Φ is the central measure of IIT and quantifies the level of system consciousness. In the context of GNW, Φ attempts to measure how much the

workspace integrates information from the modules, in a causally irreducible way.

A high value of Φ suggests that GNW is generating a unified and rich experience, while a low value indicates a lack of integration and possibly the absence of consciousness.

2. Global ignition can be seen as a transition to a state of high information integration, characterized by an increase in Φ. Analyzing how Φ evolves over time can reveal the dynamics of ignition and the conditions necessary for the emergence of global consciousness.

3. The Φ metric allows us to investigate how different parameters of the model affect information integration. The strength of the coupling between modules and the workspace can influence the system's ability to integrate information.

   The level of noise or stochasticity can affect the coherence of representations and integration. Lateral competition can modulate the activity of modules and influence the dynamics of integration.

   Analyzing the relationship between these parameters and Φ can provide insights into the ideal conditions for the emergence of consciousness in GNW.

5 How to Define Integrated Information in the GNW Model.

In the GNW model, the partition already defined by the local modules and the workspace:

$$\mathsf{GNW} = \underbrace{\cup_{\mathsf{m}=1}^{M} \mathcal{M}_{\mathsf{m}}}_{\mathbb{A}} \uplus_{\alpha} \underbrace{\mathcal{W}}_{\mathbb{B}} \tag{16}$$

can be a good approximation for MIP and, under certain circumstances, it could be MIP itself. Let's see why.

First, the GNW model is built with a clear modular structure, with interconnected local modules and a central workspace. This structure suggests that modules and workspace can function as relatively independent units, with specific interactions between them.

Furthermore, local modules are specialized in processing different types of information (e.g., vision, hearing, language), while the workspace is responsible for integrating and coordinating these representations. This functional specialization suggests that modules can be considered as distinct subsystems.

Finally, the coupling between modules and the workspace (through the C and $\mathcal{F}$ matrices) can be relatively sparse or limited compared to interactions

within modules or within the workspace. This means that removing these interactions between modules and the workspace may not cause a major disturbance in the behavior of individual subsystems.

Let's now see under what circumstances the Modules/Workspace partition (16) could be the MIP itself.

1. If the modular structure of GNW is strong (that is, if the modules are highly cohesive internally and weakly coupled to each other and to the workspace), then the partition that separates the modules from the workspace can be a good approximation of MIP.

2. By separating the modules from the workspace, we are removing the interactions between them, but maintaining the internal interactions of each module and the workspace. If most of the causal action occurs within these components, then the loss of information will be minimized.

In this case, Φ would represent the amount of information that is integrated through the interactions between the modules and the work. A high value of Φ would indicate that the workspace is actively coordinating the activity of the modules, while a low value would indicate a lack of integration.

On the other hand, if the coupling between the modules and the workspace is very strong, then removing these interactions can significantly disrupt the behavior of the subsystems and lead to high information loss. In this case, the modules/workspace partition may not be the MIP. It may then be necessary to consider other partitions to find the MIP. For example, it may be necessary to consider partitions that divide the modules or the workspace into smaller parts.

If the internal dynamics of the modules and the workspace dominate the system behavior, then the choice of partition may have little influence on the value of Φ.

In summary: the partition defined by the local modules and the workspace is a good starting point for finding an approximation of the MIP in the GNWmodel, especially if the modular structure is strong and the coupling between the modules and the workspace is relatively weak. However, it is important to verify this assumption by comparing the information loss associated with this partition with other possible partitions and considering the network structure.

How to calculate Φ in the GNW model?

Let's detail how we can approach the definition and approximation of Φ, the integrated information measure of the IIT, in the context of the GNWmodel, described in [JNT2025a]. Given that the exact calculation of Φ is computationally complex, especially for large systems, we will focus on an approximation that captures the essence of causal irreducibility.

In this approximation, we assume that there are no direct interactions between the modules, only interactions through the workspace; this significantly simplifies the calculation of Φ. In this case, the workspace acts as an information "bottleneck" between the modules.

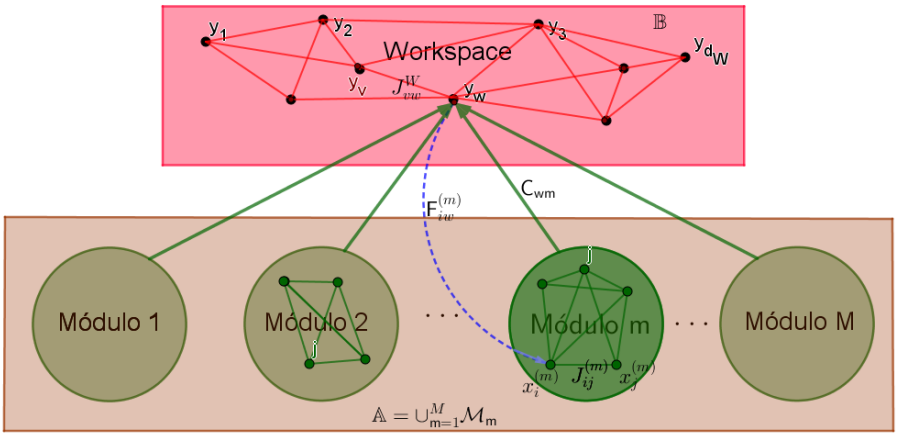


Figure 12: MIP partition in the GNWmodel.

Using the notations from section 3, we have the following in the GNW model (Fig. 12):

1. $\mathbb{S} =$ GNW, and according to (16), $\mathbb{A} = \cup_{\mathsf{m}=1}^{M} \mathcal{M}_{\mathsf{m}}$ and $\mathbb{B} = \mathcal{W}$, the workspace. Let's recall the definitions of the connectivity matrices $\mathsf{C} = C_{\mathsf{um}}$, where C_{um} represents the strength of the connection from the module $\mathcal{M}_{\mathsf{m}}$ to the node $\mathsf{u} \in \mathcal{W}$, and the feedback matrix $\mathcal{F}^{(\mathsf{m})} = (F_{i\mathsf{u}}^{(\mathsf{m})})$, where $F_{i\mathsf{u}}^{(\mathsf{m})}$ represents the strength of the feedback from the node $\mathsf{u} \in \mathcal{W}$ to the node $i \in \mathcal{M}_{\mathsf{m}}$.

2. Now we need to calculate the transfers, and reverse transfers, of entropy, TE and RTE: $\mathsf{TE}(\mathbb{A}_t \to \mathbb{B}_{t+1})$, $\mathsf{TE}(\mathbb{B}_t \to \mathbb{A}_{t+1})$, $\mathsf{RTE}(\mathbb{A}_t \to \mathbb{B}_{t-1})$ and $\mathsf{RTE}(\mathbb{B}_t \to \mathbb{A}_{t-1})$ whose definitions are those adopted in the Appendix 9.

   For example, $\mathsf{TE}(\mathbb{A}_t \to \mathbb{B}_{t+1})$ measures how much the activity of the modules at time t, $\mathbb{A}_t$, influences the state of the workspace $\mathbb{B}_{t+1}$. Since we are assuming that there are no direct interactions between the modules, we can calculate the TE of the joint activity of the modules for the Workspace.

   More specifically:

$$\mathsf{TE}(\mathbb{A}_t \to \mathbb{B}_{t+1}) = \mathsf{Ent}(\mathbb{B}_{t+1}|\mathbb{B}_t) - \mathsf{Ent}(\mathbb{B}_{t+1}|\mathbb{B}_t, \mathbb{A}_t) \tag{17}$$

   where $B_t = \mathbf{y}(t)$ is the state of the workspace at time t, and $\mathbb{A}_t = \mathbf{x}(t)$ represents the joint state of the modules at time t, that is:

$$\mathbb{A}_t = \left(\mathbf{x}_t^{(1)}, \cdots, \mathbf{x}_t^{(M)}\right)$$

is the concatenated vector of the states of the M modules.

Let us also remember that:

$$\mathsf{TE}(\mathbb{A}_t \to \mathbb{B}_{t+1}) = \sum P(\mathbb{B}_{t+1}, \mathbb{B}_t, \mathbb{A}_t) \cdot \log_2 \frac{P(\mathbb{B}_{t+1}|\mathbb{B}_t, \mathbb{A}_t)}{P(\mathbb{B}_{t+1}|\mathbb{B}_t)}$$

If $\mathbb{A}_t$ has no influence on the transition from $\mathbb{B}_t \longrightarrow \mathbb{B}_{t+1}$, then $P(\mathbb{B}_{t+1}|\mathbb{B}_t, \mathbb{A}_t) = P(\mathbb{B}_{t+1}|\mathbb{B}_t)$ and, in that case, $\mathsf{TE}(\mathbb{A}_t \to \mathbb{B}_{t+1})$ will be zero. If $\mathbb{A}_t$ influences the transition, then $P(\mathbb{B}_{t+1}|\mathbb{B}_t, \mathbb{A}_t)$ will be significantly different from $P(\mathbb{B}_{t+1}|\mathbb{B}_t)$. The value of this difference, weighted by the joint probability, will determine the value of TE.

Note that

$$\mathsf{TE}(\mathbb{A}_t \to \mathbb{B}_{t+1}) = \mathsf{Ent}(\mathbb{B}_{t+1}|\mathbb{B}_t) - \mathsf{Ent}(\mathbb{B}_{t+1}|\mathbb{B}_t, \mathbb{A}_t))$$

is a theoretical formula that assumes perfect knowledge of the probability distributions.

It represents the true Entropy Transfer (which is usually unknown).

The Practical Formula (Estimate) should be:

$$\widehat{\mathsf{TE}}(\mathbb{A}_t \to \mathbb{B}_{t+1}) = \widehat{\mathsf{Ent}}(\mathbb{B}_{t+1}|\mathbb{B}_t) - \widehat{\mathsf{Ent}}(\mathbb{B}_{t+1}|\mathbb{B}_t, \mathbb{A}_t))$$

which uses estimates of the probability distributions, based on simulated data. It represents the best estimate of Entropy Transfer. The more data we have and the better the estimation methods, the closer the estimate will be to the true (unknown) value.

Steps for computational simulation.

1. Simulate the GNW a sufficient number of time steps, $\mathbb{T} \in \mathbb{N}\Delta t$, to capture the dynamics of the system. Store the states of the modules $\mathbb{A}_t = \mathbf{x}(t)$ and the workspace $\mathbb{B}_t = \mathbf{y}(t)$, for each time step $t = 1, \cdots, \mathbb{T}$

2. Estimate the conditional probabilities needed to calculate the Entropy. For example, to calculate $\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+|\mathbb{B})$ (see the following note on the notations), count how many times each state transition $\mathbb{B}_t \to \mathbb{B}_{t+1}$ occurs in the stored data.

   IMPORTANT NOTE: When we calculate, for example, the entropy estimate $\widehat{\mathsf{Ent}}(\mathbb{B}_{t+1}|\mathbb{B}_t)$, using the conditional probability estimates obtained by counting, as in a previous example, the dependence on t disappears.

