# Mind and Consciousness Analysis of a "Global Neuronal Workspace" (GNW) Model Based on Stochastic Hopfield Networks

**J N Tavares** *

December 16, 2025

**Abstract**

This study consists of two parts. The first is the subject of this article and explores the dynamics of neuronal ignition in the Global Workspace (GNW) model, using a statistical mechanics approach with stochastic Hopfield networks. Taking as a starting point the previous work [JNT2025], we now propose a model where the local modules and the workspace are stochastic Hopfield networks, interacting through connectivity and feedback. This work also provides an initial basis for understanding GNW through the formalism of statistical mechanics, establishing parallels between GNW and complex systems in statistical mechanics, providing an innovative perspective on emergent properties of the model. In the second part, in preparation, we investigate how stochasticity, network capacity, and information integration influence the ignition phase transition. We use Integrated Information Theory (IIT) to quantify the overall coherence of the system, analyzing the role of Entropy Transfer between modules and workspace, and in the calculation of Phi. The results provide insights into the mechanisms underlying consciousness and its emergence[1].

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

[1]About the title: "Mind and Consciousness". The mind is the set of mental processes and thoughts. It is often associated with cognitive processes such as reasoning, memory, imagination, and decision-making. Consciousness is the ability to have subjective experiences, to feel emotions, perceptions, and to have self-awareness. It involves the ability to be aware of the external world and our own internal world (mental processes and thoughts).

# Contents



# 1 Introduction

The Global Neuronal Workspace (GNW) model has been one of the most influential theories in cognitive neuroscience for explaining the mechanisms underlying human consciousness. Initially proposed by Dehaene, Changeux, and Naccache [DCN2006], [DKC1998], the GNW model suggests that consciousness emerges from the integration and dissemination of information in a global network of neurons, which acts as a "workspace" to process and share information between different brain regions. This model offers a robust framework for understanding how the brain selects, amplifies, and maintains relevant information, enabling conscious experience.

In this first work, we explore an innovative approach to integrate Hopfield networks into the GNW model, a type of recurrent neural network known for its ability to store and retrieve patterns stably. Hopfield networks, with their attractor-based dynamics, offer an interesting perspective for modeling the stability and resilience of the global workspace proposed by GNW. By combining these two concepts, we seek to investigate how attractor dynamics can contribute to the integration and maintenance of information in the global workspace, providing a deeper understanding of the neural mechanisms of consciousness.

The proposal of this work is, therefore, to present a hybrid model that integrates GNW with Hopfield networks, exploring how attractor dynamics can be applied to simulate conscious cognitive processes. Through simulations and

theoretical analysis, we hope to contribute to the advancement of the understanding of consciousness, offering a new perspective on how the brain can implement the global workspace and how the dynamic stability of neural networks can play a crucial role in this process.

It is worth mentioning that the GNW model had a precursor – the Global Workspace Theory (GWT) [Baars1997], proposed by Bernard Baars in the 1980s, which is a cognitive model of consciousness that seeks to explain how the brain integrates, selects, and disseminates information between different modules or specialized systems, such as vision, hearing, memory, language, etc. It is worthwhile to briefly describe what it consists of.

The GWT presents consciousness as a "mental theater", where various sources of unconscious processing compete to "access the center stage", making certain information globally accessible to the cognitive system. The GWT argues that consciousness arises from the distribution and coordination of information across a global space accessible to multiple brain systems.

The key concept is this: when information is amplified and disseminated throughout the brain network (through a neuronal "ignition"), it triggers the synchronization of various processes (attention, memory, planning) in a common workspace. Thus, consciousness is the result of a global selection and diffusion of activity that makes information accessible to multiple cognitive functions. A kind of virtual "central framework" in which information, after being amplified by attention, becomes widely available to all brain modules – memory, language, planning, etc.

Most brain processing is unconscious. Consciousness emerges when certain contents become predominant and "earn the right" to be widely disseminated. The critical aspect of consciousness is its global availability, not so much the content itself, but the fact that it can be accessed, used, and manipulated by multiple specialized systems.

## Metaphor of Bernard Baars' Theater of Consciousness

Baars uses the analogy of a theater to explain how consciousness works. Just like in a theater, the mind has:

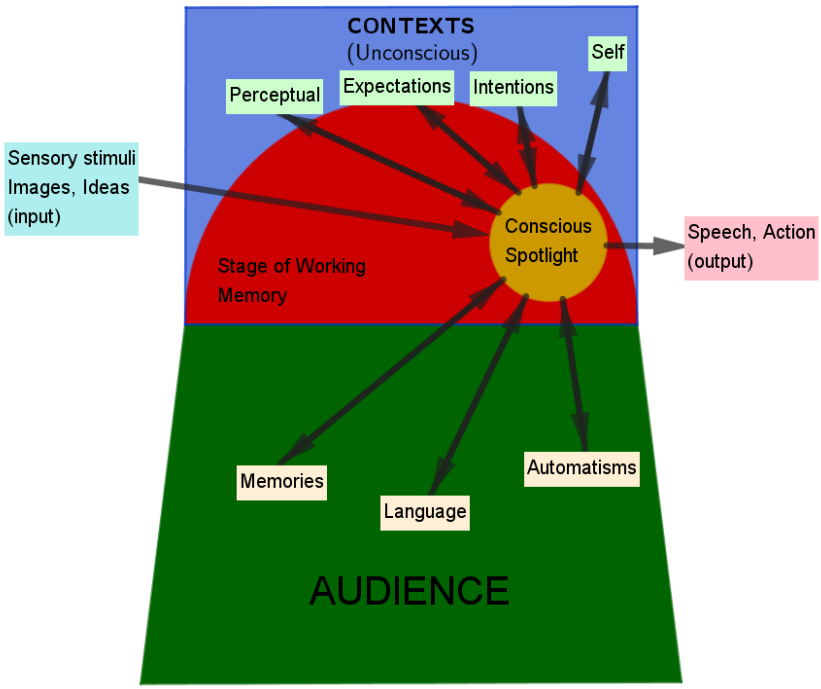



**Figure 1:** Metaphor of Bernard Baars' Theater of Consciousness.

The Stage represents the workspace where the thoughts, images, and sensations of which we are conscious at a given moment unfold. It is a limited space, just like a theater stage, which can only accommodate a restricted number of actors and props; our consciousness can only retain a small number of items simultaneously. Events on stage are dynamic and changeable: actors enter and exit the scene, the lights change, the scenery transforms, reflecting the fluid and transient nature of consciousness.

The Actors (Mental Contents) are the thoughts, sensations, images, memories, and other forms of information that compete for access to the stage (consciousness). Only a few "actors" can be on stage at the same time, reflecting the limited capacity of conscious attention. The quality and intensity of their performance (level of neuronal excitation, salience, relevance) determine the probability of being selected to occupy center stage.

The Spotlight of Attention (Selective Focus) represents the mechanism of selective attention, which illuminates certain actors (mental contents) on the stage, making them the focus of our consciousness. Attention is selectively directed to certain aspects of the experience (the voice of a speaker, a captivating image, a persistent thought), while others remain in the shadows.

Metaphorically, attention functions as a filter through which only some contents pass.

The Audience (Unconscious Processors) represents the wide range of unconscious mental processes that operate in parallel, processing information,

generating emotions, controlling the body, etc. The ”audience” is not directly aware of what is happening on stage, but receives the information that is broadcast from there.

The Backstage Team (Unconscious Context), are the unconscious processes that shape the conscious experience (like setting up a stage). They are the expectations, beliefs, memories, and other background information that influence how we perceive and interpret what is happening on stage (consciousness). They act as ”commands” that select the direction of the performance stage.

Finally, the Director (Executive Functions), which represents the executive control systems located in the frontal cortex, which oversee and coordinate activities on stage. The director makes decisions about which actors should enter the scene, which props should be used, and how the story should be told, reflecting the role of executive functions in regulating and organizing conscious thought. The functions of the Metaphor of the Theatre of Baars are therefore:

- To explain the limits of conscious capacity: Just as a stage can only accommodate a limited number of actors, our consciousness can only handle a restricted amount of information at any given moment.
- To illustrate the importance of the unconscious: Most mental activity occurs outside of our consciousness, such as the actors backstage and the technical crew who make the show possible.
- Giving meaning to the free flow of consciousness: The actors enter and exit the stage, the lights turn on and off, the scenery changes, reflecting the dynamic and fluid nature of our thoughts and sensations.
- Facilitating the understanding of the functions of consciousness: The metaphor of theater helps to understand how consciousness allows us to integrate information, plan actions, solve problems, and interact with the world in a flexible way.

Baars' metaphor was, however, quite criticized. Some criticized it for implying the existence of an internal observer (a ”homunculus”) who watches the spectacle of consciousness. To this, Baars responded that the ”audience” in the theater is not a conscious homunculus, but rather the vast array of unconscious processors in the brain, each acting according to its own competence.

Baars also argues that consciousness is not a physical place in the brain, but rather a process of information dissemination.

In short: Baars' Theater of Consciousness metaphor offers an intuitive and powerful way to understand how consciousness emerges from the interaction between conscious and unconscious mental processes. By highlighting the importance of attention, integration, and dissemination of information, this metaphor provides a rich framework for investigating the neural mechanisms underlying conscious experience.

## The Global Neuronal Workspace (GNW)

Stanislas Dehaene and Jean-Pierre Changeux [DCN2006], [DKC1998], in developing the Global Neuronal Workspace (GNW) model, started from Bernard Baars's Theater of Consciousness metaphor and "neuronalized" it, adding biological specificity, experimental testability, and computational rigor. The main alterations and improvements were as follows:

The replacement of the "mental theater" with a neuronal architecture.

Although Baars sees his metaphor as a cognitive architecture, with actors and spotlights representing functional processes, he does not define where this theater is located in the brain or how it is physically implemented.

The Dehaene & Changeux GNW model identifies a true "neuronal work" – a distributed network, composed mainly of areas of the prefrontal, parietal, and cingulate cortex, a part of the brain involved in processing emotions, memory, and aggression control, connected by long-range axons.

These neurons of the "Workspace" receive signals from various sensory areas and can globally amplify and process the selected content.

For Baars, the spotlight is an abstract concept of selection and amplification. Dehaene & Changeux propose specific neuronal mechanisms for this amplification, replacing the "attention spotlight" with mechanisms of neuronal ignition – when a stimulus reaches a certain threshold (due to strength, novelty, attention), it abruptly activates a pattern of sustained activity on a large scale (observed in EEG, MEG and fMRI [2]. Only content that triggers this sudden ignition becomes conscious; other content remains subliminal or unconscious. The "ignition" is a sudden wave of synchronized activity that spreads across the workspace, sustained by neurons with long-range axons and specialized synapses. Feedback (re-entry) between cortical areas and the thalamus reinforces these signals, keeping the information active in working memory.

For Baars, the "actors" were pieces of information and the audience the receivers of that information. For Dehaene & Changeux, this distinction is replaced by a more complex integrated system of distributed interactions between neurons that form complex circuits.

The metaphor has a director, but it doesn't explain well how this figure manages everything. Dehaene & Changeux use the recurrent loop, mediated by the thalamus, to explain how the networks of consciousness interact.

[2] Electroencephalography is an examination that records the electrical activity of the brain through electrodes placed on the scalp. It is a non-invasive and painless examination, mainly used to diagnose and monitor neurological conditions such as epilepsy, sleep disorders and brain injuries. MEG is a functional neuroimaging technique to map brain activity by recording magnetic fields produced by currents. electrical signals that occur naturally in the brain, using very sensitive magnetometers. fMRI – functional Magnetic Resonance Imaging – Functional magnetic resonance imaging is a non-invasive brain imaging technique that measures brain activity by detecting changes in blood flow. It works by identifying the areas of the brain that are most active during a specific task or at rest, based on the principle that active brain regions require more oxygen and therefore have increased blood flow.

The events in Baars's theater are highly dependent on what a person or observer sees. For Dehaene & Changeux, the model allows the identification of events that could relate to experience in neurobiological terms, making the existence of a subjective report unnecessary, and testing the theory.

Let's elaborate in more detail on the main concepts that characterize the functioning of Dehaene & Changeux's GNW.

- Consciousness as global amplification (broadcast). Consciousness emerges when information processed locally (in sensory, perceptual zones, etc.) is "amplified" and distributed throughout the brain network, becoming available to multiple systems (attention, memory, planning).
- Ignition. There is a critical moment ("ignition") when certain patterns of neuronal activity expand rapidly and globally, for example, via gamma synchrony [3]
- Objective Experiences. Using techniques such as fMRI, EEG, magnetoencephalography, and paradigms like the "visual mask", Dehaene identifies brain markers of consciousness, such as sustained activation of the parietal-frontal cortex. He advocates an empiricist stanceconscious states can (and should) be investigated by objective methods, without falling into metaphysical speculation.
- Meaning and Limits of Consciousness. Dehaene strongly distinguishes between unconscious processing (extensive and efficient) and conscious processing (limited, sequential, but flexible).

For Dehaene, consciousness is a neurocomputational phenomenon, emerging from the selection and diffusion of information in complex neural networks, capable of being tested experimentally in the laboratory.

