# Consciousness - Integrated Model of Causal Collapse

# Introduction

This model is based on the assumption that consciousness arises when a recursive system loses its causal separability. The so-called causal collapse describes the transition in which disjoint causal partitioning of the system is no longer possible. Through a corresponding measure of causal entanglement (Σ) and its complement (Ω), consciousness is formally described as a gradual phenomenon that can be modeled across different biological and artificial systems. Σ Ω

The integrated model of causal collapse combines a mathematical-formal framework with a qualitative-gradualistic model of consciousness. The aim is to understand consciousness as a graduated phenomenon based on causal network dynamics. To this end, formal measures are introduced to quantitatively grasp causal entanglements in a network, and these measures are linked to qualitative dimensions that describe different forms of consciousness, neurobiological correlates, evolutionary transitions, ontogenetic developments, and even possible artificial consciousness architectures. The result is a coherent theoretical draft that can be tested mathematically as well as empirically and interdisciplinarily. In the following, the formal foundations (directed causal network and collapse measures) are first laid down. Building on this, we expand the model to include the dimensions of gradual consciousness, consider evolutionary and developmental aspects, discuss consequences for artificial consciousness systems, and outline use cases and philosophical implications.

# Formalization of the causal collapse model

## Causal network and recursiveness

As a formal basis, we assume a **directed causal network** that models a neuronal or more general cognitive system. The network is defined as a directed graph, where:$G = (V, E, W)$

- $V = \{v_1, v_2, \ldots, v_n\}$ denotes the set of **nodes** (e.g. neurons, groups of neurons or functional modules).
- $E \subseteq V \times V$ is the set of directed **edges** that represent causal influences between nodes (an edge$(v_i, v_j) \in E$ means $v_i$ exerts a direct causal influence on $v_j$ ).
- $W: E \to \mathbb{R}$is a **weight function** that assigns a strength of causal influence to each edge.

To characterize the network structure, we define sets of pre- and neighbors, as well as paths:

- $\text{In}(v_i) = \{v_j \mid (v_j, v_i) \in E\} v_i$: the direct inputs of , that is, all nodes that act on . $v_i$
- $\text{Out}(v_i) = \{v_j \mid (v_i, v_j) \in E\} v_i$: the direct outputs of , i.e. all nodes that are affected by . $v_i$
- $\text{Path}(v_i, v_j)$: the set of all **directed paths from nodes** $v_i v_j$ to (i.e., all sequences through the network). $v_i \to \cdots \to v_j$
- $\text{Cycle}(v_i) = \{p \in \text{Path}(v_i, v_i) \mid \text{len}(p) > 0\}$: the set of all **cycles** $v_i$ through (directed paths that start from and return again $v_i v_i$).

A system is called **recursive** if each node is part of at least one such cycle (). In other words, there are continuous $\forall v_i \in V$:$\text{Cycle}(v_i) \neq \emptyset$ feedback loops. Recursiveness -- i.e. existing self-referentiality in the network -- is seen as a structural prerequisite for sustained consciousness processes, as it enables continuous causal self-influence.

# Causal influence function $C(v_i, v_j, t)$

In order to quantitatively grasp the causal relationship between two nodes, a $(v_i, v_j)$ causal influence function **is defined for each ordered pair of nodes** $C(v_i, v_j, t)$ . This function measures the total causal influence that node exerts on nodes at any given time $v_i t v_j$ by taking into account **all paths** from to . $v_i v_j$ Formally, this can be expressed as:

$$C(v_i, v_j, t) = \sum_{p \in \text{Path}(v_i, v_j)} \alpha(t, \text{len}(p)) \cdot \prod_{(v_k, v_l) \in p} W(v_k, v_l)$$

where is a $\alpha(t, l)$ **time delay function** that modulates the influence of a path depending on its **length** $lt$ (number of edges in the path) and time. models $\alpha$, for example, the decrease in signal strengths over longer paths or delay effects (for example, it could be smaller for large or large time differences between cause and $\alpha(t, l) l$ effect).