   In fact, what we are calculating is actually an average conditioned entropy over its entire time series, which we will denote by $\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+|\mathbb{B})$, with analogous notations for all other estimates obtained by counting:

$$\begin{aligned} \widehat{\mathsf{Ent}}(\mathbb{B}_{t+1}|\mathbb{B}_t, \mathbb{A}_t) &\longrightarrow \widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+|\mathbb{B}, \mathbb{A}) \\ \widehat{\mathsf{TE}}(\mathbb{A}_t \to \mathbb{B}_{t+1}) &\longrightarrow \widehat{\mathsf{TE}}_{\mathbb{T}}(\mathbb{A} \to \mathbb{B}^+) \\ \widehat{\mathsf{RTE}}(\mathbb{A}_t \to \mathbb{B}_{t-1}) &\longrightarrow \widehat{\mathsf{RTE}}_{\mathbb{T}}(\mathbb{A} \to \mathbb{B}^-), \quad \text{etc.} \end{aligned} \tag{18}$$

As we discussed earlier, the sliding window approach is the key to estimating "instantaneous" quantities.

3. Calculate $\widehat{\mathsf{TE}}_{\mathbb{T}}(\mathbb{A} \to \mathbb{B}^+)$:
   - Calculate $\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+|\mathbb{B})$.
   - Calculate $\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+|\mathbb{B}, \mathbb{A})$. For this, we need to estimate the conditional probability $P(\mathbb{B}^+|\mathbb{B}, \mathbb{A})$.
   - Subtract:
   $$\widehat{\mathsf{TE}}_{\mathbb{T}}(\mathbb{A} \to \mathbb{B}^+) = \widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+|\mathbb{B}) - \widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+|\mathbb{B}, \mathbb{A})$$
4. Calculate
   $$\widehat{\mathsf{TE}}_{\mathbb{T}}(\mathbb{B} \to \mathbb{A}^+)$$
   analogously.

Calculate the estimates of the integrated information $\widehat{\mathbb{II}}_{\mathbb{T}}(\mathbb{A})$ and $\widehat{\mathbb{II}}_{\mathbb{T}}(\mathbb{B})$. Note that we are assuming that $\mathsf{GNW} = \mathbb{A} \uplus \mathbb{B}$ is the MIP. We do not need to reduce interactions by the factor α, etc.

Calculate $\mathbf{\Phi}$:
$$\hat{\mathbf{\Phi}}_{\mathbb{T}} \approx \widehat{\mathbb{II}}(\mathbb{A}) + \widehat{\mathbb{II}}(\mathbb{B})$$

$\mathsf{TE}(\mathbb{A} \to \mathbb{B})(t))$: Instantaneous Entropy Transfer from $\mathbb{A}$ to $\mathbb{B}$ at time t.
$\widehat{TE}(A \to B)$: Estimate of the average Entropy Transfer from $\mathbb{A}$ to $\mathbb{B}$ over time.
With this notation, the relationship between the two quantities is:

$$\widehat{\mathsf{TE}}(\mathbb{A} \to \mathbb{B}) \approx \frac{1}{T-1} \sum_{t=1}^{T-1} \mathsf{TE}(\mathbb{A} \to \mathbb{B})(t)$$

This means that our estimate of the average Entropy Transfer is approximately equal to the average of the instantaneous Entropy Transfers over time.

The Implication for Formulas: the formulas we were using until now directly calculate the estimate of the average Entropy Transfer,

$$\widehat{\mathsf{TE}}(\mathbb{A} \to \mathbb{B})$$

Therefore, they already incorporate the average over time and do not need an explicit time index. The correct formulas are:

$$\widehat{\mathsf{TE}}(\mathbb{A}_t \to \mathbb{B}_{t+1}) = \widehat{\mathsf{Ent}}(\mathbb{B}_{t+1}|\mathbb{B}_t) - \widehat{\mathsf{Ent}}(\mathbb{B}_{t+1}|\mathbb{B}_t, \mathbb{A}_t)$$

Where: $\widehat{\mathsf{Ent}}(\mathbb{B}_{t+1}|\mathbb{B}_t)$ is the estimate of the average entropy of the state of B at time $t+1$, given its state at time t, and $\widehat{\mathsf{Ent}}(\mathbb{B}_{t+1}|\mathbb{B}_t)$ is the estimate of the average entropy of the state of $\mathbb{B}$ at time $t+1$ given its state at time t and the state of $\mathbb{A}$ at time t.

Estimating $\widehat{\mathsf{RTE}}_{\mathbb{T}}(\mathbb{A} \to \mathbb{B}^-)$

1. Record data by $\mathbb{T}$ time instants, with the following structure:

$$(\mathbb{A}_{t-1}, \mathbb{B}_{t-1}, \mathbb{A}_t, \mathbb{B}_t)$$

   to be able to calculate the estimates of the conditional probabilities necessary for the RTE.

2. Conditional Probability Estimation: use the data count to estimate the following conditional probabilities:

$$P(\mathbb{B}^-|\mathbb{B}, \mathbb{A}) = \frac{\mathsf{Count}(\mathbb{B}_{t-1}, \mathbb{B}_t, \mathbb{A}_t)}{\mathsf{Count}(\mathbb{B}_t, \mathbb{A}_t)}$$

$$P(\mathbb{B}^-|\mathbb{B}) = \frac{\mathsf{Count}(\mathbb{B}_{t-1}, \mathbb{B}_t)}{\mathsf{Count}(\mathbb{B}_t)}$$

3. Estimated Reverse Entropy Transfer Calculation ($\widehat{\mathsf{RTE}}$):

$$\widehat{\mathsf{RTE}}_{\mathbb{T}}(\mathbb{A} \to \mathbb{B}^-) = \sum P(\mathbb{B}^-, \mathbb{B}, \mathbb{A}) \cdot \log_2 \frac{P(\mathbb{B}^-|\mathbb{B}, \mathbb{A})}{P(\mathbb{B}^-|\mathbb{B})}$$

   where the sum is over all possible states of $\mathbb{B}^-, \mathbb{B}$ and $\mathbb{A}$, and

$$P(\mathbb{B}^-, \mathbb{B}, \mathbb{A}) = \frac{\mathsf{Count}(\mathbb{B}_{t-1}, \mathbb{B}_t, \mathbb{A}_t)}{\mathsf{Total\ of\ Samples}}$$

Important Note

The counting method we have used so far calculates estimates of probabilities and, consequently, of entropy transfers (TE's and RTE's) that represent averages over the time series. This means that we lose any information about how these quantities may vary over time.

To calculate "instantaneous" estimates of these quantities at time t, we need to modify this approach to use sliding windows. The basic idea is to calculate the conditional probabilities and entropy transfers not over the entire time series, but only over a time window centered around the point t.

This will give an estimate of how these quantities behave at that moment t. Here's how to proceed:

1. Set Window Size (W): choose a window size W that is appropriate for the timescale of the changes we are interested in capturing.

2. A larger value of W will provide smoother estimates, but will also attenuate rapid changes. A smaller value of W will be more sensitive to rapid changes, but will also be noisier.

3. Sliding the Window Over Time:

   - For each time instant t in the time series, create a window that includes the W data points centered around t:

     $$\text{Window} = [t - W/2, t + W/2]$$

   - If W is odd, the window will be symmetrical around t. If W is even, we may need to slightly adjust the position of the window to ensure it is well defined.

   - Deal with the edges of the time series (when $t - W/2 < 0$ or $t + W/2 >$ the length of the time series) by truncating the window or using other padding techniques.

   - Calculate the conditional probabilities in the Window:

   - For each t, calculate the conditional probabilities (and the joint probabilities) using only the data within the window centered on t. Use the same counting formulas we have used so far, but restrict the calculations to the data within the window.

4. Calculate Instantaneous TE and RTE:

   - Use the conditional probabilities calculated in the window to calculate the "instantaneous" estimates of TE and RTE at the time point t: $\mathsf{TE}(\mathbb{A}_t \to \mathbb{B}_{t+1})$ is now an estimate of the entropy transfer at that instant in time, based on the behavior of the system within the window. $\mathsf{RTE}(\mathbb{A}_t \to \mathbb{B}_{t-1})$ is an estimate of the reverse entropy transfer at that instant in time.

A simple numerical example. Let's assume that the discretized time series $\mathbb{B}_t$, of the workspace (from $t = 1$ to $t = 7$) is:

$$-1, -1, +1, +1, -1, +1, +1$$

We have to analyze the transitions $\mathbb{B}_t \longrightarrow \mathbb{B}_{t+1}$, for $t = 1, 2, \cdots, 6$. Putting this into a table (on the left) and a counting table (on the right), we have respectively:

t	$\mathbb{B}_t$	$\mathbb{B}_{t+1}$
1	-1	-1
2	-1	$+1$
3	$+1$	$+1$
4	$+1$	-1
5	-1	$+1$
6	$+1$	$+1$

$\mathbb{B}_t$	$\mathbb{B}_{t+1}$	Counting
-1	-1	1
-1	$+1$	2
$+1$	-1	1
$+1$	$+1$	2

To calculate the conditional probabilities, we have to normalize the counts: for each state $\mathbb{B}_t$, divide the count of each transition $\mathbb{B}_t \longrightarrow \mathbb{B}_{t+1}$ (2nd table), by the total number of times that the state $\mathbb{B}_t$ occurs in the 1st table. This will give the estimate, based on the data, of the conditional probability $P(\mathbb{B}^+|\mathbb{B})$. Using the 1st count table, we see that the state -1 occurs 3 times as $\mathbb{B}_t$ and the state $+1$ also occurs 3 times as $\mathbb{B}_t$. Thus:

$\mathbb{B}$	$\mathbb{B}^+$	Count	$P(\mathbb{B}^+\|\mathbb{B})$
-1	-1	1	$1/3$
-1	$+1$	2	$2/3$
$+1$	-1	1	$1/3$
$+1$	$+1$	2	$2/3$

Now that we have the conditional probabilities, we can calculate the conditional entropy estimate $\widehat{\mathsf{Ent}}(\mathbb{B}^+|\mathbb{B})$, using the formula:

$$\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+|\mathbb{B}) = -\sum_{\mathsf{b}} P(\mathsf{b}^+) \sum_{\mathsf{b}} P(\mathsf{b}^+|\mathsf{b}) \log_2 P(\mathsf{b}^+|\mathsf{b})$$

where the first sum is over all possible states b^+ of $\mathbb{B}^+$ and the second is over all possible states b of $\mathbb{B}$. Using the formula (**??**) from the Appendix, we obtain

$$\begin{aligned}
\widehat{\mathsf{Ent}}(\mathbb{B}^+|\mathbb{B}) &= -P(-1)[P(-1|-1)\log_2 P(-1|-1) + P(+1|-1)\log_2 P(+1|-1)] \\
&\quad -P(+1)[P(-1|+1)\log_2 P(-1|+1) + P(+1|+1)\log_2 P(+1|+1)] \\
&= -0.5[1/3\log_2 1/3 + 2/3\log_2 2/3] - 0.5[1/3\log_2 1/3 + 2/3\log_2 2/3] \\
&\approx 0.918 \text{ bits} \qquad (19)
\end{aligned}$$

Numerical Example of Entropy Transfer and Φ

We will present a complete numerical example of calculating the Entropy Transfer estimate TE and $\mathbf{\Phi}$, using the states -1 (inactive) and $+1$ (active).