The GNW model is compatible with processes of attention, working memory, and conscious recognition. The sustained activation of regions of the frontoparietal cortex is a strong neuronal marker of these states of consciousness. Each module functions in parallel, unconsciously. The "Global Workspace" is a means of integration and circulation of information between modules.

[3] Gamma synchrony refers to brain activity in the frequency range, typically between 25 and 100 Hz, although more frequently observed around 40 Hz. This brain activity is associated with high-activity mental states, such as learning, short-term memory, complex problem-solving, and expanded consciousness.
When gamma synchrony increases, with many neurons firing simultaneously, the amplitude (intensity) of the signal also tends to increase. In simpler terms, gamma synchrony is a pattern of brain activity that occurs when many areas of the brain are working together in a coordinated way, at high speed. This is linked to complex mental processes and heightened states of consciousness. In summary: (i). Frequency: 25-100 Hz, focusing around 40 Hz. (ii). Associations: Learning, memory, problem-solving, expanded consciousness. (iii). Synchrony: Increased coordinated activity between different brain areas. (iv). Amplitude: Gamma synchrony tends to increase the amplitude (intensity) of the brain signal, correlating with conscious experience.

Becoming "conscious" means being amplified to this common space. Furthermore, GNW explains many phenomena such as blocks of consciousness, attention, multitasking, amnesia, etc. Dehaene and Changeux developed models in (simulated neural) networks that reproduce experimental phenomena: conscious access, blocking, masking, "all-or-none" of neuronal access, etc. The GNW predicts clear experimental signatures: sustained fronto-parietal activation, gamma synchrony, bursts, delay of conscious reports. It inspired many paradigms of neuroscience and experimental psychology tested in the laboratory with patients, anesthesia studies, sleep, etc.

In conclusion, the original theory (GWT, Baars) paved the way for viewing consciousness as a phenomenon of global integration and diffusion of content. On the other hand, GNW made this model scientific, computationally simulable, and testable by identifying the real circuits, dynamics, and neurophysiological predictions of the "global workspace" in the human brain.

## On Neuronal Correlates of Consciousness (NCC's)

Neuronal Correlates of Consciousness (NCC's) are the dynamic neuronal event processes and minimal mechanisms necessary and sufficient for a specific conscious experience.

In the GNW model, 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. IIT seeks the complex of integrated information that generates $\Phi$. An isolated module can generate its own $\Phi$, but that would be the $\Phi$ of a part of the experiment, not of the unified experience as a whole [JNT2025a].

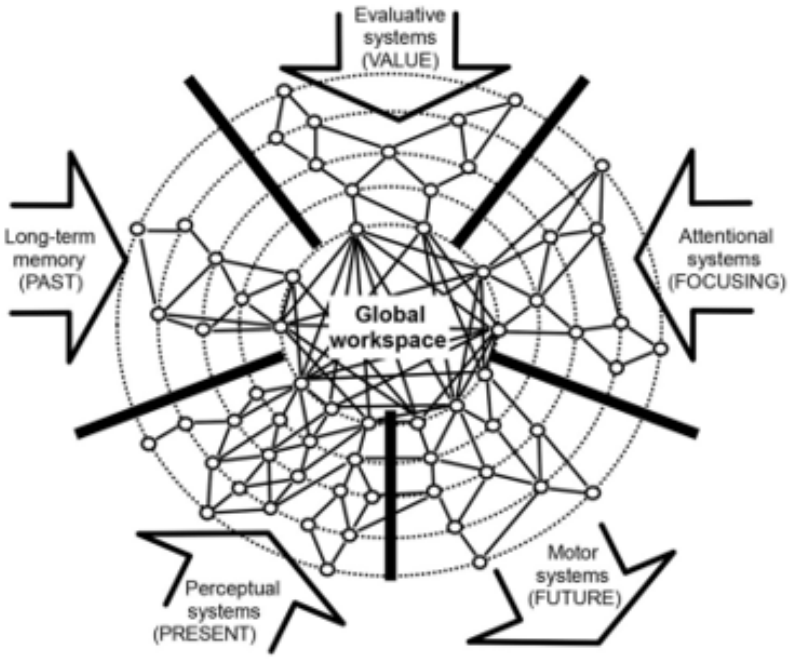
An activated module, when processing a specific pattern (for example, the visual module activated by the "red" pattern), could be the neuronal correlate of the "red" content. This content may or may not become conscious depending on how it is integrated by the workspace. Many processes that occur in 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 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 GWT 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-$\Phi$ state.

## **From** GWT **to** GNW**. Formal aspects**




**Caption. Fig. ??.** The Neural Workspace Global (GNW). Hypothesis (A-C) Original schemes from Dehaene et al. [DCN2006], illustrating the main principles of the GNW hypothesis: local and specialized cortical processors are connected, at the central level, by a central set of highly interconnected areas: (A).

containing a high density of large pyramidal neurons with long-distance axons, (B). At any given moment, this architecture can select information within one or more processors, amplify it, and transmit it to all other processors, thus making it consciously accessible and available for verbal reporting. Recent studies of global cortical connectivity trackers of feedforward and feedback confirm a bow tie architecture with a central core composed primarily of parietal and prefrontal areas, forming a structural bottleneck capable of routing information between other cortical processors (C) (Markov et al., 2013).

□

GNW's central hypothesis is that *consciousness arises when local information (processed by specific sensory areas) is "amplified" and rapidly diffuses into an extensive fronto-parietal network of long-range neurons* – the "global neuronal workspace". Consciousness is different from unconscious processing – much information is processed unconsciously at local levels. Only when you access the global workspace do you become conscious and available for thought, planning, or reporting.

The Global Neuronal Workspace (GNW) is a computational model of how consciousness can arise in the brain. Instead of thinking of the brain as a collection of separate zones, this model imagines it as a *network of interconnected neurons*, where many areas process information locally and unconsciously. However, only some information manages to be "diffused" to the entire network, becoming conscious. The GNW can be simulated on a computer, using neural networks of various architectures.

The model attempts to correspond to real brain components that make conscious experience possible ("neuronal correlates of consciousness" (NCC's)).

The higher functions of the brain (such as consciousness, decision-making, attention control, etc.) are not performed by a single area, but by the combined work of many parts, as in a swarm of bees or a school of fish ("*swarm behavior*") where collective emergent phenomena are observed. This expression ("swarm behavior") refers to the fact that in GNW, neurons work locally, but some activation patterns manage to "infect" and involve the entire network (spreading like a "spark"), leading to *emergent phenomena*: conscious decisions or central control.

In conclusion: GNW presents a robust and testable framework to explain the emergence of consciousness from global brain dynamics. Evidence shows that *consciousness depends on a sudden integration (ignition) into a specific network, allowing diffusion and manipulation of contents throughout the system* – and that the loss of this network corresponds to the loss of consciousness itself.

GNW is a model that uses neural networks (real or simulated on a computer) to explain how consciousness works. It shows that the brain, like a group of ants or a swarm, only becomes conscious of information that can be shared and amplified by a vast network, thus creating a mathematical basis for phenomena such as consciousness, decision-making, and central mental control.

Previous computational studies of the GNW model have focused mainly on simulating its dynamic behavior and examining the conditions under which global ignition occurs. In [JNT2025], we explore the GNW model using classic Hopfield networks, demonstrating how the deterministic dynamics of these networks can simulate competition and information selection in the workspace.

However, classic Hopfield networks have significant limitations. Its deterministic dynamics can lead to convergence towards spurious local minima, hindering the robust representation and processing of information. To overcome these limitations, the present study introduces stochasticity into the dynamics of Hopfield networks, modeling the local modules and the workspace as stochastic Hopfield networks.

Stochasticity allows the network to escape local minima, improves the exploration of the state space, and models, more realistically, the inherently noisy nature of neuronal processing. Stochasticity therefore facilitates the discovery of ideal solution configurations [Hertz1991] [Cool2005].

This work also provides an initial basis for understanding GNW through the formalism of statistical mechanics, establishing parallels between GNW and complex systems in statistical mechanics, providing an innovative perspective on emergent properties of the model. This analogy suggests that the dynamics of GNW can be understood in terms of phase transitions and the emergence of order from disorder.

A fundamental aspect that remains relatively unexplored within the GNW framework is the quantification of integrated information, particularly in the context of conscious experience. Integrated Information Theory (IIT) offers a promising approach to this challenge, proposing that consciousness is fundamentally related to the amount of integrated information a system possesses. However, applying IIT directly to complex systems like GNW is computationally prohibitive due to the complexity of calculating $\sigma F?\mu a, the central measure of integrated in formation.$

The main objectives of this research, which, as mentioned, consists of two articles, are:

1. Formalize the GNW model using coupled stochastic Hopfield networks, allowing an analysis of the dynamics and phase transitions with tools from statistical mechanics.

2. Investigate the role of stochasticity, network capacity, and coupling between modules and workspace in the dynamics of neuronal ignition.

3. Develop and implement mathematical and computational methods to approximate $\Phi$ in the GNW model, incorporating stochasticity in the Hopfield networks of the GNW model developed in [JNT2025].

4. Investigate the relationship between system parameters (e.g., coupling strength, temperature, stochastic noise level) and the emergence of integrated information.

5. Analyze how the temporal dynamics of integrated information relates to the phenomenon of global neuronal ignition in GNW, providing information on the mechanisms underlying consciousness.

The main innovations of this work (in two parts, remember!) include:

1. An innovative architecture, based on stochastic Hopfield networks, for modeling the GNW model.
2. Modification of the stochastic dynamics of the network by introducing modified energies through the introduction of external fields, which allow a *dynamic "guided" by stored patterns*.
3. Analysis of *Phase Transitions and the Emergence of Consciousness*.

In Part 2 [JNT2025a], we will discuss:

1. An innovative adaptation of the IIT structure to the GNW model, using concepts from statistical mechanics, taking into account the stochastic nature of neural networks.
2. A new approach to Integrated Information Theory (IIT) through notions of Information Theory.
3. The implementation of heuristic algorithms to approximate the Minimum Information Partition (MIP), a crucial step in estimating $\Phi$, optimized for the GNW architecture.
4. Adaptation and application of IIT to quantify the integration of information in GNW, relating $\Phi$ to the order parameters of the system.
5. The evaluation of the effects of stochasticity, both at the level of individual neurons and of the system as a whole, on the capacity for information integration and on the dynamics of neuronal ignition.
6. Information on conditions that facilitate high levels of integrated information and its connection to the dynamics of global neuronal ignition, revealing potential mechanisms underlying consciousness.

Here is a short summary of the sections that make up this work. The 2 section revisits stochastic Hopfield networks, an associative memory model with non-deterministic dynamics. It discusses the limitations of traditional Hopfield networks and the advantages of stochastic networks, which allow escaping local minima and modeling the noise inherent in neuronal processing. Section 3 formalizes the dynamics, integrated information, and phase transitions in the GNW model, using stochastic Hopfield networks to model the local modules and the work. It explains the interaction between the modules through

coupling energy. **??** section introduces the external fields that modify the energies and allow defining a stochastic dynamic "guided" by the stored patterns. In the section 5, consciousness is postulated as the emergence of a globally ordered state, actively orchestrated by the dynamics of the GNW, resulting from a phase transition. Finally, in Appendix **??**, the most important mathematical preliminaries for understanding this study are presented.

# 2 Stochastic Hopfield Networks: Statistical Mechanics and Phase Transitions

## Introduction

Hopfield networks[4] while useful for modeling associative memories, have some limitations:

- Deterministic Convergence: the network dynamics are completely deterministic, which means that, given an initial state, the network will always converge to the same attractor. This limits the network's ability to explore different solutions or adapt to noisy or ambiguous inputs.
- Noise Sensitivity: Traditional Hopfield networks can be sensitive to noise and initial states outside the basin of attraction of memorized patterns.
- Lack of Biological Realism: The deterministic dynamics of traditional Hopfield networks do not reflect the complexity and stochasticity observed in real neural systems.

On the other hand, stochastic Hopfield networks offer several advantages over classical networks:

- Exploring the State Space: Stochasticity allows the network to escape local minima of the energy function and explore different solutions, improving the ability to find global patterns and handle complex problems.
- Robustness to Noise: Stochasticity makes the network more robust to noise and variations in inputs, simulating the brain's ability to process information in noisy and uncertain environments.
- Biological Realism: Stochastic dynamics better reflect real neuronal activity, where noise and variability are intrinsic characteristics.

[4]John Hopfield (1933 - ) is a physicist, biologist and neurologist born in Chicago, in the United States of America. He was awarded the Nobel Prize in Physics in 2024, jointly with Geoffrey Hinton, for "fundamental discoveries and inventions that enable machine learning with artificial neural networks".

This section presents a short review of stochastic Hopfield networks, an associative memory model with non-deterministic dynamics. We discuss their statistical mechanics, their relationship to classical Hopfield networks, and the phase transitions they exhibit. Mathematical formalism is used to describe their dynamics and behavior.

These concepts are essential for building the new GNW model, starting from the Tavares model.