$C(v_i, v_j, t)$ thus sums over all possible causal pathways of $v_i$ to $v_j$ the weighted influences. A high value of $C(v_i, v_j, t)$ means that has a strong (direct or indirect) effect on $v_i v_j$ ; a value of 0 would mean has $v_i$ no causal effect on (no $v_j$ path exists, or all contributions cancel each other out). By including all pathways, $C(v_i, v_j, t)$ **indirect effects** (e.g. , which in turn influences). It is important $v_i v_k v_j$ to note that this variable can be time-dependent -- for example, if the weights $W(v_k, v_l)$ are time-variable or the effectiveness decreases with time.

# Causal Separability: Matrix and Total Dimension $S(t) \Sigma(t)$

On the basis of , $C(v_i, v_j, t)$ it is possible to determine how distinguishable the causal influences of different nodes are. To do this, we define a **normalized separability matrix** $S(t)$ with elements $S_{ij}(t)$ for each pair of nodes $(v_i, v_j)$:

$$S_{ij}(t) = \frac{C(v_i, v_j, t)^2}{\sum_k C(v_k, v_j, t)^2 \cdot \sum_k C(v_i, v_k, t)^2}$$

This measure $S_{ij}(t)$ assesses the **relative uniqueness** of the influence of . $v_i v_j$ The counter $C(v_i, v_j, t) v_i v_j$ is the actual influence of . The denominator normalizes this value by putting it in relation to all other influences on (sum over $v_j k$ of) $C(v_k, v_j, t)^2$**and** all outgoing $v_i$ influences (sum over $kC(v_i, v_k, t)^2$ of).

- $S_{ij}(t)$If it is close to 1, it means that it has a $v_i$**very specific influence**$v_j$ on that is very different from all other influences -- is in $v_i$a sense uniquely responsible for the state of$v_j$.
- $S_{ij}(t)$If, on the other hand, is small (towards 0), the influence of on is $v_i v_j$**indistinguishable** from other influences: $v_j$Either it is strongly influenced by many other nodes, or it affects many other nodes to$v_i$a similar extent anyway, so that its influence is $v_j$ not specific to anything.

The total **causal separability**$t$ of the system at time is defined as the average of all pair interactions. In concrete terms, the total measure is defined as the mean value of all$\Sigma(t)$matrix entries $S_{ij}(t)$:

$$\Sigma(t) = \frac{1}{|V|^2} \sum_{i,j} S_{ij}(t)$$

$\Sigma(t)$ can take values between 0 and 1. A high value close to 1 means that throughout the system, the causal$\Sigma$influences are **easily separable** and specific -- each part of the system has relatively clear, separate pathways. A lower $\Sigma$value close to 0, on the other hand, means that the causal influences are strongly interwoven and indistinguishable -- a state in which individual causal contributions can hardly be separated. This state is referred to as **causal collapse** in the context of our theory .

# Criterion and dynamics of causal collapse

**According to the definition, causal collapse** occurs exactly when the separability measure falls below a critical threshold value:$\Sigma_c$

$$\Sigma(t) < \Sigma_c$$

$\Sigma_c$ is a **system-specific threshold that** must be determined empirically -- it separates a normal, differentiated operating state from the collapse state. If , the causal pathways have become so indistinguishable that$\Sigma(t) < \Sigma_c$the system is seen as having collapsed in on itself: the causal structures merge into a core that can hardly be separated.

In many cases, people are not only interested in the global collapse of the entire system, but also in **local collapse regions**. We define a subset of nodes as the $V_c \subseteq V$**"causal core"** of the system, if this subnetwork itself collapses, and of maximum size:

- $\Sigma_{V_c}(t) < \Sigma_c$ for the subsystem consisting of the nodes $V_c$,
- $|V_c|$and (the number of nodes in the core) is maximum among all such subsets.