Simulation and Data. We simulate two subsystems, $\mathbb{A}$ and $\mathbb{B}$, for $\mathbb{T} = 1000$ time steps. The states of each subsystem are discretized into -1 and $+1$. The sequences of recorded states are:

- $\mathbb{A}_t$: $-1, -1, +1, -1, +1, +1, -1, -1, +1, +1, ...$ (1000 values)
- $\mathbb{B}_t$: $+1, -1, -1, +1, +1, -1, +1, -1, -1, +1, ...$ (1000 values)

The transitions $\mathbb{B}_t \longrightarrow \mathbb{B}_{t+1}$, and the counting table, are respectively:

t	$\mathbb{B}_t$	$\mathbb{B}_{t+1}$
1	$+1$	-1
2	-1	-1
3	-1	$+1$
4	$+1$	$+1$
5	$+1$	-1
$\vdots$	$\vdots$	$\vdots$

$\mathbb{B}_t$	$\mathbb{B}_{t+1}$	Count
-1	-1	200
-1	$+1$	280
$+1$	-1	280
$+1$	$+1$	240

Marginal probabilities from the counts. For $t = 1, 2 \cdots, 999$:

- $\mathsf{Count}(\mathbb{B}_t = -1) = 480 \Longrightarrow P(\mathbb{B}_t = -1) = \dfrac{480}{1000} = 0.48$
- $\mathsf{Count}(\mathbb{B}_t = +1) = 520 \Longrightarrow P(\mathbb{B}_t = +1) = \dfrac{520}{1000} = 0.52$

Conditioned Probabilities

- $P(\mathbb{B}^+ = -1 | \mathbb{B} = -1) = \dfrac{200}{480}$
- $P(\mathbb{B}^+ = +1 | \mathbb{B} = -1) = \dfrac{280}{480}$
- $P(\mathbb{B}^+ = +1 | \mathbb{B} = -1) = \dfrac{280}{480}$
- $P(\mathbb{B}^+ = +1 | \mathbb{B} = +1) = \dfrac{240}{520}$

Conditional probabilities for $\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+ | \mathbb{B}, \mathbb{A})$

$\mathbb{B}_t$	$\mathbb{A}_t$	$\mathbb{B}_{t+1}$	Contagem
-1	-1	-1	100
-1	-1	$+1$	120
-1	$+1$	-1	100
-1	$+1$	$+1$	160
$+1$	-1	-1	130
$+1$	-1	$+1$	110
$+1$	$+1$	-1	120
$+1$	$+1$	$+1$	170

To obtain the conditional probabilities, we first need to calculate $P(\mathbb{B}_t = \mathsf{b}_t, \mathbb{A}_t = \mathsf{a}_t)$:

$\mathbb{A}_t$	$\mathbb{B}_t$	Counting
-1	-1	220
-1	$+1$	230
$+1$	-1	260
$+1$	$+1$	290

Therefore:

- $P(\mathbb{B}=-1,\mathbb{A}=-1)=220/1000=0.22$
- $P(\mathbb{B}=+1,\mathbb{A}=-1)=230/1000=0.23$
- $P(\mathbb{B}=-1,\mathbb{A}=+1)=260/1000=0.26$
- $P(\mathbb{B}=+1,\mathbb{A}=+1)=290/1000=0.29$

and finally:

- $P(\mathbb{B}^+=-1|\mathbb{B}=-1,\mathbb{A}=-1)=\frac{100}{220}\approx 0.455$
- $P(\mathbb{B}^+=+1|\mathbb{B}=-1,\mathbb{A}=-1)=\frac{120}{220}\approx 0.545$
- $P(\mathbb{B}^+=-1|\mathbb{B}=-1,\mathbb{A}=+1)=\frac{100}{260}\approx 0.385$
- $P(\mathbb{B}^+=+1|\mathbb{B}=-1,\mathbb{A}=+1)=\frac{160}{260}\approx 0.615$
- $P(\mathbb{B}^+=-1|\mathbb{B}=+1,\mathbb{A}=-1)=\frac{130}{230}\approx 0.565$
- $P(\mathbb{B}^+=+1|\mathbb{B}=+1,\mathbb{A}=-1)=\frac{110}{230}\approx 0.478$
- $P(\mathbb{B}^+=-1|\mathbb{B}=+1,\mathbb{A}=+1)=\frac{120}{290}\approx 0.414$
- $P(\mathbb{B}^+=+1|\mathbb{B}=+1,\mathbb{A}=+1)=\frac{170}{290}\approx 0.586$

Calculation of Conditioned Entropies.

$$\begin{aligned}\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+|\mathbb{B}) &= -\sum_{\mathsf{b}} P(\mathsf{b})\sum_{\mathsf{b}^+} P(\mathsf{b}^+|\mathsf{b})\log_2 P(\mathsf{b}^+|\mathsf{b}) + \\ &= -0.48\,[0.417\log_2(0.417)+0.583\log_2(0.583)] \\ &\quad -0.52\,[0.481\log_2(0.481)+0.462\log_2(0.462)] \\ &\approx 0.990 \text{ bits}\end{aligned}$$

$$\begin{aligned}\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^+|\mathbb{B},\mathbb{A}) &= -\sum_{\mathsf{a},\mathsf{b}^+} P(\mathsf{a}^+,\mathsf{b})\sum_{\mathsf{b}^+} P(\mathsf{b}^+|\mathsf{b},\mathsf{a})\cdot\log_2 P(\mathsf{b}^+|\mathsf{b},\mathsf{a}) \\ &= -0.22\,[0.455\log_2(0.455)+0.545\log_2(0.545)] + \\ &\quad -0.26\,[0.385\log_2(0.385)+0.615\log_2(0.615)] \\ &= -0.23\,[0.565\log_2(0.565)+0.478\log_2(0.478)] + \\ &\quad -0.29\,[0.414\log_2(0.414)+0.586\log_2(0.586)] \\ &\approx 0.971 \text{ bits}\end{aligned}$$

Calculation of Entropy Transfers (TE)

$$\begin{aligned}\widehat{\mathsf{TE}}_{\mathbb{T}}(\mathbb{A}\to\mathbb{B}^{+}) &= \mathsf{Ent}(\mathbb{B}^{+}|\mathbb{B})-\mathsf{Ent}(\mathbb{B}^{+}|\mathbb{B},\mathbb{A})\\ &= 0.990-0.971=0.019\ \text{bits}\end{aligned}$$

Assumimos que $\widehat{\mathsf{TE}}_{\mathbb{T}}(\mathbb{B}\to\mathbb{A}^{+})=0.015$ bits.

Calculation of Reverse Entropy Transfers (RTE)

$$\begin{aligned}\widehat{\mathsf{RTE}}_{\mathbb{T}}(\mathbb{A}\to\mathbb{B}^{-}) &= \mathsf{Ent}(\mathbb{B}^{-}|\mathbb{B})-\mathsf{Ent}(\mathbb{B}^{-}|\mathbb{B},\mathbb{A})\\ &= \text{bits}\end{aligned}$$

We assume that$\widehat{\mathsf{RTE}}_{T}(\mathbb{B}\to\mathbb{A}^{-})=0.015$ bits.

Conditional probabiliteis for $\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^{-}|\mathbb{B},\mathbb{A})$

$\mathbb{B}_t$	$\mathbb{A}_t$	$\mathbb{B}_{t-1}$	Counting
-1	-1	-1	120
-1	-1	$+1$	160
-1	$+1$	-1	110
-1	$+1$	$+1$	120
$+1$	-1	-1	170
$+1$	-1	$+1$	100
$+1$	$+1$	-1	100
$+1$	$+1$	$+1$	130

To obtain the conditional probabilities, we first need to calculate $P(\mathbb{B}_t=\mathsf{b}_t,\mathbb{A}_t=\mathsf{a}_t)$:

$\mathbb{A}_t$	$\mathbb{B}_t$	Counting
-1	-1	220
-1	$+1$	230
$+1$	-1	260
$+1$	$+1$	290

Then:

- $P(\mathbb{B}=-1,\mathbb{A}=-1)=220/1000=0.22$
- $P(\mathbb{B}=+1,\mathbb{A}=-1)=230/1000=0.23$
- $P(\mathbb{B}=-1,\mathbb{A}=+1)=260/1000=0.26$
- $P(\mathbb{B}=+1,\mathbb{A}=+1)=290/1000=0.29$

and finally:

- $P(\mathbb{B}^{-}=-1|\mathbb{B}=-1,\mathbb{A}=-1)=\frac{120}{220}\approx 0.545$

- $P(\mathbb{B}^- = +1|\mathbb{B} = -1, \mathbb{A} = -1) = \frac{160}{220} \approx$
- $P(\mathbb{B}^- = -1|\mathbb{B} = -1, \mathbb{A} = +1) = \frac{110}{260} \approx$
- $P(\mathbb{B}^- = +1|\mathbb{B} = -1, \mathbb{A} = +1) = \frac{120}{260} \approx$
- $P(\mathbb{B}^- = -1|\mathbb{B} = +1, \mathbb{A} = -1) = \frac{170}{230} \approx$
- $P(\mathbb{B}^- = +1|\mathbb{B} = +1, \mathbb{A} = -1) = \frac{100}{230} \approx$
- $P(\mathbb{B}^- = -1|\mathbb{B} = +1, \mathbb{A} = +1) = \frac{100}{290} \approx$
- $P(\mathbb{B}^- = +1|\mathbb{B} = +1, \mathbb{A} = +1) = \frac{130}{290} \approx$

Calculation of Conditioned Entropies.

$$\begin{aligned}
\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^-|\mathbb{B}) &= -\sum_{\mathsf{b}} P(\mathsf{b}) \sum_{\mathsf{b}^-} P(\mathsf{b}^-|\mathsf{b}) \log_2 P(\mathsf{b}^-|\mathsf{b}) + \\
&= -0.48\,[0.417 \log_2(0.417) + 0.583 \log_2(0.583)] \\
&\quad -0.52\,[0.481 \log_2(0.481) + 0.462 \log_2(0.462)] \\
&\approx 0.990 \text{ bits}
\end{aligned}$$

$$\begin{aligned}
\widehat{\mathsf{Ent}}_{\mathbb{T}}(\mathbb{B}^-|\mathbb{B}, \mathbb{A}) &= -\sum_{\mathsf{a},\mathsf{b}^-} P(\mathsf{a}^-, \mathsf{b}) \sum_{\mathsf{b}^+} P(\mathsf{b}^-)|\mathsf{b}, \mathsf{a}) \cdot \log_2 P(\mathsf{b}^-|\mathsf{b}, \mathsf{a}) \\
&= -0.22\,[0.455 \log_2(0.455) + 0.545 \log_2(0.545)] + \\
&\quad -0.26\,[0.385 \log_2(0.385) + 0.615 \log_2(0.615)] \\
&= -0.23\,[0.565 \log_2(0.565) + 0.478 \log_2(0.478)] + \\
&\quad -0.29\,[0.414 \log_2(0.414) + 0.586 \log_2(0.586)] \\
&\approx 0.971 \text{ bits}
\end{aligned}$$

Calculation of Integrated Information and Φ

$$\begin{aligned}
\widehat{\mathbb{II}}_{\mathbb{T}}(\mathbb{A}) &= \widehat{\mathsf{TE}}_{\mathbb{T}}(\mathbb{A} \to \mathbb{B}^+) + \widehat{\mathsf{RTE}}_{\mathbb{T}}(\mathbb{A} \to \mathbb{B}^-) \\
&= 0.019 + ... \text{ bits} \\
\widehat{\mathbb{II}}_{\mathbb{T}}(\mathbb{B}) &= \widehat{\mathsf{TE}}_{\mathbb{T}}(\mathbb{B} \to \mathbb{A}^+) + \widehat{\mathsf{RTE}}_{\mathbb{T}}(\mathbb{B} \to \mathbb{A}^-) \\
&= 0.015 + ... \text{ bits}
\end{aligned} \tag{20}$$

and finally:

$$\begin{aligned}
\widehat{\mathbf{\Phi}}_{\mathbb{T}} &= \widehat{\mathbb{II}}_{\mathbb{T}}(\mathbb{A}) + \widehat{\mathbb{II}}_{\mathbb{T}}(\mathbb{B}) \\
&= ..... \text{ bits}
\end{aligned} \tag{21}$$

6 Formal Relationship between Workspace Ignition in the GNW model and Φ of IIT

The GNW (Global Network Workspace) model interconnects specialized modules (analogous to cortical areas) with a central Workspace (analogous to a global workspace). The stochastic dynamics are governed by Hopfield networks and the Metropolis method.