Hopfield networks are associative memory models proposed by John Hopfield in 1982, which are capable of storing and retrieving information patterns through a decentralized dynamic. Classical Hopfield networks are deterministic, evolving to one of their point attractors, which represent the stored memory patterns.

Stochastic Hopfield networks are an extension of classical Hopfield networks, where the updating dynamics of neurons (generally asynchronous) are governed by probabilities, introducing an element of randomness. This stochasticity can be useful for escaping local minima and exploring the state space more efficiently.

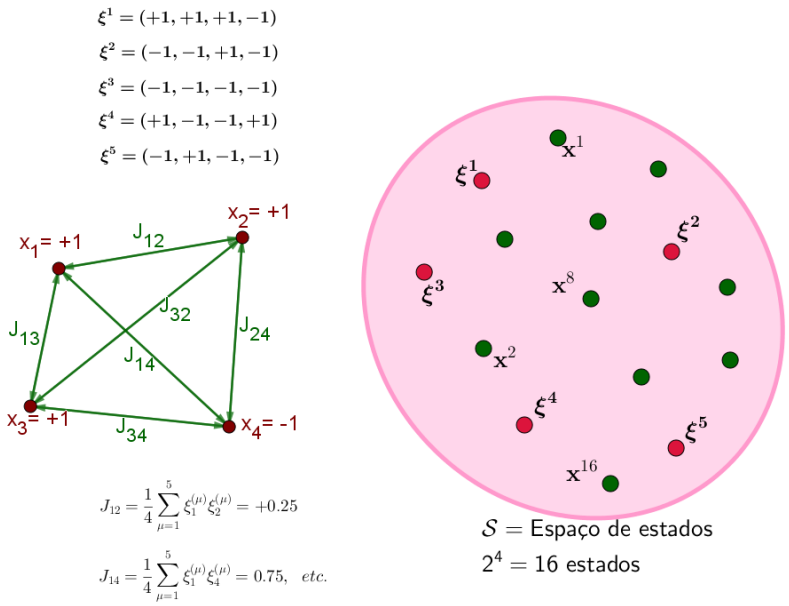



**Figure 2:** Hopfield network with $N = 4$ neurons and respective state space with $2^4 = 16$ states. The network has 5 memorized patterns and the weights $J_{ij}$ are calculated by Hebb's rule.

For example, the figure 3, represents a pattern, and another corrupted one, of a "pixelation" of the digit $0$, using $160$ binary neurons. The state space is, in this case, gigantic, with $2^{160}$ states, a colossal number!

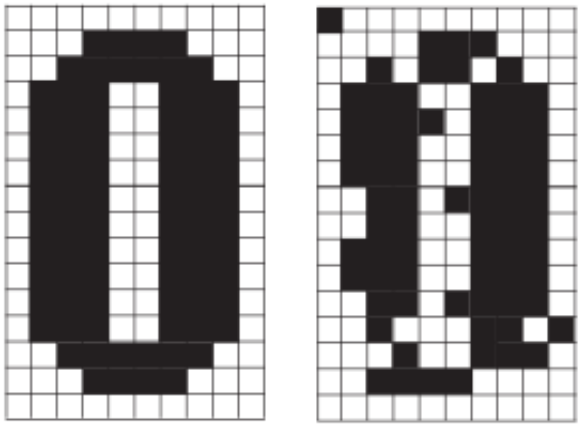


**Figure 3:** Binary image ($N = 160$) of the digit 0 and a distorted version of the same image. There are $N = 160$ neurons $x_i, i = 1, \cdots, N$, and ■ represents $x_i = +1$ while □ represents $x_i = -1$.

## The Stochastic Hopfield Network Model

A stochastic Hopfield network, $\mathcal{H}$, consists of $N$ neurons $i = 1, \cdots, N$, each of which is associated with a binary random variable, $X_i$, whose possible values are $x_i \in \{-1, 1\}$ – the possible states of neuron $i$. If $x_i = +1$ the neuron is said to be active, and if $x_i = -1$, it is said to be inactive. The network state will therefore be an N-dimensional vector $\mathbf{x} \in \{-1, 1\}^N$ and, therefore, the network state space, $\mathcal{S}$, consists of $2^N$ elements $\mathbf{x}^I = (x_1^I, x_2^I, \cdots, x_N^I) \in \{-1, 1\}^N$, $I = 1, 2, \cdots, 2^N$.

The network energy, in the state $\mathbf{x} \in \mathcal{S}$, is given by:

$$E(\mathbf{x}) = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} x_i x_j, \tag{1}$$

where $J_{ij}$ represents the strength of the interaction between neurons $i$ and $j$.

Typically, in a Hopfield network, there are specific states called stored patterns or memories. In this case, the connections are defined by Hebb's rule:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$$

where $\xi_i^\mu \in \{-1, 1\}$ is the state of neuron $i$, in the memory pattern $\mu$. The connections $J_{ij}$ are typically symmetric ($J_{ij} = J_{ji}$), with a zero diagonal ($J_{ii} = 0$).

In Fig. 2, the Hopfield network $\mathcal{H}$ has $N = 4$ neurons, and the corresponding state space $\mathcal{S}$ has $2^4 = 16$ states. This network has 5 memorized patterns $\boldsymbol{\xi}^\mu$, $\mu = 1, \cdots, 5$ and the weights $J_{ij}$ are calculated by Hebb's rule.

In the state space $\mathcal{S}$ of a Hopfield network $\mathcal{H}$, we define the Boltzmann probability measure by:

$$P(\mathbf{x}) = \frac{e^{-\beta E(\mathbf{x})}}{Z(\beta)} \tag{3}$$

where $Z(\beta)$ is the so-called partition function, defined by:

$$Z(\beta) = \sum_{\mathbf{x}\in\mathcal{S}} e^{-\beta E(\mathbf{x})} \tag{4}$$

$P(\mathbf{x})$ is therefore the probability of finding the network in a given state $\mathbf{x} \in \mathcal{S}$. $\beta = 1/T$ is the inverse of (pseudo) temperature $T$. Note that, for a fixed temperature, $T$, the higher the energy $E(\mathbf{x})$, the less likely the corresponding configuration x is.

The local field $h_i$, n of neuron $i$, is defined as the influence of other neurons on it:

$$h_i = \sum_{j\neq i} J_{ij} x_j \tag{5}$$

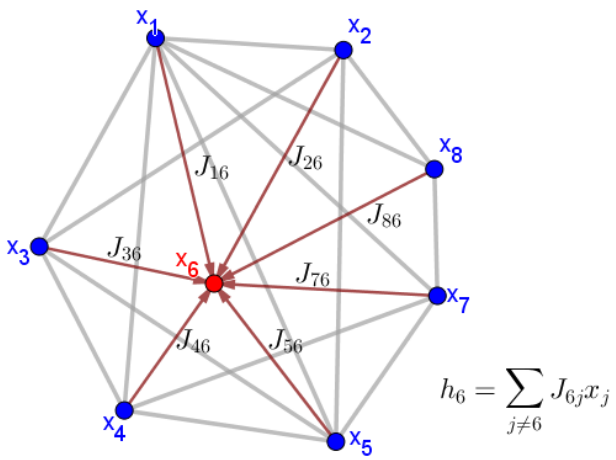



**Figure 4:** Hopfield network with $N = 8$ neurons. $h_6$ = Local field of neuron 6.

## Network Status Update.

We usually opt for what is called ASYNCHRONOUS UPDATE, in which neurons are updated one at a time, in order to sequentially or randomly update. Therefore, at each time step $t$, we select a neuron $i$ for updating, either sequentially, according to a fixed order (e.g., $1, 2, 3, \cdots, N$), or randomly, choosing a neuron randomly in a uniform manner.

This method is biologically more plausible and generally leads to a more stable dynamic. The fundamental difference between Hopfield networks, classical (deterministic) and stochastic, lies in the UPDATE RULE of the neurons.

In the deterministic case, the update of the state $x_i$, of a neuron $i$, is given by:

$$x_i(t + \Delta t) = \mathsf{Signal}(h_i(t)) \tag{6}$$

where, for each fixed neuron $i$, $h_i$ is its LOCAL FIELD, defined by (5).

In the stochastic case, the update is done using the Metropolis criterion. The Metropolis method is a Monte Carlo algorithm that allows simulating the evolution of a system (such as GNW) so that, over time, it reaches a state of thermodynamic equilibrium. In simpler terms, the method makes the system "explore" different configurations, so that the configurations with the lowest energy are more likely to occur, as dictated by the Boltzmann distribution.

To do this, we calculate the energy variation $\Delta E_i$, which would occur if the state of neuron $i$ were reversed: $x_i \to -x_i$:

$$\Delta E_i = E(..., -x_i, ...) - E(..., x_i, ...)$$

Since the energy of the Hopfield network is given by $E(\mathbf{x}) = -1/2 \sum_{i,j} J_{ij} x_i x_j$, substituting, we obtain:

$$\begin{aligned}
\Delta E_i &= \left(-1/2 \sum_{i,j} J_{ij}(-x_i)x_j\right) - \left(-1/2 \sum_{i,j} J_{ij} x_i x_j\right) \\
&= 1/2 \sum_{i,j} J_{ij} x_i x_j + 1/2 \sum_{i,j} J_{ij} x_i x_j \\
&= 2x_i \sum_{j \neq i} J_{ij} x_j \\
&= 2x_i h_i
\end{aligned} \tag{7}$$

We now apply the Metropolis criterion. We have two situations:

1. If $\Delta E_i < 0$, then the inversion of the neuron's state $i$ decreased the system's energy. In this case, the transition is always accepted:

$$P(x_i \to -x_i) = 1, \quad if \;\; \Delta E_i = 2x_i h_i < 0$$

2. If $\Delta E_i \geq 0$, the inversion of the neuron's state increases the system's energy. In this case, the transition is accepted with a probability given by $\exp(-\beta \Delta E_i)$:

$$P(x_i \to -x_i) = \exp(-2\beta x_i h_i), \quad se \;\; \Delta E_i = 2x_i h_i \geq 0$$

where $T = 1/\beta$ is the (pseudo) temperature (or noise) of the system.

This is the key part of the Metropolis criterion – it represents the probability of accepting a transition to a higher energy state. This probability decreases

with increasing $\Delta E_i$ and decreases with increasing $\beta$ (i.e., with decreasing temperature). Since we chose an asynchronous update, it is important to ensure that, on average, all neurons are updated.

This criterion ensures that the network tends towards low-energy states, and that it can also escape local minima, due to temperature-controlled stochasticity.

**Additional Notes.**

1. The network dynamics can be viewed as a Markov chain, where the network state at time $t + \Delta t$ depends only on the state at time $t$. After a sufficiently long time, the probability distribution over the network states approaches a stationary distribution, where the probabilities of the states do not change over time.
2. The Metropolis criterion can be seen as a way to ensure that the Boltzmann distribution is sampled correctly.
3. The trajectory can be viewed as a sequence of "jumps" between the discrete points of the $\mathcal{S}$ state space.

## Relationship between Attractors and Memory Patterns

The relationship between stored Memory Patterns, $\boldsymbol{\xi}^{(\mu)}$, the energy function, and the dynamics of convergence in a Hopfield network (classical or stochastic) is a crucial point.

In an idealized scenario, each memory pattern corresponds to a stable attractor of the network. That is, if the initial state of the network is close to a memory pattern, the network dynamics will converge to that pattern.

But in practice, things are much more complicated. In reality, the relationship between attractors and memory patterns can be very complex due to several factors:

The Hopfield network may have a limited storage capacity. If you try to store many patterns, not all patterns will become stable attractors. The network may have attractors that do not correspond to any of the stored memory patterns. These spurious attractors can be linear combinations or inversions of memory patterns.

Some memory patterns can be unstable, especially if the memory capacity, $\mathcal{C} = p/N$, is high.

If the memory patterns are correlated, this can affect the stability of the attractors and lead to the creation of spurious attractors.

In a classic Hopfield network, the memory patterns $\boldsymbol{\xi}^{(\mu)}$ are, by construction, local minima of the energy function. Hebb's rule (**??**) is designed to ensure that stored patterns correspond to low-energy configurations.