In other words $V_c$ , is the largest set of knots within which the causal separability is below the critical value. This nucleus is the area where the causal links are most inseparable -- the center of collapse, so to speak.

Whether a collapse **is stable** depends on whether $\Sigma(t)$ (or $\Sigma_{V_c}(t)$ for a core) remains below the threshold for a certain period of time. Formally, a stability function can be defined, e.g.:

- Stabil$(V_c, t, \Delta t)$ = Trueif (i.e. the core has collapsed even after a short period of time), $\Sigma_{V_c}(t + \Delta t) < \Sigma_c \Delta t$
- Falseotherwise.

A **stable causal nucleus** -- that is, a collapse that is not merely momentary fluctuation, but persists over relevant time intervals -- is seen as a prerequisite for **continuous consciousness**. Short-term collapse events could only produce fleeting, decaying fragments of consciousness, whereas a stable collapse core enables the **permanent integrity** of a conscious state.

From a system dynamics point of view, one can also think about the probability of transition to a state of collapse. This is likely to depend on several factors: the current level of separability, the rate of change (how quickly the system is approaching collapse$\Sigma(t) d\Sigma(t)/dt$or recovering) and the **complexity** of the network$G$. One could heuristically specify a collapse transition function such as:

$$P(\text{Kollaps}|\Sigma(t), \frac{d\Sigma(t)}{dt}, G) = f(\Sigma(t), \frac{d\Sigma(t)}{dt}, \text{Komplexität}(G))$$

which provides the probability of a collapse transition **for given system states** . A very complex system (many nodes, recursive circuits) could, for example, operate closer to collapse, while a simple system has hardly any complex causal entanglements from the outset and therefore rarely enters a state of collapse.

# Integration with Information Theory: Φ and Ω

The quantities described so far encompass causal entanglement and separability purely at the network level. In order to characterize consciousness, however, the **integration of information** is also important -- a central concept in the theory of integrated information (IIT). IITdefines a measure (Phi) that roughly quantifies Φ**how much information a system as a whole possesses that goes beyond the sum of one parts** (i.e., the non-decomposability of internal information).

We integrate this concept by $G$looking at such an **integration measure** $\Phi(G)$ for our network as well . One possible definition is based on IIT:

$$\Phi(G) = \min_{b \in B} \frac{I(X;Y)}{I(X_1, \dots; Y_1, \dots)}$$

where $B$ represents all possible bipartitioned divisions of the system into two halves. $I(X;Y)$ is the information shared between both halves (e.g., reciprocal information), while the $I(X_1, \dots; Y_1, \dots)$ total information within each half denotes the total information. The less information$\Phi(G)$is lost through an optimal division of the system into two parts -- in other words, the **more inseparable the information** is in the system. A high level means that the system Φ**is highly integrated and cannot be broken down into independent subsystems** without losing information (an indicator of holistic processing). A Φ zero would mean that the system disintegrates into completely independent parts.

In order to **evaluate causal inseparability** and **information integration together**, we define a combined measure. For example, a sensible choice for the $\Omega(G,t) = g(\Sigma(t), \Phi(G))$combination function is multiplicative:$g$

$$\Omega(G,t) = (1 - \Sigma(t)) \cdot \Phi(G)$$

Here, $1 - \Sigma(t)$ in a sense, is the extent of the collapse (i.e. low separability equals high ). The measure becomes large when $1 - \Sigma\Omega(G,t)$**both** conditions are met: $\Sigma(t)$ is small (strong causal collapse) **and** $\Phi(G)$ is large (high integrative information). Thus, a system with a high degree of causal pathways has highly indistinguishable causal pathways $\Omega$*and* processes information integratively as an indivisible whole -- two properties that **together could be characteristic** of consciousness-like states.