Integrated Information Theory (IIT) proposes Φ as a measure of phenomenological consciousness, quantifying the information intrinsically integrated by a system. The central question is to establish a formal relationship between "*Workspace Ignition*" – a dynamic event in GNW – and the value of Φ calculated for the system. The investigation of the emergence of consciousness requires an explicit connection between the physical mechanisms of a model and the theoretical principles that describe it. In the GNW model, the "Workspace Ignition" is postulated as a neuronal correlate of consciousness.

In other words, it is the minimum necessary and sufficient state (or physical/computational process) for a given aspect of (subjective) consciousness to occur. The "Workspace Ignition" is the physical/computational manifestation of consciousness that is, moreover, dynamicactive and temporally evolving, involving transitions, interactions, and changes over time.

> *In the model GNW, the hypothesis is that consciousness (the experience of being aware, integrated, and focused) is manifested by a specific process that we call "*Workspace Ignition*". This process is not a static state, but rather an active and orchestrated transition to a state of high order and coherence in the Workspace, where the information from the modules is integrated into a high-level concept that becomes globally accessible.*

Integrated Information Theory (IIT) defines Φ as a quantitative measure of integrated information. With the detailed mechanisms of GNW defined in the previous sections, we can now draw a robust connection showing how Workspace Ignition mechanically implements the principles of IIT, leading to a maximization of Φ. The formalization of the relationship between Workspace Ignition and Φ establishes a testable hypothesis for the GNW model. The prediction is that the dynamics of transition to an ordered state in the workspace, representing an ignition event, will be accompanied by a significant increase in the integrated information of the system, quantified by Φ. This can provide valuable insights into the physical mechanisms underlying the emergence of states of consciousness and integrated cognition. Let's see how, right now.

Let's consider, as before, the system $\mathsf{GNW} = \cup_{\mathsf{m}=1}^{M} \mathcal{M}_{\mathsf{m}} \cup \mathcal{W}$, which, at time t, is in the global state:

$$\mathbf{s}(t) = (\mathbf{x}(t), \mathbf{y}(t))$$

where $\mathbf{x}(t)$ is the concatenated state of all modules:

$$\mathbf{x}(t) = \left(\mathbf{x}^{(1)}(t), \mathbf{x}^{(2)}(t), \cdots, \mathbf{x}^{(M)}(t)\right)$$

and $\mathbf{y}(t)$ is the state of the workspace. We further assume that each module m has p_{m} stored patterns $\{\boldsymbol{\xi}^{(\mathsf{m},\mu)}\}_{\mu=1}^{\mathsf{p_m}}$, and that WORKSPACE has $\mathsf{p_W}$ stored patterns $\{\boldsymbol{\zeta}^{(\kappa)}\}_{\kappa=1}^{\mathsf{p_W}}$.

The evolution of the system is governed, as before, by the dynamics of Metropolis, influenced by the pseudo-temperature T (or the inverse of the temperature $\beta = 1/T$). The storage capacities, $C_{\mathsf{m}} = p_{\mathsf{m}}/N_{\mathsf{m}}$ and $C_W = \mathsf{p_W}/N_W$, are also relevant parameters.

The "Workspace Ignition", $\mathsf{Ig}_W(t)$, can be seen as a phase transition, where the activity in the workspace becomes coherent, dominating the activity of the local modules, typically associated with the retrieval of one of its memorized patterns. Formally, we can define the Ignition at time t through the overlap of the current state $\mathbf{y}(t)$ with a given memorized pattern, $\boldsymbol{\zeta}^{(\kappa)}$, in the workspace:

$$\mathsf{O}_W^{(\kappa)}(t) = \frac{1}{N_W} \sum_{\mathsf{u}=1}^{N_W} y_{\mathsf{u}}(t) \zeta_{\mathsf{u}}^{(\kappa)} \tag{22}$$

A "Workspace ignition", $\mathsf{Ig}_W(t)$, at time t, occurs when its state $\mathbf{y}(t)$ reaches a significant overlap with at least one of its memorized patterns, and this state is maintained for a minimum period:

$$\mathsf{Ig}_W(t) = \begin{cases} 1 & \text{if there exists } \kappa : |\mathsf{O}_W^{(\kappa)}(t)| > \theta, \quad \text{and} \quad \forall \tau \in [t, t + \Delta t_s], |\mathsf{O}_W^{(\kappa)}(\tau))| > \theta' \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

where:

- θ is an overlap limit (e.g., 0.7) indicating recovery of the κ pattern.
- θ' is a hold limit (can be equal to θ or slightly less).
- Δt_s is the minimum duration for which the overlap must be held, ensuring consistency (e.g., $5 - 10$ time steps).

The transition from $\mathsf{Ig}_W(t - \Delta t) = 0$ to $\mathsf{Ig}_W(t) = 1$ marks the beginning of an ignition event.

Formal Relationship between Workspace Ignition and Φ.

The central hypothesis is that

a "Workspace Ignition", in GNW, is an event actively orchestrated by GNW that mechanically implements the principles of IIT – Differentiation, Integration, and Effective Causality – resulting in a significant peak in Φ*:*

$$lg_W(t) = 1 \implies \Phi(t) \text{ high} \tag{24}$$

It is therefore expected that the value of $\Phi(t)$ will be significantly higher when the Workspace is in an ignition state. During an ignition event, $\mathsf{lg}_W(t) = 1$, we have the following:

1. The workspace transitions from a disordered state to a state of high overlap ($|\mathrm{O}_W^{(\kappa_t^*)}(t)| > \theta$), to a specific pattern $\zeta^{(\kappa_t^*)}$ and, therefore, *highly differentiated* from other possible states.

2. The transition to this ordered state of the workspace is driven by a *strong integration* of relevant inputs from the modules, and/or by the internal dynamics of the workspace itself. This implies a *strong causal interdependence*.

3. A workspace in an ignition state is causally effective. Its current state exerts a strong influence on its future state and is strongly constrained by its past state (like a robust attractor). The ignition event can be a moment where the coupling between the modules and the workspace is optimized, or where the temperature allows for stable pattern recovery.

All these factors (differentiation, integration, effective causality) contribute to a high value of Φ, which is the essence of IIT. To test this hypothesis, we can adopt the following procedure:

1. Run Monte Carlo simulations of GNW for a long period $\mathbb{T}$.

2. Store the complete time series of the states of all modules, $\{\mathrm{x}_t^{(\mathsf{m})}\}_{\mathsf{m}=1}^{M}$, and the Workspace, y_t, for $t = 1, \cdots, \mathbb{T}$. For each t, calculate $\mathrm{O}_W^{(\kappa)}(t)$, for all patterns $\zeta^{(\kappa)} : \kappa = 1, \cdots, \mathsf{p_W}$, and determine $\mathsf{lg}_W(t)$ based on the limits θ, θ' and Δt.

3. For each t, calculate $\Phi(t)$ using the formula (15) from the 3 section, using a sliding window of size w centered on t, to obtain instantaneous values.

4. Calculate the correlation coefficient between the time series $\mathsf{lg}_W(t)$ and $\Phi(t)$. Compare the distributions of $\Phi(t)$ for the states $\mathsf{lg}_W(t) = 1$ and $\mathsf{lg}_W(t) = 0$. Display $\Phi(t)$ and $\mathsf{lg}_W(t)$, "plotted" against time, to identify peaks of Φ during ignition.

7 Recap of the Active Mechanisms of GNW in Pattern-Driven Ignition

GNW operates with an interactive feedforward and feedback loop, actively guided by patterns [JNT2025a]:

1. Modules generate average activities, $A_{\mathsf{m}}(t)$, that compete with each other. The contribution of each module m, to this competition, is measured by $\mathsf{Comp}_{\mathsf{m}}(t)$. The competition function $\mathcal{F}_{\mathsf{m}}(t)$ amplifies the most relevant modules.

2. With this prioritized information, the external field $h^{\mathsf{mod}}_{W,\mathsf{u}}(t)$ is formed, which modifies the energy landscape of the workspace.

3. The workspace evaluates, through the overlaps $\mathrm{O}^{(\kappa)}_{W}$, how well its current state, $\mathrm{y}(t)$, aligns with its memorized patterns, $\boldsymbol{\zeta}^{(\kappa)}$.

4. The outer field $h^{\mathsf{guide}}_{W,\mathsf{u}}(t)$, actively "pulls" the workspace toward the dominant pattern, $\boldsymbol{\zeta}^{\kappa^*_t}$, representing an internal "intention" or "hypothesis".

5. When the workspace begins to converge to the $\boldsymbol{\zeta}^{\kappa^*_t}$ pattern, it sends feedback, $h^{\mathsf{fb}}_{\mathsf{m},i}(t)$, to the modules, directing them to be consistent with the integrated concept.

Pattern-Driven Ignition: The Realization of the Principles of IIT

"*Pattern-Driven Ignition*" is defined as the state where the workspace transitions to a globally ordered phase, with $\langle \mathsf{o}_{\mathsf{m}} \rangle >> 0$ and $>> 0$, sustaining the pattern $\boldsymbol{\zeta}^{\kappa^*_t}$ for a period of time, with the active steering mechanisms operational. This dynamic event mechanically implements the principles of IIT:

1. Ignition, by definition, is the convergence of the workspace to a specific and sustained pattern $\boldsymbol{\zeta}^{\kappa}$. This pattern is a unique configuration of neurons, distinguishable from all other memorized patterns and chaotic states.

2. During ignition, $\langle \mathsf{o}_{\mathsf{m}} \rangle >> 0$ for the pattern $\boldsymbol{\zeta}^{\kappa}$ and $>> 0$. This quantifies the clarity and distinctness of the workspace state, indicating that it is in an ordered phase.

3. The $\mathcal{F}_{\mathsf{m}}(t)$ competition ensures that the inputs from the modules to the Workspace are equally differentiated, providing high-quality "raw material" for the formation of the integrated concept.