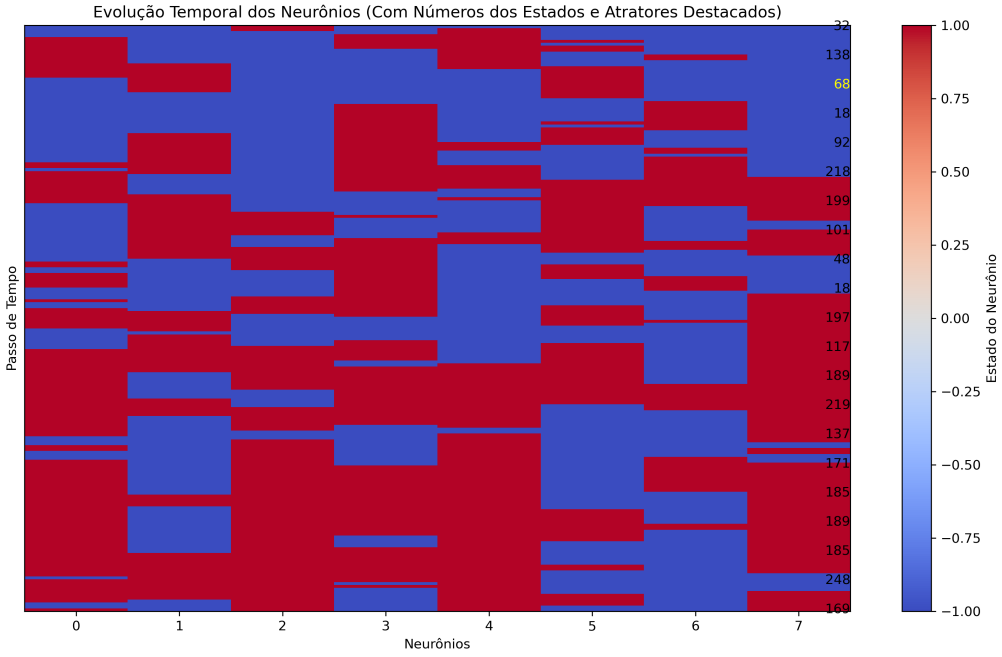



**Figure 5:** Network Status Update with 8 neurons and two stored patterns. The vertical axis represents the time steps. The horizontal axis represents the neurons (from 0 to 7). Each cell in the image represents the state of a neuron at a given time step. Inactive neurons (-1) are indicated in blue, while active neurons (+1) are indicated in red. This graph allows you to quickly visualize the evolution of the states of all neurons over time, providing an overview of network dynamics.

If the initial state is sufficiently close to a stored pattern, the deterministic dynamics will converge to that pattern, making it an attractor. In a classic network, if we start the asynchronous network update in any state, it does not necessarily converge to a memory pattern. It may converge to a spurious local minimum, which does not correspond to any stored pattern.

In a stochastic Hopfield network, memory patterns remain low-energy points, but the presence of noise (controlled by temperature $T$) changes the situation.

Memory patterns are local minima of the energy function, but the network may not remain in them indefinitely due to thermal fluctuations.

In a stochastic network, convergence to a memory pattern is still less guaranteed than in the classical network. Noise allows the network to escape local minima, but also prevents it from stabilizing at a single attractor. The network may fluctuate around an attractor for a while, but eventually it may escape and move to another attractor or to a disordered state. The probability of escaping

```
Histórico dos Estados:
Passo 0: [-1 -1 -1 -1 -1  1 -1  1] (Estado 5)
Passo 1: [-1 -1 -1 -1 -1  1 -1  1] (Estado 5)
Passo 2: [-1 -1 -1 -1 -1  1 -1 -1] (Estado 4)
Passo 3: [-1 -1 -1  1 -1  1 -1 -1] (Estado 20)
Passo 4: [-1 -1 -1  1 -1  1  1 -1] (Estado 22)
Passo 5: [-1 -1 -1  1 -1  1  1 -1] (Estado 22)
Passo 6: [-1 -1 -1  1 -1  1  1 -1] (Estado 22)
Passo 7: [-1 -1 -1  1 -1  1  1 -1] (Estado 22)
Passo 8: [-1 -1  1  1 -1  1  1 -1] (Estado 54)
Passo 9: [-1 -1  1  1  1  1  1 -1] (Estado 62)
Passo 10: [-1 -1  1  1 -1  1  1 -1] (Estado 54)
Passo 11: [-1 -1  1  1 -1  1 -1 -1] (Estado 52)
Passo 12: [-1 -1  1  1  1  1 -1 -1] (Estado 60)
```

**Figure 6:** First 12 steps of updating the Network Status.

patterns (increased with temperature) makes the transition much easier.

The precision with which stored patterns can be represented in a stochastic network depends on the network's capacity $\mathcal{C} = p/N$, and also depends on the temperature $T$.

In summary:

- Memory patterns are local energy minima.
- Convergence to a pattern is not guaranteed due to stochasticity.
- Temperature controls how likely the network is to escape local minima and explore different regions of the state space.
- Attractors are stable states of the network, to which the dynamics converge over time. Ideally, each stored memory pattern corresponds to an attractor. Analyzing the stability of the attractors and the influence of temperature is crucial to understanding the network's behavior.
- At low temperature ($T \approx 0$), dynamics become more deterministic. The network converges rapidly to one of the local minima of the energy function, which ideally corresponds to a memory pattern.
- At high temperature ($T >> 0$), the dynamics become completely random. The network does not converge to any specific attractor and explores the state space randomly. Ergodicity is guaranteed when the temperature is high. The lower the temperature, the more regions there are and the weaker the connection between these different regions.
- At intermediate temperature ($T \approx T_c$), the network manages to balance the exploration of the state space with convergence to the attractors. The dynamics are more complex and the network can fluctuate around an attractor for a while before jumping to another.

The energy function $E$ plays a fundamental role in determining the attractors. The attractors correspond to the local minima of this function. The mean-field equations provide an approximation of the equilibrium state of the network, but do not necessarily describe all attractors. The solutions of the mean-field equations correspond to the fixed points of the network dynamics, which may or may not be stable attractors. To identify the attractors, it is necessary to analyze the stability of the solutions of the mean-field equations (We do not discuss them in this work).

A solution is stable if the network converges to it from nearby initial states. The solutions of the mean-field equations (the stable states) can be characterized by their overlap with the stored memory patterns. A high overlap with a pattern indicates that the stable state corresponds to that pattern.

# 3 Global Workspace Model Dynamics with Stochastic Hopfield Networks

The Global Workspace Neuronal Model (GNW) is an influential theoretical framework in cognitive neuroscience, proposing that consciousness emerges when information from various specialized modules is integrated and amplified in a global workspace.

In this section we will formalize the dynamics of this model, using stochastic Hopfield networks to model the local modules and the workspace.

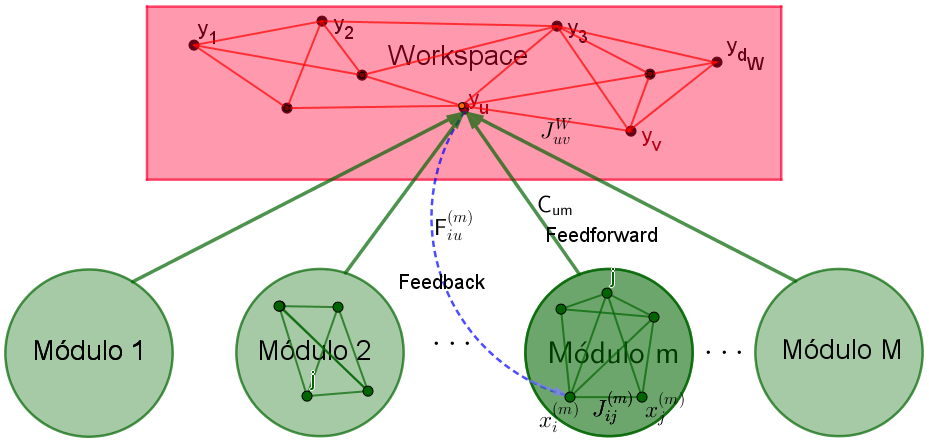



**Figure 7:** GNWModel. Connections of the neurons of each module to the neurons of the Workspace, their respective weights and feedback from the neurons of the Workspace to the neurons of the modules. Inter-module competition is not represented.

## GNW as a Coupled System of Stochastic Hopfield Networks.

As in [JNT2025], the GNW model is represented as a coupled system of Hopfield Networks, but this time, each main component (local modules and workspace) is modeled as a Hopfield Network with stochastic dynamics. The notations are those used in the cited work. The system components are (Fig. 7):

- Local Modules. There are $M$ local modules $\mathsf{m} = 1, 2, \cdots, M$. Each module $\mathsf{m}$ is modeled as a Hopfield network with $N_{\mathsf{m}}$ neurons. The state of neuron $i$ in module $m$ is $x_i^{(\mathsf{m})} \in \{-1, 1\}$. The energy of module $\mathsf{m}$, in the state $\mathbf{x}^{(\mathsf{m})} \in \{-1, 1\}^{N_{\mathsf{m}}}$, is given by:

$$E_{\mathsf{m}}\left(\mathbf{x}^{(\mathsf{m})}\right) = -\frac{1}{2}\sum_{i,j} J_{ij}^{(\mathsf{m})} x_i^{(\mathsf{m})} x_j^{(\mathsf{m})}$$

  where $J_{ij}^{(\mathsf{m})}$ are the interactions between neurons $i$ and $j$ in module $\mathsf{m}$.

- Each module has $p_{\mathsf{m}}$ stored patterns $\boldsymbol{\xi}^{(\mathsf{m},\mu)}, \mu = \cdots, \mathsf{p}_{\mathsf{m}}$. Synaptic connections are given by Hebb's rule:

$$J_{ij}^{(\mathsf{m})} = \frac{1}{N_{\mathsf{m}}}\sum_{\mu} \xi_i^{(\mathsf{m},\mu)} \xi_j^{(\mathsf{m},\mu)}$$

- Workspace. The workspace is modeled as another Hopfield network with $N_W$ neurons, where the state of neuron $\mathsf{u}$ in the workspace is $y_{\mathsf{u}} \in \{-1, 1\}$. The energy of the workspace, in the state $\mathbf{y} \in \{-1, 1\}^{N_W}$, is given by:

$$E_W(\mathbf{y}) = -\frac{1}{2}\sum_{\mathsf{u}\neq\mathsf{v}} J_{\mathsf{uv}} y_{\mathsf{u}} y_{\mathsf{v}}$$

  where $J_{\mathsf{uv}}$ are the interactions in Workspace. The workspace has $p_W$ stored patterns $\zeta^{(\kappa)}, \kappa = \cdots, \mathsf{p}_{\mathsf{W}}$. Synaptic connections are given by Hebb's rule:

$$J_{\mathsf{uv}} = \frac{1}{N_W}\sum_{\alpha} \zeta_{\mathsf{u}}^{(\kappa)} \zeta_{\mathsf{v}}^{(\kappa)}$$

- Connectivity. The connectivity matrix $C = (C_{\mathsf{um}})$ has dimension $N_W \times M$, where $C_{\mathsf{um}}$ represents the influence of the module $\mathsf{m}$, as a whole (via average activity or lateral competition of the module $\mathsf{m}$) on the neuron $\mathsf{u}$ of the workspace.

- The average activity of module $\mathsf{m}$, $A_{\mathsf{m}}$, is defined by:

$$A_{\mathsf{m}} = \frac{1}{N_{\mathsf{m}}}\sum_{i=1}^{N_{\mathsf{m}}} x_i^{(\mathsf{m})}$$

and the competition of module m, as:

$$\mathsf{Comp}_{\mathsf{m}} = A_{\mathsf{m}} - \frac{1}{M-1}\sum_{\mathsf{n}\neq\mathsf{m}} A_{\mathsf{n}} \tag{8}$$

where the sum is over all other modules $\mathsf{n} \neq \mathsf{m}$. $\mathsf{Comp}_m$ represents the difference between the activity of module m and the average activity of the other modules. If $\mathsf{Comp}_{\mathsf{m}}$ is positive, the module m is more active than the average activity of the other modules. If it is negative, you are less active.

$A_{\mathsf{m}}(t)$ measures the total magnetization of module m. If the memorized pattern is $\boldsymbol{\xi} = (1, 1, 1, -1, -1, -1)$, the magnetization is $0$. But if the current state is $(1, 1, 1, -1, -1, -1)$, $A_{\mathsf{m}} = 0$. If the current state is $(1, -1, 1, -1, 1, -1)$, $A_{\mathsf{m}} = 0$. Both have zero magnetization, but the first is a pure pattern and the second is not. Therefore, $A_{\mathsf{m}}(t)$ and $\mathsf{Comp}_{\mathsf{m}}(t)$ alone do not distinguish whether a module is in a clear, memorized pattern or just in a high/low magnetization state.

In the following section, we will see how to correct this deficiency in the model, opting for pattern-driven dynamics. We will see what this means: the role of competition in network dynamics, as well as that of *feedforward connectivity* $C_{\mathsf{um}}$ and *feedback* $F_{i\mathsf{u}}^{(\mathsf{m})}$.

This model will be completed in the **??** section. But for now, let's analyze the role of stored patterns in the GNW dynamics.

# 4 Role of Patterns and Connectivity in GNW with Lateral Competition

This section aims to integrate memorized patterns and connectivity into the GNW model through a new competition function $\mathcal{F}_{\mathsf{m}}(t)$ that modulates the input to the workspace.

To correct the anomaly described in the previous section, we must ensure that *competition between modules is explicitly guided by the overlaps of the modules with their own memorized patterns, and therefore directly reflects their coherence with the learned patterns, which is fundamental for a truly pattern-driven dynamic.*

Therefore, for the GNW dynamic to be effectively "pattern-driven", the selection and prioritization of modules (competition) must be based on their ability to activate and sustain memorized patterns.