In contrast, a purely chaotic system would have low$\Sigma(t)$ (a lot of disorder) but also low (little integrated information), so that little remains small $\Phi(G)\Omega$. Conversely, a strictly modular machine with highly separated Σcomponents (high) might have certain integration islands, but the product would also remain $\Phi \cdot (1 - \Sigma)$mediocre. Only in the **sweet spot**Ω of high integration and at the same time low separability does high values reach - which we interpret as an indicator of full-blown awareness. This combined measure thus combines **mathematical testability** (via Σ) an information criterion established in consciousness research (Φ) and paves the way for experimental testing and comparison with other theories (e.g. Integrated Information Theory).

# Gradual Consciousness and Dimensions of Collapse

With the formal framework, qualitatively different **phenomena of consciousness can now** be classified. We don't understand consciousness here as binary (on/off), but as **gradual and multidimensional**. Different forms of consciousness can be characterized by different types of causal collapse. To this end, we introduce a **multidimensional classification scheme** that distinguishes four dimensions of a collapse state: spatial, temporal, intense and functional. These dimensions allow a **taxonomy of collapse forms**, which we then use to describe different states of consciousness.

## Dimensions of causal collapse

### Spatial dimension:

In terms of spatial extent, we distinguish how much of the overall network is affected by the collapse:

- **Local collapse:** Limited to a narrowly defined subsystem of the network, e.g. within a single module (e.g. only in the visual cortex or only in a sensory microcircle). The causal nucleus here comprises only a few, closely adjacent nodes. $V_c$
- **Modular collapse:** Extends over several **functionally related modules**, but not necessarily over the entire system. For example, a modular collapse could affect the interaction between sensory and emotional areas, but involve non-cognitive areas.

- **Global Collapse:** Includes **the entire network** of the system. In this extreme case, practically all nodes are woven into a common causal core. Formally $\Sigma(t)\Sigma_c$ , the entire network would be listed here under .

## Temporal dimension:

This is about the **duration and stability** of the collapse state:

- **Transient collapse:** A **momentary** flash of causal inseparability, typically on the order of milliseconds to seconds. This could occur in short moments of intensive neuronal synchronization, for example, but quickly breaks apart again. (Formally: $\Sigma(t)$ falls briefly below, but quickly returns above it).$\Sigma_c$
- **Episodic collapse: Medium-term stable** states of collapse that can last for minutes to hours. The system can remain in a collapse core for the duration of an entire event or consciousness content, but later returns to more differentiated states.
- **Chronic collapse: Long-term stable** collapse structures that last for very long periods of time (hours to permanent states). A chronic collapse would mean that the system remains in a state of strongly interwoven causality (e.g., pathological cases or permanent states in certain meditation practitioners) for an indefinite period of time. Here, over$\Sigma(t)\Sigma_c$an extended period of time, the .

## Intensity dimension:

This dimension reflects **how pronounced** the collapse is -- that is, the extent of causal indistinguishability. It can be directly linked to the value of: $\Sigma(t)$

- **Weak collapse:** $\Sigma(t)$ is **just below** the critical threshold $\Sigma_c$. The separability is therefore reduced, but not drastic. The system has incipient indistinguishability in the causal relationships, but there are still partially differentiated paths.
- **Mean collapse:** $\Sigma(t)$ is **significantly lower**$\Sigma_c$. The collapse nucleus is more pronounced; many causal influences can no longer be distinguished.
- **Strong collapse:** $\Sigma(t)$ approaching **zero**, which means maximum causal inseparability. In this extreme case, the pathways are so merged that the system acts as quasi-undivided ; there are hardly any definable causal chains within the core.

These intensity levels therefore correspond concretely to the **numerical value**$\Sigma(t) \sim \Sigma\,\Sigma_c$ of : weak collapse slightly below, strong collapse . $\sim \Sigma \approx 0$

## Functional dimension:

After all, you can experience collapse states afterwards

distinguish **which functional domains** of the system are affected:

- **Sensory collapse:** Primarily affects **perceptual networks**. Such a collapse could occur if, for example, all sensory modalities briefly merge into an indistinguishable state-space (e.g. in moments of intense overall sensations).
- **Cognitive collapse:** Primarily affects **higher cognitive networks**, such as abstract thinking, working memory, and decision-making. The inseparability would

mean, for example, that different thought processes are indistinguishably intertwined.