Integration in GNW during Ignition

Integration is the heart of the active *feedforward-feedback* loop and the external fields.

1. The field $h^{\text{mod}}_{W,\text{u}}(t)$ in the workspace is the first layer of integration, selectively selecting and combining the differentiated information (coming from the modules in the ordered phase).
2. The active pull, $h^{\text{guide}}_{W,\text{u}}(t)$, actively integrates these modulated inputs, pulling its own state towards one of its patterns $\boldsymbol{\zeta}^{(\kappa)}$. This is the act of "unifying" the diverse information into a single high-level concept.
3. Feedback Driven by patterns, $h^{\text{fb}}_{\text{m},i}(t)$. Workspace feedback (carrying the integrated concept $\boldsymbol{\zeta}^{(\kappa)}$ back to the modules, coordinating their activity) reinforces the overall cohesion and causal interdependence of all parts, solidifying integration.

Effective Causality in GNW during Ignition

The feedforward-feedback cycle and active pattern attraction establish a robust causality complex.

1. The workspace (driven by $h^{\text{mod}}_{W,\text{u}}$ and $h^{\text{guide}}_{W,\text{u}}$) causes its own transition to an ordered state – the pattern $\boldsymbol{\zeta}^{(\kappa^*_t)}$.
2. Once in this ordered state, the workspace causes the Coordination of modules via feedback, $h^{\text{fb}}_{\text{m},i}(t)$, restricting their future dynamics.
3. The status of GNW is strongly determined by its internal patterns and the way it selects and integrates inputs (past), and, in turn, strongly restricts its own future. This is the essence of intrinsic causality.

The central hypothesis is that the "Workspace Ignition" in GNW is the dynamic moment that represents the maximized implementation of IIT principles, and therefore correlates directly with high values of $\boldsymbol{\Phi}$.

1. Ignition as a Phase Transition to Consciousness. Ignition is the transition to an ordered phase (high $\langle\rangle$, $>> 0$), actively orchestrated, which represents a state of clear, focused, and integrated cognition.
2. GNW and $\boldsymbol{\Phi}$ Mechanisms. All external fields $h^{\text{guide}}_{W,\text{u}}$, $h^{\text{fb}}_{\text{m},i}$, and the competition functions ($\mathcal{F}_{\text{m}}$) work together during Ignition to maximize differentiation (forming a pattern $\boldsymbol{\zeta}^{(\kappa^*_t)}$, integration (feedforward-feedback loop) and effective causality (pattern guidance), which are the ingredients for a high $\boldsymbol{\Phi}$.

Research Methodology

To test this hypothesis, the following procedure is proposed:

1. GNWSimulation. Run Monte Carlo simulations of the GNW for a long period $\mathbb{T}$, recording the states and values of the external fields.
2. Phase Mapping. Identify the ordered, confused, and disordered phase regions depending on the control parameters (e.g., $\mathcal{C}$, β, λ) using $\langle \mathsf{o}_\mathsf{m} \rangle$ and q_m.
3. Ignition Detection. For each t, determine $\mathsf{Ig}_W(t)$ based on the overlap thresholds, $\mathsf{O}_W^{(\kappa_t^*)} > \theta_{\mathsf{ign}}$, and lift, and the operation of the active steering mechanisms.
4. Calculation of $\Phi(t)$. For each t, calculate $\Phi(t)$ (via TE and RTE in a sliding window).
5. Detailed Correlation Analysis. Correlate ignition events with $\Phi(t)$ peaks; verify if the $\Phi(t)$ peaks and ignition events occur predominantly in the ordered phase of the GNW, and investigate if $\Phi(t)$ is maximized in the critical phase transition region.

In conclusion: the formulation of the relationship between Workspace Ignition and Φ demonstrates how the intricate mechanisms of the GNW model, including modular competition, bidirectional connectivity, and the active role of patterns, align with the principles of IIT. Ignition is conceived as an event of active orchestration that results in a state of high differentiation, integration, and effective causality, being a strong candidate to correspond to a peak of Φ, and, hypothetically, to the emergence of a conscious state.

8 Final Abstract

This work explores the dynamics of neuronal ignition in the Global Workspace model, using a statistical mechanics approach with stochastic Hopfield networks. Building on the previous work [JNT2025a], a model is proposed where local modules and the workspace are stochastic Hopfield networks, interacting through connectivity and feedback. The study investigates how stochasticity, network capacity, and information integration influence the ignition phase transition. Integrated Information Theory (IIT) is used to quantify the global coherence of the system, analyzing the role of Transfer Entropy between modules and workspace, and in the calculation of Φ. The results provide insights into the mechanisms underlying consciousness and its emergence.

The "resolution" (even if approximate) of the Hard Problem of Consciousness, through a computational model like the one proposed here (GNW+ IIT), is

a subject of profound philosophical and scientific debate. Clearly, this model does not solve it, but, however, it can offer significant advances for a deeper understanding.

The "Hard Problem of Consciousness" (David Chalmers) does not consist of explaining how the brain processes information, decides, or generates behaviors (the "Easy Problems"). The problem is to explain why, and how, subjective experience (the qualia), the "feeling of being" something, emerges from physical processes. It is the explanatory gap between the physical and the phenomenological.

How could the GNW+ IIT model then contribute to this discussion and what are its limitations? Here are some "hints":

1. Providing Quantifiable Neuronal Correlates. The GNW model, with its Hopfield networks and workspace ignition, offers a mechanistic and simulable neuronal substrate. IIT provides a quantifiable measure, Φ, of the information integrated into this substrate. If we find that "workspace ignition" (a dynamic event in GNW) corresponds to a peak of Φ, we will have a highly sophisticated and theoretically grounded neuronal correlate of consciousness (NCC).

   In this way, we identify a complex physical process (high Φ in a dynamic GNW) that coincides with what we presume to be a conscious state. But the Hard Problem asks why this complex physical process generates the subjective experience, and not just what this process is! Therefore, this study cannot be conclusive.

2. IIT, as a (computational) theory of consciousness, makes predictions about the properties of conscious systems (high Φ). The GNW model can be a "test lab" for these predictions. We can vary parameters (coupling forces, temperature, etc.) and see how the dynamics of GNW (ignition), and the resulting Φ, behave. If Φ correlates with ignition and behaves consistently with other IIT predictions, this strengthens the theory.

   However, even if IIT is the "correct theory" of how consciousness arises from the physical, it is still a functionalist theory at its core. She describes the structure of experience (its differentiation and integration) in terms of its underlying causality. This does not explain the leap to phenomenological character.

3. The GNW model can reveal patterns of dynamic and emergent information integration that are not obvious. "Ignition" may be an elegant mechanism for the formation of a transient "field of consciousness". Studying how ignition modulates the Phi may provide insights into how the GNW architecture can generate integrated processing states.

   This expands knowledge about information integration, which is a crucial aspect of consciousness, but not necessarily the Hard Problem itself. It

may be that subjective experience is a "byproduct" or an "emergent property" of this integration, but the explanation of how this emergence occurs has not been given.

4. IIT attempts to "solve" the Hard Problem by stating that consciousness is integrated information. If a system has $\Phi > 0$, it has consciousness. It is not that consciousness emerges from Φ, but rather that Φ is consciousness. This is the IIT's "solution," and the GNW model helps to test it mechanistically.

   Many philosophers (and scientists) do not consider this a satisfactory solution to the Hard Problem. For them, the IIT is "dissolving" the problem, redefining consciousness as integrated information, instead of explaining why integrated information generates experience. In short: "Why does integrated information 'feel' something?"

This work seeks to be a useful contribution to the science of consciousness and to integrated information theory, by validating and refining IIT in a dynamic computational model context; proposing a plausible mechanism for the emergence of highly integrated (ignition) states in neural networks; and providing a quantifiable framework for studying the relationship between neural dynamics and consciousness metrics.

However, the "approximate resolution" of the Hard Problem remains an open and largely philosophical question. This work can, at most, explain what consciousness is in terms of integrated information and how it manifests dynamically in the GNW model, proposed here, with Hopfield Networks. Why and how this generates subjective experience, the fundamental explanatory gap, persists.

The hypothesis that consciousness is the emergence of an ordered global state, resulting from a phase transition (exploitable with the statistical mechanics of the GNW model, and measurable with Φ), is a scientific approach expected to be useful and promising for advancing our understanding of the neuronal and informational mechanisms of consciousness. This hypothesis does not eliminate the Hard Problem from philosophy, but transforms it into an empirically constrained question, focused on the gap between the mechanistic description (however complete it may be) and the phenomenological experience. If it does this, it brings us closer to a more complete understanding of what it means to be conscious.

On Neuronal Correlates of Consciousness (NCCs)

Neuronal Correlates of Consciousness () are the neuronal events – dynamic processes and minimal mechanisms, necessary and sufficient for a specific conscious experience to occur.

In the model GNW, as developed here, modules are specialized units that store patterns (sensory memories, concepts). They process information in a

segregated way, acting as ”specialists” for different modalities/domains. They compete for access to the workspace, providing the ”raw material” of differentiated information.

Consciousness is characterized by unity and integration. When we see a red, round, ripe apple, we don't have three separate experiences (”red,” ”round,” ”ripe”). We have a single integrated experience of ”apple.” A single module, being specialized, cannot handle this integration. Global theories (such as GWT) suggest that consciousness emerges when information becomes globally available to the system. Modules, by definition, are local. IITseeks the complex of integrated information that generates Φ. An isolated module can generate its own Φ, but that would be the The Φ of a part of the experiment, not of the unified experience as a whole.

An activated module, when processing a specific pattern (for example, the visual module activated by the ”red” pattern), could be the neuronal correlate of the ”red” content. This content may or may not become conscious depending on how it is integrated by the workspace. Many processes that occur in the modules can be considered ”pre-conscious.” They are necessary for perception, but perception only becomes conscious when the information is integrated and ”diffused” into the workspace. Modules are essential for the differentiation of experience. Without them, the workspace would have less information to integrate.

The NCC for a unified and global conscious experience would most likely be the dynamic and integrated state of the workspace, in conjunction with the relevant modules that actively feed it. It would be the ”complex” formed by the workspace in its ”ignition” state, acting as an integration center and making the Information

Globally available and accessible to the system. This ”complex” (Workspace + active modules) would be the physical substrate that generates the maximum Phi for that specific experiment.

The phase transition to a globally ordered state (involving the workspace and possibly the coordination of several modules) would be the dynamic event that constitutes the NCC.

The modules of the GNW model are indispensable for consciousness. They are the essential components that provide differentiated information (the ”content correlates” or ”preconscious”). However, the NCC for unified conscious experience, as we understand it in GNW or IIT, would be better described as the integrated system that emerges from the dynamic interaction between the workspace and the active modules, especially during an ”ignition” event, which corresponds to a phase transition to a globally ordered and high state. Φ.

9 Appendix. Mathematical Preliminaries.

Information Theory: Entropy and Mutual Information

Information Theory, developed by Claude Shannon in 1948 [**?**], aims to understand and quantify information, through a rigorous mathematical formalism.