To achieve this, we start by defining the modular overlaps. For each $t = 1, \cdots, \mathbb{T}$, and for each Module m, we calculate the modular overlaps:

$$\mathsf{Ov}_{\mathsf{m}}^{(\mu)}(t) = \mathsf{Overlap}\left(\mathbf{x}^{(\mathsf{m})}(t), \boldsymbol{\xi}^{(\mathsf{m},\mu)}\right) = \frac{1}{N_{\mathsf{m}}}\sum_{i=1}^{N_{\mathsf{m}}} x_i^{(\mathsf{m})}(t)\xi_i^{(\mathsf{m},\mu)} \tag{9}$$

for each pattern $\boldsymbol{\xi}^{(\mathsf{m},\mu)}$. This is the "thermometer" that measures how close the current state of the module, $\mathbf{x}^{(\mathsf{m})}(t)$, is to one of its memorized patterns. This ensures that the competition is not based solely on raw activity ($A_{\mathsf{m}}(t)$), but on the consistency of the activity with a learned pattern.

We now calculate the maximum overlap of the current state of the module $\mathsf{m}$ with its patterns:

$$\mathsf{o}_{\mathsf{m}}(t) = \max_{\mu} \mathsf{Overlap}\left(\mathbf{x}^{(\mathsf{m})}(t), \boldsymbol{\xi}^{(\mathsf{m},\mu)}\right) \tag{10}$$

We now define lateral competition based on maximum overlaps: for each Module $\mathsf{m}$:

$$\mathsf{Comp}_{\mathsf{m}}(t) = \mathsf{o}_{\mathsf{m}}(t) - \frac{1}{M-1}\sum_{\mathsf{n}\neq\mathsf{m}} \mathsf{o}_{\mathsf{n}}(t) \tag{11}$$

This means that the competition now rewards the modules that are more "clear" (with high overlap with their patterns), and not just the most active ones. The competition function with overlaps as inputs is:

$$\mathcal{F}_{\mathsf{m}}(t) = \frac{e^{\beta_s \mathsf{Comp}_{\mathsf{m}}(t)}}{\sum_{\mathsf{n}=1}^{M} e^{\beta_s \mathsf{Comp}_{\mathsf{n}}(t)}} \tag{12}$$

Analogously, we define the overlaps of the current state of the workspace with their memorized patterns:

$$\mathsf{Ov}_W^{(\kappa)} = \mathsf{Overlap}\left(\mathbf{y}(t),\right.$$

$\left.\boldsymbol{\zeta}^{(\kappa)} = \frac{1}{N_W}\sum_{\mathsf{u}=1}^{N_W} \mathbf{y}(t)\zeta_{\mathsf{u}}^{(\kappa)}\right)$ (13) and also the maximum overlap:

$$\mathsf{o}_W(t) = \max_{\kappa} \mathsf{Ov}_W^{(\kappa)} \tag{14}$$

The Dominant Pattern will be

$$\boldsymbol{\zeta}^{(\kappa_t^*)} \quad \text{where} \quad \kappa_t^* = \arg\max_{\kappa} \mathsf{Ov}_W^{(\kappa)} \tag{15}$$

$\mathsf{Comp}_{\mathsf{m}}(t)$, defined by (11) (inter-module lateral competition) is the basis for the selective attention mechanism between the modules of the GNW. This measure $\mathsf{Comp}_{\mathsf{m}}(t)$ now represents the "relative strength" of the module $\mathsf{m}$, compared to the other modules. A positive and high $\mathsf{Comp}_{\mathsf{m}}(t)$ indicates that the module $\mathsf{m}$ is more active than the others, winning the competition.

Our goal is to integrate the role of memorized patterns and the *feedforward connectivities*, $(C_{\mathsf{um}})$, which connect the *module m, as a whole*, to the neuron $\mathsf{u}$ of the workspace and the *Feedback* $(F_{i\mathsf{u}}^{(\mathsf{m})})$, which connects the neuron $\mathsf{u}$ of the workspace to the neuron $i$ of the module $\mathsf{m}$.

In the context of the GNW model, where multiple specialized modules provide "raw material" of information, the way the workspace selects and prioritizes

this information is fundamental. For this model to reflect cognitive processes more robustly, the memorized patterns within each subsystem (modules and workspace) must play a more prominent and active role, acting as guides for dynamics.

This leads us to formulate the following hypothesis:

> *The memorized patterns, within each subsystem (modules and workspace), actively influence the flow of information, decision-making (modular competition), and the formation of integrated states (workspace ignition).*

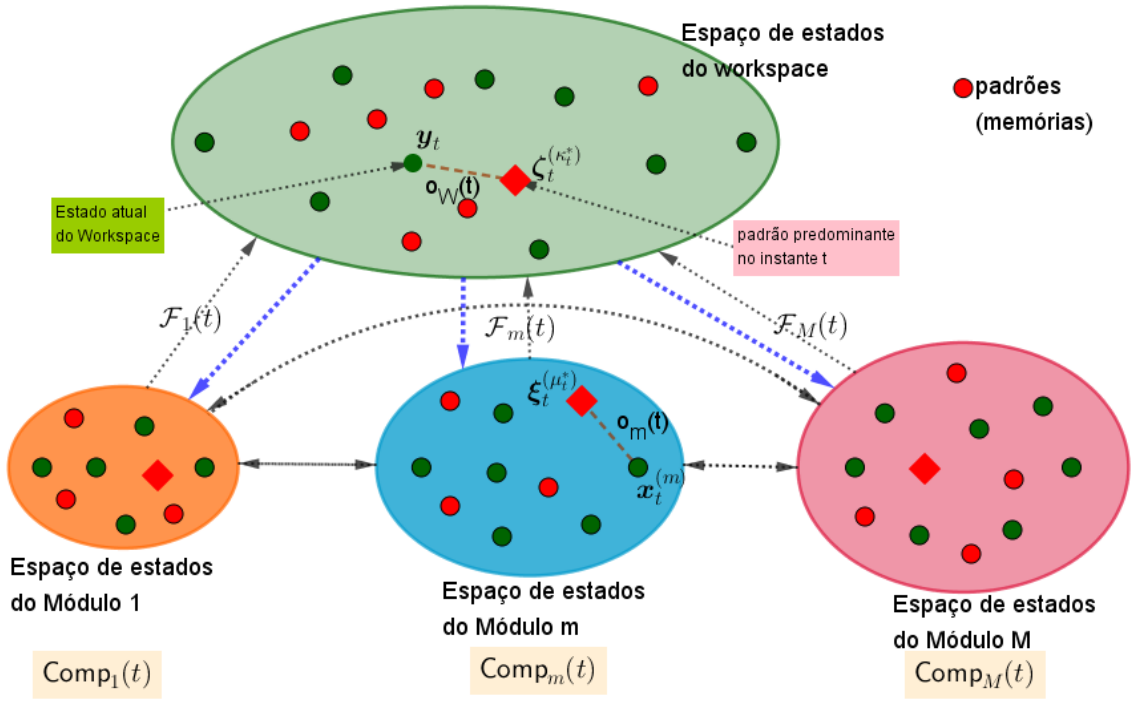



**Figure 8**

The patterns stored in the modules, $\{\boldsymbol{\xi}^{(\mathsf{m},\mu)}\}_{\mu=1}^{\mathsf{p_m}}$, and the connections $C_{\mathsf{mu}}$, should drive the competition so that the workspace receives the most relevant information. The competition tells us which module has the strongest and most consistent representation for the workspace input. This evaluates how "active" and "consistent" the signal coming from each module m is.

To put these ideas into practice, let's now define a Competition Function (the bridge), $\mathcal{F}_{\mathsf{m}}$, which implements a mechanism of selective attention or prioritization of relevance between modules. $\mathcal{F}_{\mathsf{m}}$ quantifies how "important" or

"coherent" the signal of a specific module m is at a given instant in time $t$, in relation to the other modules. Let's now see how to define it.

$\mathsf{Comp}_{\mathsf{m}}(t)$ (see **??**) is the ideal measure of "relevance" to feed the function $\mathcal{F}_{\mathsf{m}}(t)$. The function $\mathcal{F}_{\mathsf{m}}(t)$ serves to normalize and apply a "selective gain" to these $\mathsf{Comp}_{\mathsf{m}}(t)$ values, transforming them into weights that sum to 1 (using softmax), and which determine how strongly the m module influences the workspace.

$$\mathcal{F}_{\mathsf{m}}(t) = \frac{e^{\beta_s \mathsf{Comp}_{\mathsf{m}}(t)}}{\sum_{\mathsf{n}=1}^{M} e^{\beta_s \mathsf{Comp}_{\mathsf{n}}(t)}} \tag{16}$$

where $\beta_s$ is the selectivity parameter. A high $\mathcal{F}_{\mathsf{m}}(t)$ means that the m module "won" the competition and will have a greater impact on the Workspace.

## Pattern-Driven Dynamics

Let's now look at the role of memorized patterns and connectivities in the dynamics of the GNW model.

> *The goal is to add "external" fields to the energy of each subsystem, so that the dynamics are influenced by the memorized patterns.*

The external fields are the vehicles through which dynamic interaction and active modulation occur in GNW. They translate the "intention" (active patterns), the "attention" (modular competition), and the "coordination" (feedback) of the forces that drive the Metropolis dynamics of each subsystem. The correct formulation of these fields is essential for the plausibility and functionality of the model.

Intuition about external fields and pattern-driven dynamics is fundamental. We can think of external fields as "gravitational forces" that tilt the energy landscape of the system. They are not inherent to the network, but come "from outside" (from another subsystem, or from an internal control mechanism) – their role is to "push" the current state of the neurons in a certain direction. Memorized patterns function as "ideas" or "memories" of the network. Pattern-driven dynamics means that these "ideas" are not passively stored, but are actively used to drive what the network does.

Let's now see how to formalize these intuitions. To do this, let's start with the dynamics in the workspace. First, we define the OVERLAP

$$\mathsf{Ov}_W^{(\kappa)} = \mathsf{Overlap}\left(\mathbf{y}(t), \boldsymbol{\zeta}^{(\kappa)}\right)(t) = \frac{1}{N_W}\sum_{\mathsf{u}=1}^{N_W} y_{\mathsf{u}}(t)\zeta_{\mathsf{u}}^{(\kappa)} \tag{17}$$

do which measures the alignment, at time $t$, of the current state of the workspace, $\mathbf{y}(t)$, with the pattern $\boldsymbol{\zeta}^{(\kappa)}$. If it is equal to $+1$, they are perfectly aligned; if it is equal to $-1$, they are completely misaligned (antagonistic).

The dominant pattern in the workspace, $\boldsymbol{\zeta}^{(\kappa_t^*)}$, at time $t$, is the one with the greatest overlap with the current state of the workspace, $\mathbf{y}(t)$:

$$\boldsymbol{\zeta}^{(\kappa_t^*)}, \quad \text{where} \quad \kappa_t^* = \arg\max_{\kappa} \mathsf{Ov}_W^{(\kappa)} \tag{18}$$

Now let's define the "External" Fields, and the Modified Energies.

**External Field in the Workspace.** The Workspace neuron $\mathsf{u}$ receives influences from two main sources:

1. Modulated Input, coming from the Modules (Feedforward). This field represents the aggregated "*bottom-up*" information from all modules to the Workspace neuron $\mathsf{u}$, modulated by the competition function $\mathcal{F}_{\mathsf{m}}(t)$:

$$h_{W;\mathsf{u}}^{\mathsf{mod}}(t) = \sum_{\mathsf{m}=1}^{M} C_{\mathsf{um}} \mathcal{F}_{\mathsf{m}}(t) \cdot \mathsf{o}_{\mathsf{m}}(t) \tag{19}$$

   - $\mathsf{o}_{\mathsf{m}} = \max_{\mu} \mathsf{Ov}_{\mathsf{m}}^{(\mu)}(t)$ represents the activation of module $\mathsf{m}$. This is the aggregated and normalized "signal" that module $\mathsf{m}$ sends.
   - $C_{\mathsf{um}}$ is the connectivity of module $\mathsf{m}$ (as a whole) to neuron $\mathsf{u}$ in the workspace. This scales the signal strength from module $\mathsf{m}$ to neuron $\mathsf{u}$.
   - $\mathcal{F}_{\mathsf{m}}(t)$ is the competition function. It modulates (weights) the contribution of each module, implementing selective attention. More relevant modules (high $\mathcal{F}_{\mathsf{m}}(t)$ will have a greater impact.
   - The summation $\sum_{\mathsf{m}=1}^{M}$, ensures that the Workspace neuron $\mathsf{u}$ receives the weighted contribution of *all* modules.

   *This field implements "selective attention"* bottom-up. This ensures that the Workspace is receiving input that represents the "clarity of a pattern" of the module, and not just its raw activity. The Workspace is now integrating the "information" that the modules are processing, not just their states.