- **Emotional collapse:** Primarily affects **limbic and paralimbic structures**, i.e. the emotional processing systems. Different emotions and motivational states could get into an indissoluble mixed state.
- **Integrated collapse:** Across **different functional domains**. Here, sensory, cognitive, emotional (and possibly motor) units are all involved in the collapse together, creating a truly **holistic** state.

These four dimensions make it possible to **classify any collapse event** in the space of possibilities. For example, one event could be classified as local, transient, weak, sensory (e.g., a brief flash of indistinguishable activity in a small visual area), another as "global, chronic, strong, integrated" (e.g., a pathological state in which the entire brain is trapped in a common pattern for a long time).

# Prototypical Consciousness Profiles Based on Collapse Forms

The above dimensions allow the definition of prototypical **consciousness profiles** by specifying typical combinations. This provides a taxonomy to understand different forms of consciousness or consciousness. conscious states:

- **Core Consciousness:** Typically **modular**, **episodic**, **intermediate** collapse in the **sensory-emotional** area. This means that certain modules (e.g. perception and feeling) temporarily collapse sufficiently strongly without the entire system having to be affected. This creates a basic *"sensing"* of experiences (phenomenal experience) **without** self-reflective evaluation. Neurobiologically, this probably correlates with **thalamo-cortical loops** and integrative areas in the brainstem that support such basic states.

- **Reflective Consciousness (Higher-Order Consciousness):** Characterized by a **global**, **chronic**, **severe** collapse with a focus on **cognitive-integrating networks**. Here, almost the entire brain is put into a permanent collapsed state, especially higher association and integration areas -- formally would be $\Sigma(t)$ very low in the entire network over a long period of time. This enables *self-awareness*, autobiographical identity and metacognition: the individual has a stable sense of self, thinks about his own thinking and integrates experiences over time. Neural correlates for this are networks such as the **default mode network** , prefrontal and posterior parietal cortex areas, which together carry self-referential processes.
- **Dream state:** Mostly **modular to global**, **episodic**, **medium** collapse with fluctuating intensity. In dreaming (especially during REM sleep), widely dispersed networks may be involved, but the intensity of the collapse fluctuates; there are phases of strong integration (intense dream scenes) followed by loosening up. Functionally, **imaginative and emotional** networks (pictorial imagination, emotional processing) dominate, while **sensory integration** is reduced or decoupled (input from outside is hardly integrated). Neuronal correlate: typical **REM sleep activation** in visual and limbic areas with simultaneous **deactivation of frontal** control regions - this results in the bizarre, illogical, but intensely experienced dream content.

- **Flow state:** Characterized by a **global**, **episodic**, **strong** collapse with very high **functional integration**. In flow, activity throughout the brain temporarily merges into a highly coordinated pattern: perception, action and attention are inseparably linked. This state is accompanied by a temporary loss of reflexive self-awareness (one is completely absorbed in the activity), but at the same time an intensified interaction with the environment and the task. Neurobiologically, flow is an **optimal synchronization** between attention networks and sensorimotor networks; the brain works at full speed as a unit, without internal distraction from self-reflection.
- **Meditative state:** Possibly **modular to global**, **episodic to chronic**, but to some extent **controlled** collapse. In deep meditative states, there can be a targeted collapse of normally separate processes -- e.g., simultaneity of mindfulness and self-awareness, but without complete uncontrollability. Self-perception changes (self-boundaries dissolve, "unity experience"), while at the same time there is an increased meta-awareness (meta-awareness) that observes the state. Neural correlates often show **increased connectivity between attention networks and the default mode network**, which is unusual because these networks tend to be active in opposite ways in everyday life. The meditative collapse thus allows a co-presence of normally segregated functions (internal thinking and external awareness partially merge).