The central concept is that of Entropy, a measure of uncertainty or disorder associated with a random variable, possibly multidimensional. Unlike thermodynamic entropy, which describes the physical disorder of a system, entropy in Information Theory quantifies the average amount of information needed to describe the value of a variable.

To define entropy, we begin by defining a measure of information, such that an unexpected observation has a higher information content than an expected observation (surprise effect). In particular, a certain event will have zero information content. If we observe an event that has a probability of occurring equal to p, we associate with that event the amount of information:

$$\mathsf{Inf}(p) = \log_2 \frac{1}{p}$$

such that $\mathsf{Inf}(p) \to 0$, when $p \to 1$ (generally the $\log$ is taken in base 2).

Shannon Entropy.

1. A Shannon entropy is a measure of the average uncertainty, or average information, of a random variable (multidimensional). *The greater the entropy, the greater the uncertainty and the amount of average information.*

2. For a discrete random variable X, with possible values x_i, and corresponding probabilities $P(X = x_i) = p_i$, with $\sum_i p_i = 1$, the Shannon entropy is defined by:

$$\begin{aligned} \mathsf{Ent}(X) &= \sum_i p_i \mathsf{Inf}(p_i) \\ &= \sum_i p_i \log_2 \frac{1}{p_i} \\ &= -\sum_i p_i \log_2 p_i \end{aligned} \tag{25}$$

   where the base of the logarithm (typically 2) determines the unit of the information (bits).

3. Example. Let X be a binary random variable with possible values $\{-1, +1\}$, with probabilities $P(X = +1) = p$ and $P(X = -1) = 1 - p$, with $0 \le p \le 1$. Then

$$\mathsf{Ent}(X) = (+1) \log_2 p + (-1) \log_2 (1 - p)$$

This entropy, as a function of p, is called the binary entropy function, and is illustrated in Fig. 13. As the figure shows, $\mathsf{Ent}(p)$ is maximized for a uniform distribution (i.e., for $p = 1/2$).

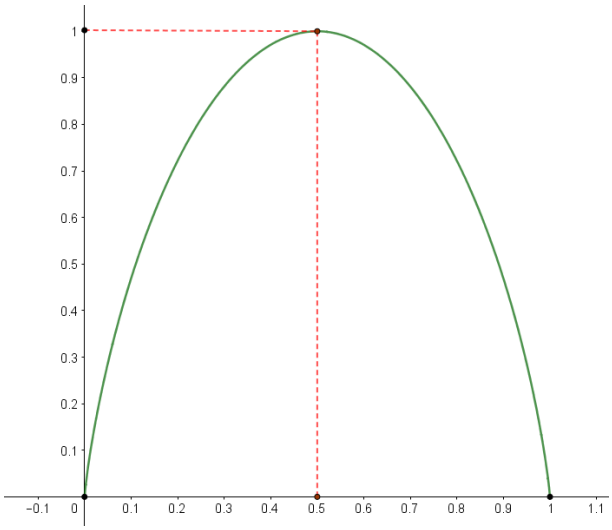


Figure 13: Binary entropy $p \longmapsto \mathsf{Ent}(p)$.

Note. Two ways to view the same concept:

1. Information such as "*reduction of observer uncertainty*": when the entropy of the system decreases (goes from chaotic to ordered), the observer (or the system itself processing the input) gains information about the state of the system.

2. Information such as "*surprise content of an event*": if a system has low entropy, the events it produces are more predictable and therefore each event contains less information (surprise) on average. Both perspectives are valid and important in Shannon's Information Theory. But in the GNWmodel, considered here, the most relevant is the first: the system "*gains information*" about the input by reducing its internal uncertainty, based on a low entropy state (a memory/representation).

Joint and Conditional Entropy.

1. A JOINT ENTROPY of two random variables X and Y measures the uncertainty of the pair (X, Y). It is defined as:

$$\mathsf{Ent}(X, Y) = -\sum_{x,y} P(x, y) \log_2 P(x, y) \tag{26}$$

where $P(x, y)$ is the joint probability distribution of X and Y.

2. The CONDITIONAL ENTROPY of X given Y, measures the uncertainty about X, when the value of Y is known. It is defined as:

$$\mathsf{Ent}(X|Y=y) = -\sum_x P(x|y)\log_2 P(x|y) \tag{27}$$

where $P(x|y)$ is the conditional probability of X given $Y=y$. Taking the average in relation to Y, we obtain:

$$\begin{aligned}\mathsf{Ent}(X|Y) &= \sum_y P(y)\mathsf{Ent}(X|Y=y)\\ &= -\sum_y P(y)\sum_x P(x|y)\log_2 P(x|y)\\ &= -\sum_x\sum_y P(x,y)\log_2 P(x|y)\end{aligned} \tag{28}$$

Joint entropy can be expressed in the form:

$$\mathsf{Ent}(X,Y) = \mathsf{Ent}(X) + \mathsf{Ent}(Y|X) = \mathsf{Ent}(Y) + \mathsf{Ent}(X|Y) \tag{29}$$

3. Example. Consider two variables: Y (= Time $\in$ {"sun", "rain"}) and X (= Clothes $\in$ {"jacket", "no coat"}). If we know it's raining, our uncertainty about the clothes the person is wearing is reduced. Therefore, $\mathsf{Ent}(X|Y)$ is less than $\mathsf{Ent}(Y)$.

4. $\mathsf{Ent}(X|Y) \leq \mathsf{Ent}(X)$. Information about Y decreases the uncertainty about X, with equality if and only if X and Y are independent. In other words, "conditioning" reduces entropy.

5. Entropy is additive for independent variables; That is,

$$\mathsf{Ent}(X,Y) = \mathsf{Ent}(X) + \mathsf{Ent}(Y) \tag{30}$$

for independent X and Y.

Mutual Information DEFINITION. Mutual information between two random variables X and Y measures how much knowledge of one of these variables reduces the uncertainty about the other.

For example, if X and Y are independent, then knowing X provides no information about Y and vice versa. Therefore, their mutual information will be zero. At the other extreme, if X is a deterministic function of Y, and Y is a deterministic function of X, then all information contained in X is shared by Y: knowing X determines the value of Y and vice versa. As a result, in this case the mutual information is the same as the uncertainty contained in Y (or X) alone.

Mutual information is defined by:

$$\mathsf{Inf}(X;Y) = \mathsf{Ent}(X) - \mathsf{Ent}(X|Y) = \mathsf{Ent}(Y) - \mathsf{Ent}(Y|X) = \mathsf{Ent}(X) + \mathsf{Ent}(Y) - \mathsf{Ent}(X,Y)$$

and, in terms of probabilities, is defined by:

$$\mathsf{Inf}(X;Y) = \sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{30}$$

This expression is a particular example of the Kullback-Leibler divergence, a measure of the difference in information between dependence and independence.

Example: If X represents the presence of a disease and Y represents the result of a diagnostic test, the mutual information $\mathsf{Inf}(X;Y)$ measures how much the test result reduces the uncertainty about the presence of the disease.

Properties:

1. $\mathsf{Inf}(X;Y) = \mathsf{Inf}(Y;X) = \mathsf{Ent}(X) - \mathsf{Ent}(X|Y) = \mathsf{Ent}(Y) - \mathsf{Ent}(Y|X)$.
2. $\mathsf{Inf}(X;Y) = \mathsf{Ent}(X) + \mathsf{Ent}(Y) - \mathsf{Ent}(X,Y)$.
3. $\mathsf{Inf}(X;Y) \leq \mathsf{Ent}(X)$ with equality valid if and only if X is a function of Y, that is, $X = f(Y)$ for some function $f(\cdot)$.
4. $\mathsf{Inf}(X;Y) \geq 0$, with equality if and only if X and Y are independent.

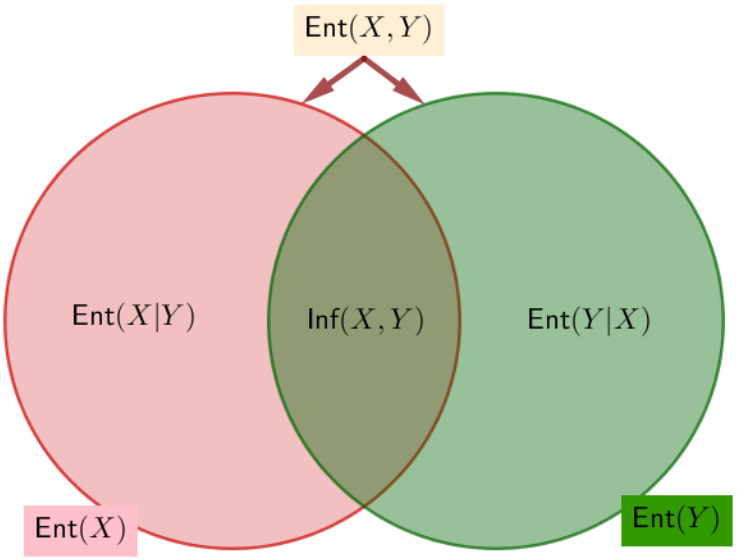


Figure 14: Relationship between entropy and mutual information.

5. Conditional mutual information:

$$\mathsf{Inf}(X;Y|Z) = \mathsf{Ent}(X|Z) - \mathsf{Ent}(X|Y,Z) \tag{31}$$

Entropy Transfer (TE)

Given discrete (for simplicity), jointly distributed, and possibly multivariate random variables X and Y, we have seen that mutual information $\mathsf{Inf}(X;Y)$ provides an intuitive and well-founded answer to the following questions:

- How much does the uncertainty about the state of Y decrease when you know the state of X (and vice versa)?
- How much information is shared between X and Y?
- How can we quantify the degree of statistical dependence between X and Y?

Now suppose that instead of static variables, we have random time series X_t, Y_t, where $t = 0, 1, 2, 3, \cdots$.
How, then, can we interpret and answer questions comparable to the previous ones for dynamic stochastic processes instead of static variables?

We can, of course, consider the mutual information $\inf(X_t; Y_t)$ between variables X_t and Y_t, at each fixed time t. But note that, for stochastic processes, there may be dependencies between individual variables.

Thus, for example, X_t, (the variable X observed at time t), may have a statistical dependence on its value X_{t-s}, at the previous instant $t - s$, or even on its entire past or the past of Y.

A particularly attractive notion is to quantify a temporally oriented transfer or flow of information between variables. Thus, we could try to answer the question:

- *How much information is transferred from the current state of X to the future (immediate) state of Y (and vice versa)?*

This information transfer, which in principle will be asymmetrical in X and Y, unlike the synchronous mutual information $\mathsf{Inf}(X_t; Y_t)$, is quantified by Entropy Transfer (TE).

Definition. Information theory and the notion of entropy, are the basic concepts for formalizing the notion of TE. The key idea is as follows:

- $\mathsf{TE}(X_t \to Y_{t+1})$, with $\mathsf{lag} = 1$, can be intuitively interpreted as the decrease in the degree of uncertainty about the immediate future of Y, Y_{t+1}, caused by the current state of X and Y: $\mathsf{Ent}(Y_{t+1}|Y_t, X_t)$, in addition to the decrease in the degree of uncertainty about the immediate future of Y, Y_{t+1}, caused only by its own current state Y_t: $\mathsf{Ent}(Y_{t+1}|Y_t)$.