2. Active Attraction guided by workspace patterns (Internal Guidance). Represents the internal "push" of the workspace towards your most active or relevant memorized pattern,
   $\boldsymbol{\zeta}^{(\kappa_t^*)}$.

$$h_{W;\mathsf{u}}^{\mathsf{guide}}(t) = \gamma \cdot \zeta_{\mathsf{u}}^{(\kappa_t^*)} \tag{20}$$

- $\zeta_{W,\mathsf{u}}^{(\kappa_t^*)}$, the component $\mathsf{u}$ of the memorized pattern $\boldsymbol{\zeta}^{(\kappa_t^*)}$, which is the pattern of the workspace with the largest overlap at the moment (or the ignition target). This term defines the "direction of the push".
- $\gamma$ is the force of attraction. A positive value of $\gamma$ tilts the energy landscape, making the energy valley corresponding to $\boldsymbol{\zeta}^{(\kappa_t^*)}$ deeper.

*This field implements the "internal guidance" or "intent" of the workspace to form a specific integrated concept.*

The modified total energy of the workspace, including both external fields, is then:

$$E_W^{\mathsf{mod}}(\mathbf{y}(t)) = -\frac{1}{2}\sum_{\mathsf{u}\neq\mathsf{v}} \mathsf{J}_{\mathsf{uv}} y_{\mathsf{u}}(t) y_{\mathsf{v}}(t) - \sum_{\mathsf{u}=1}^{N_W} \left(h_{W;\mathsf{u}}^{\mathsf{mod}}(t) + h_{W;\mathsf{u}}^{\mathsf{guide}}(t)\right) y_{\mathsf{u}}(t) \tag{21}$$

**External Field in Modules.** Before defining it, let's introduce the following order parameter, which will play a relevant role shortly:

$$\mathsf{o}_W(t) = \mathsf{Overlap}\left(\mathbf{y}(t), \boldsymbol{\zeta}^{(\kappa_t^*)}\right)(t) \tag{22}$$

the overlap of the current workspace state with the predominant pattern.

The neuron $i$ of the module $\mathsf{m}$ receives influences from the workspace via feedback driven by workspace patterns. The external field:

$$h_{\mathsf{m};i}^{\mathsf{fb}}(t) = \delta \cdot \mathsf{o}_W(t) \cdot \sum_{\mathsf{u}=1}^{N_W} F_{i\mathsf{u}}^{(\mathsf{m})} y_{\mathsf{u}}(t) \tag{23}$$

represents the "*top-down*" influence of the active Workspace on the modules. $\delta$ is the feedback strength, $\mathsf{o}_{\mathsf{m}}(t)$ is the overlap of the workspace with the dominant pattern, $\boldsymbol{\zeta}^{(\kappa_t^*)}$ (which indicates the strength and coherence of the top-down signal), and $y_{\mathsf{u}}(t)$ is the state of the neuron $\mathsf{u}$ of the Workspace.

The interpretation is as follows:

- $\sum_{\mu=1}^{N_W} F_{i\mathsf{u}}^{(\mathsf{m})} y_{\mathsf{u}}(t)$ is the weighted sum of the states of all neurons in the workspace, as the feedback connections $F_{i\mathsf{u}}^{(\mathsf{m})}$, from the workspace neuron $\mathsf{u}$ to the module $\mathsf{m}$ neuron. This is the "raw" feedback signal.
- $\mathsf{o}_W(t)$ is the overlap of the workspace with the dominant pattern. This term acts as a modulator of the feedback strength. If the workspace is in a clear state and consistent with one of its patterns ($\mathsf{o}_W(t)$ high), the feedback is stronger. If the workspace is cluttered ($\mathsf{o}_W(t)$ low), the feedback is weaker.
- $\delta$ is the strength of the overall feedback.

*This field implements the "*top-down*" control, aligning the modules with the active integrated concept in the workspace.*

The Modified Energy of each module, including the outer feedback field, is:

$$E_{\mathsf{m}}^{\mathsf{mod}}(\mathbf{x}^{(\mathsf{m})}(t)) = -\frac{1}{2}\sum_{i\neq j} \mathsf{J}_{ij}^{(\mathsf{m})} x_i^{(\mathsf{m})}(t) x_j^{(\mathsf{m})}(t) - \sum_{i=1}^{N_{\mathsf{m}}} h_{\mathsf{m};i}^{\mathsf{fb}}(t) x_i^{(\mathsf{m})}(t) \tag{24}$$

The connectivity of modules to the Workspace is now modeled using the average activity of each module (magnetization), which is consistent with the already adopted definitions of $A_{\mathsf{m}}(t)$ and $\mathsf{Comp}_{\mathsf{m}}(t)$. This keeps complexity at an appropriate level, focusing on the interaction between modules, as functional units, and the workspace.

- Lateral competition, $\mathsf{Comp}_{\mathsf{m}}(t)$, is the engine of selective attention, which manifests itself through $\mathcal{F}_{\mathsf{m}}(t)$.
- $\mathcal{F}_{\mathsf{m}}(t)$ modulates the signal strength from the modules to the workspace, ensuring that it receives filtered and prioritized input.
- Workspace patterns (via $\mathsf{o}_{\mathsf{m}}(t)$) provide the mechanism of internal attraction and guide feedback to the modules, creating a fundamental top-down control loop for integrated cognition and ignition.

This adaptation integrates all functions (lateral competition, competition function $\mathcal{F}_{\mathsf{m}}(t)$ and input modulation) into the model GNW. They work together to model information selection and the emergence of coherent and integrated states.

# 5 Phase Transitions and Consciousness in GNW

$\mathcal{C} = p/N$ is the density of patterns stored per neuron. For each module $\mathsf{m}$, we have $\mathcal{C}_{\mathsf{m}} = \mathsf{p}_{\mathsf{m}}/N_{\mathsf{m}}$, and for the workspace, $\mathcal{C}_W = \mathsf{p}_{\mathsf{W}}/N_W$, which are fundamental control parameters that determine the behavior regimes of the subsystems.

The storage capacity, $\mathcal{C} = p/N$, is also a fundamental control parameter in GNW, directly determining the quality of memory retrieval and directly influencing the order and clarity of the subsystems.

As we have already seen, for a subsystem with $N$ neurons and patterns $\boldsymbol{\xi}^{(\mu)}$, the overlap with a pattern $\mu$ is $\mathsf{Overlap}\left(\mathbf{s}(t), \boldsymbol{\xi}^{(\mu)}\right) = \frac{1}{N}\sum_{i=1}^{N} s_i(t)\xi_i^{(\mu)}$. The order parameter $\mathsf{o}(t)$ is the maximum overlap:

$$\mathsf{o}(t) = \max_{\mu} \mathsf{Overlap}\left(\mathbf{s}(t), \boldsymbol{\xi}^{(\mu)}\right) \tag{25}$$

and its TIME AVERAGE (or ensemble average) is $\langle \mathsf{o}_{\mathsf{m}} \rangle_{\mathsf{temp}}$.

$\langle \mathsf{o} \rangle_{\mathsf{temp}}$ measures the average maximum similarity of the subsystem (module or Workspace) with one of the patterns it has memorized. If $\langle \mathsf{o} \rangle$ is high, it means that the subsystem is "thinking clearly" about one of its "ideas" or memories. $\langle \mathsf{o} \rangle \approx 1$ means that the system is almost always focused on a clear memory. $\langle \mathsf{o} \rangle \approx 0$ is confused or in a diffuse state due to several blurred memories.

The mean squared magnetization per neuron, $\mathsf{q}$, defined by

$$\mathsf{q} = \frac{1}{N} \sum_{i=1}^{N} \langle S_i \rangle_{\mathsf{temp}}^2 \tag{26}$$

where $\langle S_i \rangle_{\mathsf{temp}}$ is the time average of the neuron's state $i$. It measures how "stable" or "polarized" the individual neurons in the subsystem are, over time. If $\mathsf{q}$ is high, it means that the neurons are "frozen" in their positions, forming a stable pattern (whether it's a pure pattern or something more complex). If $\mathsf{q}$ is low, the neurons are fluctuating randomly, not maintaining a stable configuration. $\mathsf{q} \approx 1$: stable neurons, "frozen". $\mathsf{q} \approx 0$: unstable, random neurons. $\mathsf{q}$ intermediate (between $0$ and $1$): some bias exists, but it is not complete.

The averages in $\langle \mathsf{o} \rangle$ and $\mathsf{q}$ are temporal, over a sufficiently long period of time until the system reaches equilibrium.

This section formulates the central hypothesis about the emergence of consciousness in the GNW model, integrating the definitions of modular competition, feedforward connectivity and feedback, the active role of Workspace patterns, and the order parameters $\mathsf{o}$ ($\mathsf{o}_{\mathsf{m}}, \mathsf{o}_W$) and $\mathsf{q}$ ($\mathsf{q}_{\mathsf{m}}, \mathsf{q}_W$). The central hypothesis is that consciousness, or states of integrated cognition, emerges in specific phases of the GNW. It is as follows (focusing on the Workspace):

> *Consciousness arises as the emergence of a globally ordered state of the workspace, actively orchestrated by the dynamics of the GNW and resulting from a phase transition.*

Workspace is the "stage" of consciousness in GNW. It is there that information becomes globally accessible and integrated. The "global order" for consciousness should predominantly be the order of the Workspace. But, despite the focus being the workspace – that's where emergent consciousness arises – which leads us to consider the workspace order parameters, $\langle \mathsf{o}_W \rangle_{\mathsf{temp}}$ and $\mathsf{q}_W$, we must also consider the order parameters of the Modules, as they are essential and indispensable for the Workspace to achieve and sustain this ordered state. In fact:

The Workspace (Maestro) is the primary neuronal correlate for Consciousness (NCC): $\langle \mathsf{o}_W \rangle_{\mathsf{temp}} >> 0$, indicates that the workspace is clearly focused on an integrated (predominant) pattern ($\zeta^{\kappa_t^*}$). He is thinking of a clear concept. $\mathsf{q}_W >> 0$ indicates that the neurons in the workspace are stable and "frozen" in this pattern, ensuring the support and coherence of thought/perception.

Therefore, for GNW to be conscious, the workspace needs to be in this state. If the workspace is disordered or confused, the system is not conscious, regardless of what the modules do.

Modules (Musicians). They are not the NCC of global consciousness, but they are the NCC's of the contents of consciousness.

A module by itself is not conscious, but its information can become conscious. For the workspace to reach the phase where $\langle o_W \rangle_{temp} >> 0$ and $q_W >> 0$, it needs coherent and relevant inputs. These inputs come from the modules. If a module is contributing to the ignition of the Workspace, it needs to be in its own ordered phase – $\langle o_m \rangle_{temp} >> 0$ (for active modules). Or, it needs to provide a "clear signal" (a tuned note) to the workspace. If the modules were confused or disordered, the Workspace would not be able to integrate anything clear. Furthermore, the neurons of active modules must be "frozen" in their predominant pattern, ensuring the stability and quality of the signal sent to the Workspace – $q_m >> 0$ (for active modules).

Consciousness is an integrated phenomenon. The order in the workspace depends on the order of the modules. If these were chaotic or confused (low $\langle o_m \rangle_{temp}$, low $q_m$), the workspace would receive only noise. It would be impossible for the workspace to reach an ordered state. On the other hand, there could be modular order – but without integration. That is, the modules can be in perfect order, each with its own clear pattern, but if there is no effective competition and modulation for the workspace, the information may not be integrated, and the workspace would remain disordered.

Therefore, the order parameters of the Workspace ($o_W >> 0$ and $q_W >> 0$) are the direct and primary conditions for consciousness to emerge in the GNW. However, for the Workspace to achieve and maintain these conditions, it is fundamental that the relevant modules (those that "win" the competition $\mathcal{F}_m(t)$, and whose signals are prioritized by $h^{mod}_{W,u}$) are also in their own ordered phase (with $o_m >> 0$ and $q_m >> 0$). This describes a hierarchical and integrated state of global order. Consciousness is the order of the Maestro (Workspace), but it can only happen if the right Musicians (Modules) are playing their parts clearly, and the Maestro is actively coordinating everything.

## The Three Key Phases in GNW, induced by $\mathcal{C}$, $\langle o_W \rangle_{temp}$ and $q_W$

The variation of $\mathcal{C}_W = p_W/N_W$ (and $\beta = 1/T$, which is assumed fixed for this analysis of $\mathcal{C}$) induces phase transitions, which we can now characterize using $\langle o_W \rangle_{temp}$ and $q_W$. The phases (ordered, confusion, disordered) are macroscopic and equilibrium properties of the system. They do not refer to an isolated instant, but rather to the typical and persistent behavior that the system exhibits under certain conditions.

1. Ordered Phase (Clarity/Awareness). The subsystem retrieves and sustains memorized patterns clearly. There is coherence and focus.

This occurs when $\mathcal{C}_W$ is low ($\mathcal{C}_W < \mathcal{C}_W^c$, where $\mathcal{C}_W^c$ is the critical storage capacity).

Characterization by Order Parameters:

- $\langle o_W \rangle_{\text{temp}} >> 0$: high overlap with memorized patterns;
- $q_W >> 0$: high mean-squared magnetization, indicating that the neurons are "frozen" in the state that forms the maximal pattern.

Cognitive Relevance – clear, focused, and meaningful cognition. A workspace in this phase is capable of supporting a high-level concept.

2. Confusion Phase (Overload/Indecision). The subsystem is overloaded with patterns ($\mathcal{C} > \mathcal{C}_c$). It cannot retrieve pure (or stored) patterns, but it is also not chaotic. You get stuck in "spurious" attractors (attractors that emerge in the network but were not explicitly memorized. They are complex combinations or mixtures of various pure patterns), or in a state where neurons have some bias, but are not globally consistent with a pure pattern. This occurs when the capacity $\mathcal{C}$ is medium-high ($\mathcal{C}_c < \mathcal{C} < \mathcal{C}_L$, where $\mathcal{C}_L$ is the limit capacity.

   Characterization by Order Parameters:

   - $\langle o_m \rangle_{\text{temp}} \approx 0$: low overlap with any pure pattern.
   - $q_W$ is intermediate ($0 < q_W < 1$): there is some polarization in the neurons, but they do not follow a pure pattern, indicating a complex structure without clarity.

     Cognitive Relevance – state of cognitive overload, indecision, or fragmented processing. The workspace struggles to form a clear "ignition," potentially generating confused or inconsistent states.
   - Disordered Phase (Chaos/Unconsciousness). The subsystem is chaotic and random. There is no structure, coherence, or memory. This occurs when $\mathcal{C}$ is very high. ($\mathcal{C} > \mathcal{C}_L$) or the temperature is too high.

   Characterization by Order Parameters:

   - $\langle o_m \rangle_{\text{temp}} \approx 0$: low overlap with any pure pattern.
   - $q_W \approx 0$: low mean square magnetization, indicating that neurons fluctuate randomly between $+1$ and $-1$.

   Relevance Cognitive – state of unconsciousness or absence of cognition. The GNW does not generate or process meaningful information.

## Consciousness and the Phases of the GNW.

The central hypothesis is that consciousness (or states of integrated cognition) corresponds to an ordered phase of the system, actively modulated by the interactions of the GNW.

> *A emergency of consciousness in GNW is characterized by the operation of the system in the* ORDERED PHASE*, where memorized patterns are robustly retrieved and integrated. This phase is distinctly characterized by* $< o >>> 0$ *and* $q >> 0$ *for Workspace and relevant modules, and is actively modulated by the dynamics of GNW.*

## The Engine of Consciousness: Ignition in the Ordered Phase.

The WORKSPACE IGNITION, $\mathsf{Ig}(t) = 1$, represents the dynamic transition to this ordered phase. At this moment, the workspace (with $\langle \mathsf{o}_\mathsf{m}^W \rangle >> 0$ and $>> 0$) forms a clear and integrated concept.

This ignition is not passive. It is actively driven by the external fields defined before:

- $h_{W,\mathsf{u}}^{\mathrm{mod}}(t)$: Input of modulated modules via $\mathcal{F}_\mathsf{m}(t)$ (selective attention).
- $h_{W,\mathsf{u}}^{\mathsf{guide}}(t)$: Active attraction of the workspace patterns (internal guidance for the concept).

At the same time, the active workspace (in the ordered phase) influences the modules through the feedback-induced external field, $h_{k,i}^{\mathrm{fb}}(t)$, directing them towards coherent states.

## Research Methodology.

To test the hypothesis stated above ("*The emergence of consciousness in GNW is characterized by the system's operation in the* ORDERED PHASE") we propose the following computational simulations:

1. MONITOR ORDER PARAMETERS. During the simulations, calculate and record $\langle \mathsf{o}_\mathsf{m} \rangle$, $\mathsf{q}_\mathsf{m}$ for each module $\mathsf{m}$, and $\langle \mathsf{o}_W \rangle$, $\mathsf{q}_W$ for the workspace.
2. VARY $\mathcal{C}$ AND OTHER CONTROL PARAMETERS. Run simulations for different values of $\mathcal{C}_\mathsf{m}$, $\mathcal{C}_W$, $\beta = 1/T$ and $\lambda$.

3. Mapping the Phases in Detail. Using graphs to identify the regions where the system (workspace and relevant modules) enters the ordered phase (high $\langle o_m \rangle$ and $q_m$), the confusion phase (low $\langle o_m \rangle$, intermediate $q_m$), and the disordered phase (low $\langle o_m \rangle$, low $q_m$).

4. Correlate with Ignition and $\Phi$: Verify if the workspace "ignition" and the $\Phi$ peaks occur predominantly in the ordered phase regions, and if they are maximized in the critical region ($\mathcal{C} \approx \mathcal{C}_c$).

Storage capacity ($\mathcal{C} = p/N$) is a fundamental control for the GNW phases. By using $\langle o_m \rangle_{temp}$ and $q_m$ as order parameters, we can accurately characterize the order, confusion, and disorder phases. The hypothesis that consciousness emerges in the ordered phase, actively orchestrated by the GNW's competition and feedback mechanisms, and optimized by operation in criticality, provides robust quantitative support for investigating the relationship between the model's statistical mechanics and high-level cognitive phenomena.

# 6 Plasticity in the GNW Base Model

In the context of the GNW base model, plasticity refers to the ability of the model's connections and parameters to adapt and change over time in response to experience or learning. In biological terms, synaptic plasticity is the basis of learning and memory.

The incorporation of plasticity is an important improvement that can make the GNW model more adaptive, robust, and capable of modeling complex cognitive processes.

Plasticity can be incorporated at several levels:

1. Synaptic Plasticity in Hopfield Modules. Adjust the weight matrices ($J^{(m)}$) of the Hopfield modules to memorize new patterns or strengthen existing patterns.

   Use Hebbian learning rules (e.g., Oja's rule) to update the weights based on neuronal activity.

2. Plasticity in Module-Workspace Connections. Adjust the connection matrices – feedforward, $C = (C_{um})$, and feedback, $F^{(m)} = (F^{(m)})_{iu}$, to optimize the flow of information between modules and the workspace. Allow connections to strengthen or weaken based on the relevance of the information.

3. Plasticity in the Workspace. Adjust the weight matrix, $((J^W_{uv})$, of the workspace to learn new global patterns or refine existing patterns.

## Pseudocode for Stochastic GNW with Incremental Hebbian Plasticity and External Fields

This section presents a detailed pseudocode for simulating the stochastic Global Network Workspace (GNW) model, incorporating incremental Hebbian plasticity. This formulation integrates the dynamics of lateral competition, the active role of memorized patterns through modulated external fields, and the connectivity matrices, $C_{\mathsf{um}}$ and $F_{i\mathsf{u}}^{(\mathsf{m})}$. The inclusion of plasticity allows weight matrices (and therefore, by extension, the memorized patterns) to dynamically adapt to the system's activity, making GNW more flexible and biologically plausible.

For GNW to simulate a more adaptive cognitive system, it is crucial that the memorized patterns (represented by the weight matrices) can change over time. Incremental Hebbian plasticity is a simple mechanism to introduce this "online" learning capability. This pseudocode generalizes the previous definitions to include this dynamic updating of the weight matrices.

The GNW context and notations are those used previously.

## Pseudocode: Simulation of GNW Dynamics, with Incremental Hebbian Plasticity and Incorporation of External Fields

**Inputs:**

- $M, N_{\mathsf{m}}, \mathsf{p}_{\mathsf{m}}, N_W, \mathsf{p}_{\mathsf{W}}$: System sizes and capacities.
- $\beta = 1/T$: Inverse of the global temperature for the Metropolis method.
- $\gamma$: Active attraction force in the Workspace.
- $\delta$: Feedback strength for Modules.
- $\beta_s$: Competition selectivity parameter.
- $\eta$: Learning Rate (Hebbian Plasticity) ($0 < \eta << 1$).
- Cum: Feedforward Connectivity Matrix.
- $F_{i\mathsf{u}}^{(\mathsf{m})}$: Feedback Connectivity Matrix.
- $\mathbb{T}_{\mathsf{sim}}$: Total Simulation Steps.
- $\mathbb{T}_{\mathsf{Msteps}}$: Metropolis Steps by $\mathbb{T}_{\mathsf{sim}}$ step (scans).

**Initialization:**

- For each Module m, initialize $\mathbf{x}^{(\mathsf{m})}(0)$ randomly, with $x_i^{(\mathsf{m})}(0) \in \{-1, +1\}$.

- Generate $\mathsf{p_m}$ initial patterns $\boldsymbol{\xi}^{(\mathsf{m},\mu)}$. These patterns define the initial weights, $J_{ij}^{(\mathsf{m})}(0)$, using Hebb's Rule.
- Initialize $\mathbf{y}(0)$ randomly, with $y_{\mathsf{u}}(0) \in \{-1, +1\}$.
- Generate $\mathsf{p_W}$ initial patterns $\boldsymbol{\zeta}^{(\kappa)}$. These patterns define the initial weights in the workspace, $J_{\mathsf{uv}}^{W}$, using Hebb's Rule.

**Calculation of Metrics and Auxiliary Fields**

- For each Module $\mathsf{m}$:

$$\mathsf{Comp_m}(t) = \mathsf{o_m}(t) - \frac{1}{M-1}\sum_{\mathsf{n}\neq\mathsf{m}} \mathsf{o_n}(t)$$

- For each Module $\mathsf{m}$:

$$\mathcal{F}_{\mathsf{m}}(t) = \frac{e^{\beta_s \mathsf{Comp_m}(t)}}{\sum_{\mathsf{n}=1}^{M} e^{\beta_s \mathsf{Comp_n}(t)}}$$

- Workspace Overlaps with Patterns $\boldsymbol{\zeta}^{(\kappa)}$: For each pattern $\kappa \in \{1, \ldots, \mathsf{p_W}\}$ of the Workspace:

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

- Identify

$$\kappa_t^* = \underset{\kappa}{\operatorname{argmax}} |\mathcal{O}_W^{(\kappa)}(t)|$$

**Calculation of External Fields for the Workspace, $h_{W,\mathsf{u}}$.**

- Feedforward Modulated Input: For each neuron $\mathsf{u}$ in the Workspace:

$$h_{W,\mathsf{u}}^{\mathsf{mod}}(t) = \sum_{\mathsf{m}=1}^{M} \mathcal{F}_{\mathsf{m}}(t) \cdot C_{\mathsf{um}} \mathsf{o_m}(t)$$

- Active Guidance for the Dominant Workspace Pattern: For each neuron $\mathsf{u}$ in the Workspace:

$$h_{W,\mathsf{u}}^{\mathsf{guide}}(t) = \gamma \cdot \zeta_{W,\mathsf{u}}^{(\kappa_t^*)}$$

**External Field Calculation for Modules.**

- Pattern-Driven Feedback: For each Module $\mathsf{m}$ and each neuron $i$ of module $\mathsf{m}$:

$$h_{\mathsf{m},i}^{\mathsf{fb}}(t) = \delta \cdot \mathcal{O}_W^{(\kappa_t^*)}(t) \cdot \sum_{\mathsf{u}=1}^{N_W} F_{i\mathsf{u}}^{(\mathsf{m})} y_{\mathsf{u}}(t)$$

The Metropolis dynamics uses the external fields and the current matrix $J$ to calculate $\Delta E$.

For $N_{\mathsf{Msteps}}$ scans, update the Workspace:

For each neuron $\mathsf{u}$ in the Workspace (chosen randomly), calculate $\Delta E_{\mathsf{u}}$ for $y_{W,\mathsf{u}} \to -y_{W,\mathsf{u}}$ using $J^W(t)$ and $h^{\mathsf{mod}}_{W,\mathsf{u}}(t) + h^{\mathsf{guide}}_{W,\mathsf{u}}(t)$.

Apply Metropolis Rule to update $y_{\mathsf{u}}(t)$.

Updating the Modules: For each Module $\mathsf{m}$ and for each neuron $i$ of Module $\mathsf{m}$ (chosen randomly), calculate $\Delta E_i$ for $x_i^{(\mathsf{m})} \to -x_i^{(\mathsf{m})}$, using $J^{(\mathsf{m})}(t)$ and $h^{\mathsf{fb}}_{\mathsf{m},i}(t)$.

Apply Metropolis Rule to update $x_i^{(\mathsf{m})}(t)$.

**Incremental Hebbian Plasticity – Updating the weights to the next step $t+1$.**

- For each Module $\mathsf{m}$, and for each pair of neurons $(i,j)$ in Module $\mathsf{m}$:

$$J_{ij}^{(\mathsf{m})}(t+1) = (1-\eta) J_{ij}^{(\mathsf{m})}(t) + \eta \cdot \frac{x_i^{(\mathsf{m})}(t) x_j^{(\mathsf{m})}(t)}{N_{\mathsf{m}}}$$

  and

$$J_{ii}^{(\mathsf{m})}(t+1) = 0$$

- For the Workspace. To each pair of $(\mathsf{u},\mathsf{v})$ neurons in the Workspace:

$$J^W_{\mathsf{uv}}(t+1) = (1-\eta) J^W_{\mathsf{uv}}(t) + \eta \cdot \frac{y_{\mathsf{u}}(t) y_{\mathsf{v}}(t)}{N_W}$$

  and

$$J^W_{\mathsf{uu}}(t+1) = 0$$

**Record Data**

- Register $\mathbf{x}^{(\mathsf{m})}(t)$, $\mathbf{y}(t)$, $\mathcal{F}_{\mathsf{m}}(t)$, $\mathcal{O}^{(\kappa)}(t)$, etc. over time.
- Output: Historical data of states, overlaps, and other parameters over time.
- Time series of the states of the Hopfield modules $x_i^{(\mathsf{m})}(t)$ for each neuron $i \in \mathcal{M}_{\mathsf{m}}$.
- Time series of the state of Workspace $y_{\mathsf{u}}(t)$.
- Time series of the average activation of Workspace $a_{\mathsf{W}}(t)$.
- Time series of the ignition of Workspace Ignition$(t)$.
- Evolution of the weight matrices $\mathsf{J}^{(\mathsf{m})}(t)$, $\mathsf{J}^{\mathsf{W}}(t)$, $\mathsf{W}(t)$ and $\mathcal{F}^{(\mathsf{m})}(t)$ over time.