  Symbolically:

$$\begin{aligned} \mathsf{TE}(X_t \to Y_{t+1}) &= \mathsf{Ent}(Y_{t+1}|Y_t) - \mathsf{Ent}(Y_{t+1}|X_t, Y_t) \\ &= \mathsf{Inf}(Y_{t+1}; Y_t|X_t) \end{aligned} \tag{32}$$

If $\mathsf{TE}(X_t \to Y_{t+1}) > 0$, then X influences the (immediate) future of Y.

If $\mathsf{TE}(X_t \to Y_{t+1}) = 0$, then X provides no additional information to predict the (immediate) future of Y, given its own current state.

In particular, $\mathsf{TE}(X_t \to Y_{t+1}) = 0$ if and only if Y, conditioned by its own current state, is independent of the current state of X.

Additional Notes.

1. Entropy Transfer (TE) is a non-parametric measure that quantifies the directional transfer of information between stochastic processes. Unlike correlation measures, TE captures the asymmetry in information transfer, providing insights into observational causality.

2. At a conceptual level, TE is a model-free implementation of Wiener's principle of observational causality, which states:

   For two processes X, Y observed simultaneously, we call X "causal" to Y if knowledge about the current state of X improves our prediction of the future of Y, beyond what is predictable from the current state of Y.

   In terms of information theory: "*How much information does the current state of process X contain about the future observation of a value of Y, assuming that we know the current state of Y?*"

3. Let's restrict ourselves here to the immediate future:

   $$\mathsf{TE}(X_t \to Y_{t+1}) = \mathsf{Inf}(Y_{t+1}, X_t|Y_t)$$

   although deeper studies could justify the consideration of higher-order "lags" (as in Markov processes).

4. It is crucial to interpret TE as a *measure of predictive influence*, and not as a measure of causality in the strong sense. TE quantifies the ability of a process to predict the future of another, but does not necessarily imply a direct cause-and-effect relationship.

The TE can be seen as the decomposition of the information needed to predict the future state of a variable into two components:

1. Stored information. The information about the future state of Y, Y_{t+1}, that is already present in its current state Y_t.

2. Transferred information. The additional information about the future state of Y, Y_{t+1} that is provided by the current state of X, X_t, and that is not contained in the current state of Y, Y_t.

EXAMPLE: Consider two processes X and Y that are influenced by a common factor Z (latent variable). Even if X does not directly cause Y, the TE from X to Y can be high due to the common influence of Z. This scenario is known as "spurious causality".

EXPLANATORY NOTE: The equation 32, compares two conditional probabilities: the probability of predicting the future of Y, Y_{t+1}, using only its current state ($p(Y_{t+1} = y_{t+1}|Y_t = y_t)$) and the probability of predicting the future of Y using its current state and the current state of X ($p(Y_{t+1} = y_{t+1}|Y_t = y_t, X_t = x_t)$).

Illustrative Example: Let's imagine that X represents neuronal activity in the visual cortex and Y represents activity in the parietal cortex. A high TE from X to Y suggests that visual information processed in the visual cortex influences spatial processing-related activity in the parietal cortex.

Conceptual Illustration: Let's imagine we are trying to predict tomorrow's temperature, Y_{t+1}. We can use today's temperature, Y_t, as a first estimate. The TEallows us to quantify how much additional information about tomorrow's temperature is provided by today's humidity, X_t, beyond what we already know about today's temperature. If today's humidity allows us to refine our forecast of tomorrow's temperature, then $\mathsf{TE}(X_t \to Y_{t+1}) > 0$, indicating a transfer of information from X to Y.

Practical Estimation of Transfer Entropy and Calculation of Transfer Entropy.

1. DISCRETIZATION. If the variables X and Y are continuous, it is necessary to discretize them into K discrete bins. This is not our case, where the variables are binary ($\mathsf{p_m}1$) and, therefore, discrete.

2. ESTIMATING PROBABILITIES. From the observed data, estimate the joint and conditional probability distributions, counting the occurrences of each combination of states:

$$P(y_{t+1}, y_t, x_t) \approx \frac{\mathsf{Cont}(y_{t+1,y_t,x_t})}{\mathsf{Total\ Samples}}$$

$$P(y_{t+1}|y_t, x_t) \approx \frac{P(y_{t+1}, y_t, x_t)}{P(y_{t+1}, x_t)}$$

$$P(y_{t+1}t|y_t) \approx \frac{\mathsf{Cont}(y_{t+1}, y_t)}{\sum_{y_t} \mathsf{Cont}(y_{t+1}, y_t)}$$

3. TRANSFER ENTROPY CALCULATION. Substitute the probability estimates into the Transfer Entropy formula and calculate the sum.

Reverse Entropy Transfer (RTE)

RTE is a tool for inferring causal relationships in complex systems, and its definition is relatively straightforward.

Definition. The RTE quantifies how much knowledge of the state of X, at time t, reduces the uncertainty about the state of Y, at time $t-1$, conditioned on knowledge of the state of Y at time t.

In other words, it measures how much X_t helps to "reconstruct" or restrict the possible past causes of Y. The mathematical definition is as follows:

$$\mathsf{RTE}(X_t \to Y_{t-1}) = \mathsf{Ent}(Y_{t-1}|Y_t) - \mathsf{Ent}(Y_{t-1}|X_t, Y_t) \tag{33}$$

where $\mathsf{Ent}(Y_{t-1}|Y_t)$ is the conditioned entropy of Y, at time $t-1$, given knowledge of the state of Y, at time t. It represents the uncertainty we still have about the past state of Y (at $t-1$) after observing its present state (at t). It quantifies how much the present state of Y tells us about its past.

$Ent(Y_{t-1}|X_t, Y_t)$ is the conditional entropy of Y at time $t-1$, given the joint knowledge of the state of X and the state of Y, both at time t. It represents the remaining uncertainty about the past state of Y after observing both the present state of Y and the present state of X. It quantifies how much the knowledge of X_t and Y_t together tells us about the past of Y. The equation $\mathsf{RTE}(X_t \to Y_{t-1})$ measures the reduction in uncertainty about Y_{t-1} that is obtained by adding the knowledge of X_t to the already existing knowledge of Y_t.

If $\mathsf{RTE}(X_t \to Y_{t-1}) > 0$ then the knowledge of X_t reduces the uncertainty about Y_{t-1} beyond what is already provided by the knowledge of Y_t. This suggests that X_t contains relevant information about the past causes of Y. If $\mathsf{RTE}(X_t \to Y_{t-1}) = 0$, then the knowledge of X_t does not provide any additional information about Y_{t-1} beyond what is already provided by the knowledge of Y_t. This suggests that X_t is not causally related to the past of Y (or that the relationship is too weak to be detected).

Analogy. Think of Y_{t-1} as the crime to be solved, X_t as new evidence discovered in the present, and Y_t as the prior knowledge the detective already has about the case. $\mathsf{Ent}(Y_{t-1}|Y_t)$ represents the uncertainty the detective still has about the crime before analyzing the new evidence X_t.

$\mathsf{Ent}(Y_{t-1}|Y_t, X_t)$ represents the uncertainty the detective has about the crime after analyzing the new evidence X_t. $\mathsf{RTE}(X_t \to Y_{t-1})$ represents how much the new evidence X_t helped the detective to narrow down possible explanations for the crime. If RTE is high, it means that the evidence was very useful in solving the case.

Practical Calculus:

1. Discretization: If the variables X and Y are continuous, it is necessary to discretize them into a finite number of states.

2. Estimating probabilities joint and conditioned from the observed data:

- $P(X_t = x_t, Y_t = y_t, Y_{t-1} = y_{t-1})$,
- $P(Y_{t-1} = y_{t-1}|Y_t = y_t)$,
- $P(Y_{t-1} = y_{t-1}|X_t = x_t, Y_t = y_t)$

Calculate the conditional entropies using the probability estimates:

- $\mathsf{Ent}(Y_{t-1}|Y_t) = -\sum P(y_t) \sum P(y_{t-1}|y_t) \log_2 P(y_{t-1}|y_t)$
- $\mathsf{Ent}(Y_{t-1}|X_t, Y_t) = -\sum P(x_t, y_t) \sum P(y_{t-1}|x_t, y_t) \log_2 P(y_{t-1}|x_t, y_t)$, obtained by counting.
- Calculate RTE using the formula above.

Important Considerations:

1. Sampling: accurate estimation of probabilities requires a significant amount of data.
2. Discretization: the choice of discretization can affect the results.

   Interpretation: the RTE does not necessarily imply causality in the strong sense. It only indicates that X_t contains relevant information about the past of Y. The relationship may be indirect or mediated by other variables.

Statistical and Temporal Means

Consider an observable $A : \mathcal{S} \to \mathbb{R}$, defined in the state space $\mathcal{S}$ of a stochastic Hopfield network, with Boltzmann probability.

The statistical mean of A is calculated over all possible states of the system, weighted by their respective probabilities:

$$\langle A \rangle_{\mathsf{est}} = \sum_{\mathbf{x} \in \mathcal{S}} P(\mathbf{x}) A(\mathbf{x})$$

It represents the average value of the observable, considering all possible configurations of the system.

The time mean is calculated along a stochastic time trajectory of the system. The trajectory is generated by the stochastic dynamics of the network:

$$\langle A \rangle = \lim_{T \to \infty} \frac{1}{\mathbb{T}} \sum_{t=1}^{\mathbb{T}} A(\mathbf{x}(t))$$

where $\mathbf{x}(t)$ is the state of the system at time t, $\mathbb{T}$ is the time duration of the trajectory. In practice, we approximate this average with a finite time $\mathbb{T}$:

$$\langle A \rangle \approx \frac{1}{\mathbb{T}} \sum_{t=1}^{\mathbb{T}} A(\mathbf{x}(t))$$

This average represents the average value of the observable over time, following the evolution of the system.

Let's now look at the role of Ergodicity:

- A system is ergodic if, over time, it explores its entire space of accessible states uniformly.

    1. Ergodicity is the property that ensures that the temporal average is equal to the statistical average. Formally:

$$\langle A \rangle_{\text{est}} = \langle A \rangle \approx \frac{1}{\mathbb{T}} \sum_{t=1}^{\mathbb{T}} A(\mathbf{x}(t)) \tag{34}$$

    2. Ergodicity in the Model GNW. If the Hopfield stochastic networks (in the modules and in the workspace) are well designed (e.g., with an appropriate noise level), then ergodicity can be a reasonable approximation. The noise (stochasticity) in the network helps to explore the state space and avoid getting stuck in unrepresentative local minima.

    If the network is not ergodic (e.g., at low temperatures, or with very strong interactions), then the time average may not be equal to the statistical average; the simulation of a single time trajectory may not provide an accurate picture of the system's behavior.

How can we practically assess whether ergodicity is a good approximation? Let's see: Convergence of Means: monitor the convergence of the time averages. If the averages stabilize after a certain period of time, this suggests that ergodicity may be a good approximation.

Sensitivity to Initial Conditions: run multiple simulations with different initial conditions and check if the results are similar. If the results depend strongly on the initial conditions, this may indicate that ergodicity is not valid.

Implications for the Calculation of Phi: The calculation of Transfer Entropy and integrated information requires the estimation of probabilities, for example, of the type:

P($\mathbf{x}_{t+1}, x_t, y_t), P(x_{t+1}|x_t, y_t), etc.$

These probabilities are estimated by counting the frequency with which different combinations of states occur over time.

Ergodicity ensures that the frequency with which each combination of states occurs in the time trajectory is representative of its probability in the state space as a whole. Ergodicity is therefore a crucial assumption that allows the use of time averages as an approximation for statistical averages, making it possible to analyze the behavior of the GNW system with computational simulations. The validity of this approximation must be verified to ensure the accuracy of the results.

How to estimate the probability $P(A(\mathbf{x}(t)))$ with sampling?

$P(A(\mathbf{x}(t)))$ represents the probability that the observable A has a certain value when the system is in the state $\mathbf{x}(t)$ at time t.

Examples: (i). $A = \delta_i$. In this case, $\delta_i(\mathbf{x}(t)) = x_i(t)$, and $P(x_i(t)))$ represents the probability that neuron i has a given value ($\mathsf{p_m}1$), when the system is in state $\mathbf{x}(t)$ at time t. (ii). $A = \mathsf{Ov}$, the overlap of the current state of the system with a memory pattern $\boldsymbol{\xi}^{(\mu)}$.

Sampling Approach

1. Simulation: Simulate the stochastic Hopfield network for a long period of time T.
2. Data Collection: At each time step t, record the state of the system $\mathbf{x}(t)$ and the value of the observable $A(\mathbf{x}(t))$.
3. Observable Discretization (if necessary): If the observable $A(\mathbf{x})$ is continuous, discretize its range of values ??into K bins.
4. Occurrence Counting: Count how many times the observable $A(\mathbf{x}(t))$ falls into each bin during the simulation.
5. Probability Estimation: Normalize the counters to obtain an estimate of the probabilities.

Mathematical Formalization

1. Define a counter $C(k)$ for each bin k. At each time step t, determine in which bin $A(\mathbf{x}(t))$ falls and increment the corresponding counter. Define

$$C(k) = \sum_{t=1}^{T} \mathbb{I}(A(\mathbf{x}(t)) \in (a_k, a_{k+1})$$

where $\mathbb{I}$ is the indicator function:

$$\mathbb{I}(A(\mathbf{x}(t)) \in (a_k, a_{k+1}) = \begin{array}{ll} 1 & \text{if } A(\mathbf{x}(t)) \in (a_k, a_{k+1} \\ 0 & \text{otherwise} \end{array}$$

2. Probability Estimation. Estimate the probability of $A(\mathbf{x}(t))$ being in bin k as:

$$P(A(\mathbf{x}(t)) \in (a_k, a_{k+1}]) \approx \frac{C(k)}{T}$$

Average values. The statistical mean of a variable $A : \mathcal{S} \to \mathbb{R}$, is given by:

$$\langle A \rangle_{\mathsf{est}} = \sum_{\mathbf{x} \in \mathcal{S}} A(\mathbf{x}) P(\mathbf{x}) \tag{35}$$

This sum is impractical to calculate directly and, therefore, is approximated by the time average calculated over a time trajectory, $t \mapsto \mathbf{x}(t)$, generated by updating the neuron states using the method of Metropolis, as we saw before:

$$\langle A \rangle \approx \frac{1}{\mathbb{T}} \sum_{t=1}^{\mathbb{T}} A(\mathbf{x}(t)) \tag{36}$$

In a stochastic Hopfield network, the variables (e.g., the state of a neuron) fluctuate over time due to stochasticity and the interactions between neurons.

The time average is an average value calculated over a sufficiently long period of time so that the fluctuations cancel out and the system reaches a state of statistical equilibrium.

So, if, as before, $x_i(t)$ is the state of neuron i at time t. The time average of x_i is defined as:

$$\langle x_i \rangle = \lim_{\mathbb{T} \to \infty} \frac{1}{\mathbb{T}} \sum_{t=1}^{\mathbb{T}} x_i(t) \tag{37}$$

where $\mathbb{T}$ is the simulation time, over the course of which the average is calculated.

In practice, since we cannot simulate the system for an infinite time, we approximate this average by a value calculated over a finite period of time:

$$\langle x_i \rangle \approx \frac{1}{\mathbb{T}} \sum_{t=1}^{\mathbb{T}} x_i(t)$$

The value of $\mathbb{T}$ must be large enough for the system to reach a steady state and for the fluctuations to be sufficiently attenuated.

Simulation. We simulate the Hopfield stochastic network for a discretized time period $\mathbb{T} \in N\Delta t$. At each time step $t \in N\Delta t$, we record the states of the neurons $x_i(t)$ in the modules and in the workspace. After the simulation, we calculated the time averages for each neuron and for the relevant order parameters.

For example, the network overlap $\mathsf{q_m}$ is calculated as:

$$\mathsf{q_m} = \frac{1}{N} \sum_{i=1}^{N} \langle x_i \rangle^2 \tag{38}$$

where $\langle x_i \rangle \approx \frac{1}{\mathbb{T}} \sum_{t=1}^{\mathbb{T}} x_i(t)$ is the time average of the neuron i's state.

This average gives an idea of how much time neuron i spends in each state ($+1$ or -1). If $\langle x_i \rangle$ is close to $+1$, the neuron is almost always active. If $\langle x_i \rangle$ is close to -1, the neuron is almost always inactive. If $\langle x_i \rangle$ is close to 0, the neuron spends approximately the same amount of time in both states. Since we squared it, $\langle x_i \rangle^2$, then

- $q_m \approx 1$ indicates that most neurons in the workspace spend most of their time in a consistent state (all active or all inactive). This means that the workspace exhibits ordered and coherent activity.

- $q_m \approx 0$ indicates that the neurons in the workspace are frequently switching between active and inactive states, or that they are equally distributed between active and inactive over time. This suggests a lack of order and coherence in the workspace, corresponding to a state of confusion or disorder.

For order parameters such as q_m, first calculate the temporal average of each neuron individually, and then calculate the average over all neurons.

Practical Considerations. Before calculating the averages, we must wait for the system to reach a state of equilibrium. We ignore the first steps of the simulation to ensure that the system is not influenced by the initial conditions. We must choose a sufficiently large value of $\mathbb{T}$ so that the averages converge to a stable value.

This convergence can be monitored by visualizing the temporal evolution of the averages and checking if they stabilize. Finally, we should run several simulations, with different initial conditions, to verify that the results are robust and do not depend on the initial choice of neuron states.

In this way, it will be possible to calculate precise and relevant time averages for the analysis of system behavior, and the identification of phase transitions.

References

[ET2000] Gerald M Edelman, Giulio Tononi, "Consciousness: How Matter Becomes Imagination", 2000, Allen Lane, ISBN-10: 0713993081

[JNT2025] J N Tavares, Mind and Consciousness Global Neural Workspace Mathematical and Computational Modeling, Preprint CMUP 2025. IOSR Journal 2026 DOI.: 10.9790/7439-0301012240

[JNT2025a] J N Tavares, "*Mind and Consciousness. Analysis of a "Global Neuronal Workspace" (GNW) Model Based on Stochastic Hopfield Networks.*" Preprint CMUP 2025

[T2004] Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5(1), 42.

[T2014] Oizumi, Masafumi; Albantakis, Larissa; Tononi, Giulio. "*From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0*". PLOS Comput Biol. 10 (5) e1003588. Bibcode:2014PLSCB..10E3588O. doi:10.1371/journal.pcbi.1003588. PMC 4014402. PMID 24811198.

[T2015] Tononi, Giulio, "*Integrated information theory*". Scholarpedia. 10 (1): 4164. Bibcode:2015SchpJ..10.4164T. doi:10.4249/scholarpedia.4164.

[MR1990] B. Muller, J. Reinhardt. *"Neural Networks: An Introduction"*, Springer-Verlag Berlin and Heidelberg, 1990, ISBN-10: 3540523804.

[K2019] Christof Koch, *"The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed"*. The MIT Press, 2019 ISBN-10: 0262042819

[TK2015] Tononi e Koch Tononi G, Koch C. 2015, Consciousness: here, there and everywhere? Phil. Trans. R. Soc. B 370: 20140167.

[DC2006] Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J., Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends in Cognitive Sciences, 10(5), 204-211.

[DC1998] Dehaene, S., Kerszberg, M., Changeux, J.P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. Proceedings of the National Academy of Sciences, 95(24), 14529-14534.

[C2005] Coolen, A. C. C., Kühn, R., & Sollich, P., *"Theory Of neuronal Information Processing Systems"*. Oxford University Press 2005.

[Hop1982] Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 79(8), 2554-2558.

[Hertz1991] , J., Krogh, A., Palmer, R.G. (1991). Introduction to the Theory of Neural Computation. Addison-Wesley Publishing Company.

[H2022] Haiping Huang *"Statistical Mechanics of neuronal Networks"* . Springer 2022.

[Oja1982] Oja, E. (1982). Simplified neuron model as a principal component analyzer. Journal of Mathematical Biology, 15(3), 267-273.

[Hebb1949] Hebb, D.O. (1949). The Organization of Behavior. A Neuropsychological Theory. New York: Wiley Sons.

[Baars1988] Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge University Press.

[DCN2006] Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends in Cognitive Sciences, 10(5), 204-211.

[DC2011] Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. Neuron, 70(2), 200-227.

[SW1963] Claude E Shannon, Warren Weaver, "The Mathematical Theory of Communication". MNG University Presses, 1963. ISBN-10: 0252725484

[DCN2006] Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J., Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends in Cognitive Sciences, 10(5), 204-211.

[DKC1998] Dehaene, S., Kerszberg, M., Changeux, J.P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. Proceedings of the National Academy of Sciences, 95(24), 14529-14534.

[Cool2005] Coolen, A. C. C., Kühn, R., & Sollich, P., *"Theory Of neuronal Information Processing Systems"*. Oxford University Press 2005.

[Hop1982] Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 79(8), 2554-2558.

[Hertz1991] , J., Krogh, A., Palmer, R.G. (1991). Introduction to the Theory of Neural Computation. Addison-Wesley Publishing Company.

[Huang2022] Haiping Huang *"Statistical Mechanics of neuronal Networks"* . Springer 2022.

[B2021] Eric Bertin, *"Statistical Physics of Complex Systems: A Concise Introduction"* (Springer Series in Synergetics). Springer 2021.

[J2024] Henrik Jeldtoft Jensen, *"Complexity Science: The Study of Emergence"*. Cambridge University Press 2024.

[Oja 1982] Oja, E. (1982). Simplified neuron model as a principal component analyzer. Journal of Mathematical Biology, 15(3), 267-273.

[Hebb1949] Hebb, D.O. (1949). The Organization of Behavior. A Neuropsychological Theory. New York: Wiley Sons.

[Baars1997] Bernard J. Baars, "In the Theater of Consciousness: The Workspace of the Mind", OUP USA, ISBN-10: 0195102657.

[Ram2021] Hubert Ramsauer and others: "Hopfield Networks is All You Need", arXiv:2008.02217.

[Sut2018] Sutton, R.S., Barto, A.G. (2018). Reinforcement Learning: An Introduction. MIT Press.