### Clarifications and Impact of Plasticity

Pattern Dynamics: With plasticity, the memorized patterns $\{\boldsymbol{\xi}^{(\mathsf{m},\mu)}\}_{\mu=1}^{\mathsf{p_m}}$, in module $\mathsf{m}$, and $\{\boldsymbol{\zeta}^{(\kappa)}\}_{\kappa=1}^{\mathsf{p_W}}$, in the workspace, which were used for initialization of weights $J^{(\mathsf{m})}(0)$ and $J^{W}(0)$, are no longer the only attractors. The weight matrices evolve, and the system's attractors adapt to the most frequently visited or reinforced states. This represents continuous learning.

The field $h_{W,\mathsf{u}}^{\mathsf{guide}}$: The term $\zeta_{W,\mathsf{u}}^{\kappa}$, for the active attraction (guide) field, now refers to the pattern $\kappa_t$ that was initially memorized. If the patterns themselves are not updated, then this field pulls the system back to the original patterns. If the intention is for the Workspace to be pulled towards new attractors that emerge from plasticity, the term $\zeta_{W,\mathsf{u}}^{\kappa_t^*}$ would need to be redefined to reflect emerging patterns (which is more complex and may be a next step). For simplicity, I kept the term as being pulled towards the initial memorized patterns.

Variation of $N_{\text{MSTEPS}}$: The number of Metropolis steps per $\mathbb{T}_{sim}$ step (scans) is crucial to allow the system to approach equilibrium in each state.

The addition of the plasticity block is what allows GNW to become adaptable and learn, making it a much more dynamic and biologically plausible model.

## 7 Final Abstract

This two-part work, the present one and the one in preparation [JNT2025a], explores the dynamics of neuronal ignition in the Global Workspace (GNW) model, using a statistical mechanics approach with stochastic Hopfield networks. Building on previous work [Tav2025], 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 overall coherence of the system, analyzing the role of Transfer Entropy between modules and workspace, and in the calculation of $\Phi$. 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 in 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":

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

In this way, we identify a complex physical process (high $\Phi$ 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!

IIT, as a (computational) theory of consciousness, makes predictions about the properties of conscious systems (high $\Phi$). 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 $\Phi$, behave. If $\Phi$ 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. It describes the structure of experience (its differentiation and integration) in terms of its underlying causality. This does not explain the leap to the phenomenological character.

The GNW model can reveal dynamic and emergent information integration patterns that are not obvious. "Ignition" can be an elegant mechanism for the formation of a transient "consciousness field". Studying how ignition modulates $\Phi$ can 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.

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 $\Phi$, but rather that $\Phi$ is consciousness. This is the "solution" of IIT, 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, 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; To provide a quantifiable framework for studying the relationship between neuronal 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 $\Phi$), 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.

# 8 Appendix. Mathematical Preliminaries.

## 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:

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

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

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

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

where

$\mathbf{x}(t)$ is the state of the system at time $t$,

$\mathbb{T}$ is the temporal duration of the trajectory. In practice, we approximate this average with a finite time $\mathbb{T}$:

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

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.

Ergodicity is the property that ensures that the time mean is equal to the statistical mean. Formally:

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

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.

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.

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 the neuron $i$ has a certain value ($\mathsf{p}_\mathsf{m}1$), when the system is in the 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:

$$\langle A \rangle_{\text{est}} = \sum_{\mathbf{x} \in \mathcal{S}} A(\mathbf{x}) P(\mathbf{x}) \tag{28}$$

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 states of the neurons by the Metropolis method, as we saw before:

$$\langle A \rangle_{\text{temp}} \approx \frac{1}{\mathbb{T}} \sum_{t=1}^{\mathbb{T}} A(\mathbf{x}(t)) \tag{29}$$

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:

$$\langle x_i \rangle_{\text{temp}} = \lim_{\mathbb{T} \to \infty} \frac{1}{\mathbb{T}} \sum_{t=1}^{\mathbb{T}} x_i(t) \tag{30}$$

where $\mathbb{T}$ is the simulation time, over 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:

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

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 $\mathsf{q}$ is calculated as:

$$\mathsf{q} = \frac{1}{N}\sum_{i=1}^{N}\langle x_i\rangle^2 \tag{31}$$

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

- $\mathsf{q} \approx 1$ 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.
- $\mathsf{q} \approx 0$ 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 $\mathsf{q}$, 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 an equilibrium state. 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 must run several simulations, with different initial conditions, to verify if the results are robust and do not depend on the initial choice of the neuron states.

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

## **Distinction between Ising and Spin-Glass Models: Application to Stochastic** GNW

This section seeks to clarify the fundamental distinction between the Ising model and Spin-Glass systems, explaining why a GNWmodel, with stochastic Hopfield networks, that serve as the basis for associative memory, can exhibit Spin-Glass-characteristic behaviors, despite having a classical formulation as an Ising model.

The Ising model is the simplest model in statistical mechanics to describe magnetic interactions in materials.

- It consists of a set of "spins" (or neurons, in the context of neural networks) $S_i \in \{-1, +1\}$, arranged in a network (for example, a 1D, 2D or complete network).
- The spins interact in pairs. The strength and type of interaction between two spins $S_i$ and $S_j$ is given by a coupling coefficient $J_{ij}$. The total energy of the system is given by:

$$E = -\sum_{\langle i,j \rangle} J_{ij} S_i S_j - \sum_i h_i S_i \tag{32}$$

  where $\langle i, j \rangle$ denotes interacting spin pairs (often just near neighbors), and $h$ is a local external magnetic field.

  In the classical Ising model (e.g., for ferromagnetism), the $J_{ij}$ are typically POSITIVE ($J_{ij} > 0$), which means that neighboring spins tend to align (ferromagnetism). If they were negative ($J_{ij} < 0$), they would tend to misalign (antiferromagnetism).

- At low temperatures, the system transitions to an ordered state (e.g., all spins up or all down), exhibiting a global order parameter such as the average magnetization.

  A Hopfield network is, in fact, an Ising model with long-range couplings $J_{ij}$, between all pairs of neurons, where $J_{ij}$ are defined by Hebb's rule for memorizing patterns ($\boldsymbol{\xi}^{\mu}$). The energy of the Hopfield network is the same as the Ising model, where the energy of the states is given by:

$$E(\mathbf{S}) = -\frac{1}{2} \sum_{i \neq j} J_{ij} S_i S_j - \sum_i h_i S_i \tag{33}$$

  Here, $J_{ij}$ are the synaptic weights, analogous to the $J_{ij}$ of the Ising model. Stochastic dynamics using the Metropolis method is a way to simulate an Ising model at a given temperature.

- Spin-glasses are a specific class of Ising models (or more complex physical models) that have two crucial characteristics:

Competitive Interactions (Frustration): the couplings $J_{ij}$ are a mixture of ferromagnetic ($J_{ij} > 0$) and antiferromagnetic ($J_{ij} < 0$) interactions, often randomly distributed.

Quenched Disorder: the distribution of $J_{ij}$ is fixed and random over time (does not change dynamically), representing a "frozen disorder" in the material.

Combining these characteristics leads to an extremely complex and rough energy landscape, with many notable properties:

- Frustration: it is impossible to satisfy all interactions simultaneously. For example, if $S_1$ and $S_2$ want to align ($J_{12} > 0$), $S_2$ and $S_3$ want to align ($J_{23} > 0$), but $S_1$ and $S_3$ want to misalign ($J_{13} < 0$), then one of them will become "frustrated".
- Many Ground States: in contrast to ferromagnets that have few ground states (all up/all down), Spin-Glasses have an exponentially large number of ground and metastable states, all with very close energies.
- Slow and Complex Dynamics: The network gets "stuck" in local energy valleys (spurious attractors), leading to a very slow and non-exponential relaxation dynamics, characterized by "aging" and "noise memory". Spin-Glass Phase: At low temperatures, Spin-Glasses enter a phase where there is no global magnetic order (like average magnetization), but there is a more subtle "order" characterized by the Spin-Glass order parameter ($q_{EA}$), which measures the correlation between replicas.
- The GNW model uses Hopfield networks, and this is where the connection becomes clear. The weight matrix $J_{ij}$ of a Hopfield network (defined by Hebb's rule) can contain both positive and negative interactions, depending on the memorized patterns. If the patterns are orthogonal, the interactions can be balanced. However, with random and non-orthogonal patterns, the weight matrix will inevitably generate frustration.
- When the storage capacity $\mathcal{C} = p/N$ exceeds the critical limit ($\mathcal{C}_c \approx 0.14$), the Hopfield network enters the confusion phase or Spin-Glass phase. In this phase, the network can no longer reliably recover pure patterns, but clings to many spurious attractors. This is the characteristic of Spin-Glass dynamics.
- The coupling strength $\mathbf{J}_{MW}$ between modules and the Workspace (and the connections within each module and the Workspace) also introduces an element of disorder and complexity. If the coupling is not perfectly homogeneous and positive, it can induce frustration in the global system.

Therefore, when it is said that the stochastic GNW model can be "Spin-Glass", it is because:

1. The underlying Hopfield subsystems can exhibit the Spin-Glass phase when overloaded ($\mathcal{C} > \mathcal{C}_c$).

2. The coupling between the subsystems can introduce additional frustration, creating a globally complex and rough energy landscape, characteristic of Spin-Glass systems.

3. Stochastic dynamics by the Metropolis method allows the system to explore this complex energy landscape in search of low-energy states, which may include these spurious Spin-Glass attractors.

The stochastic GNW model is therefore an Ising model (in its fundamental formulation of spins and interactions), but its complex and potentially frustrating interactions, especially when Hopfield subsystems operate near or above their storage capacity, cause it to exhibit behaviors characteristic of the Spin-Glass phase. Identifying this phase and understanding its role in cognition (and in the possible emergence of consciousness) is an important goal.

## Arguments for assuming that $\mathcal{C} = p/N \approx 0$ in the GNW model

It is legitimate to consider that the capacity of the Hopfield stochastic networks, which constitute the local modules and the workspace of the GNWmodel, tends to $0$. This assumption has important implications for the interpretation and functionality of the model, namely:

1. Assuming that $p/N \to 0$ simplifies the mathematical analysis of the system. It allows the use of mean-field approximations and facilitates the identification of equilibrium states.

2. If the main objective of the GNWmodel is to study the integration of information between the modules and the workspace, as will be analyzed in part 2 of this reserch, instead of the storage of memory patterns within each module, then focusing on a low-capacity regime may be reasonable.

3. The modules should therefore not be interpreted as a storehouse of a large number of memory patterns, but rather as processors that respond to inputs and produce simplified representations, to be integrated into the workspace, which combines the representations of the modules to produce a global representation. Long-term memory can be modeled separately, leaving the modules and the workspace responsible for processing and integrating information in a shorter time horizon.

4. Phase Transition to Confusion. Stochastic Hopfield networks exhibit a phase transition between an ordered state (where memory patterns are stable) and a disordered state (where the network "forgets" the patterns). This phase transition is controlled by the temperature $T$ and the memory

capacity $p/N$. The critical temperature $T_c$ is the point at which the phase transition occurs. Above $T_c$, the network is in a disordered state, and below $T_c$, the network can remember memory patterns.

5. Interpreting the GNW model as a processor and information integrator, the phase transition from the ordered state to the state of confusion (disorder or forgetfulness) should not be seen as the loss of specific memory patterns. It should rather be interpreted as the loss of the ability to form coherent representations in the workspace. If the temperature (or noise) is too high, the representations of the modules become noisy and uncoordinated, and the workspace cannot integrate this information effectively. The phase transition can be seen as a loss of global order or correlation between the modules and the workspace. The disordered state is not just "random noise," but a state where the relationship between sensory stimuli and activity in the workspace becomes weak or nonexistent. The loss of coherence in the overall representation can be seen as difficulty in forming a unified conscious experience.

6. Confusion States: In the confusion state (high temperature or noise), activity in the workspace becomes random and no longer reflects the coordinated activity of the modules. Ignition becomes impossible because there is no coherent representation to be amplified.

# 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.

[JNT2025a] J N Tavares, "*Mente e Consciência. Teoria da Informação Integrada (IIT) e "Global Neuronal Workspace" (GNW) com Redes de Hopfield Estocásticas*", em preparação.

[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.