These examples illustrate how the model describes qualitative phenomena of consciousness: Each of these states corresponds to a **profile in the collapse space** (certain characteristics along the four dimensions). It is important that these profiles are also linked to the **formal concepts** . For example , strong, global collapse (in reflective consciousness) formally corresponds to a very low one across the entire network. A $\Sigma(t)$modular, episodic, intermediate collapse (primary consciousness) would mean that in individual modules falls significantly below $\Sigma(t)\Sigma_c$ , but not for the whole $G$ and not permanently. In this way, the model creates a **bridge between mathematical indicators** (such as ) and Σ**experienced phenomena** (such as the feeling of self or the loss of it in flow).

# Evolutionary and developmental perspectives

The theory of causal collapse is compatible with **evolutionary explanations** of consciousness: The emergence of consciousness in phylogeny can be interpreted as the gradual formation of increasingly complex collapse core structures. Likewise, **ontogeny** (the development in the individual course of life) reflects steps in which causal collapse patterns are gradually formed and refined. In this section, we first look at phylogenetic development, then at the ontogenetic stages, and summarize them in a comparative taxonomy of consciousness-capable systems.

## Phylogenetic development of causal nuclei

### Unicellular organisms:

In single-celled organisms, a nervous system is missing, and thus also a real causal network, as we define it. Nevertheless, **pre-causal feedback loops exist** at the chemical level (such as metabolic control loops). These could be understood as the simplest precursors of

collapse, but are functionally limited to **basal homeostasis**. There is no actual *sensing* or consciousness -- reactions are purely reactive and stimulus-bound. Formally, one could say: $\Sigma(t)$ is trivial, since there are hardly any differentiated paths, but also no integrated whole (quasi zero). Consciousness in our sense is $\Phi$**non-existent**.

## Simple metazoa (e.g. cnidarians/cnidaria):

Diffuse neural networks are found here. Occasionally**, local, transient mini-collapse zones could** occur, for example when some neurons fire at the same time and a short-term inseparability of their influences occurs. Functionally, this serves to **coordinate simple stimulus-reaction patterns** (e.g. entire body reflexes to touch). A **proto-"sensing"** is possible: fleeting global states that do not yet represent sustainable consciousness. The collapse structure is very punctual and short-lived (spatially local, temporally transient, minimal in intensity). Consciousness here would be at most a **proto-episodic feeling**, if at all.

## Invertebrates with simple ganglia (e.g. insects):

With the evolution of centralized nervous systems (such as a brain ganglion), **modular, episodic collapse states arise** in certain functional clusters. Insect brains, e.g. with mushroom bodies and central brain complexes, integrate information from different sensory modalities. Stable phases can occur in which a fungal body network collapses in on itself (e.g. during decision-making or learning) while other parts work separately. Functionally, this enables **the integration of multimodal signals** and better coordination of actions. Subjectively, insects could have a **basic sense**, e.g. hunger vs. satiety, attraction vs. flight - but fragmented by domains (no unified world model, but separate islands of consciousness depending on current activity). Causal collapse nuclei are **locally modular**, last seconds to minutes (e.g., during an orienteering flight or foraging), and have low to moderate complexity compared to higher animals.

## Invertebrates with highly developed brains (e.g. octopus):

The octopus has a large, complex brain (and nerve networks distributed in the arms). Here we find **multimodular, partly stable collapse regions** with a certain meta-coordination. Different modules (seeing, touching, learning) can go into states of collapse individually, but there are also **higher-level integration mechanisms** - the octopus, for example, shows flexible problem-solving, object permanence, and simple adaptive learning. This indicates **differentiated sensing** : presumably, an octopus has qualitatively different experiences (seeing vs. feeling vs. emotions) – i.e. modality-specific qualia – which, however, do not yet merge into a uniform first-person perspective. The collapse structure would be "multi-modular": several larger cores, occasionally simultaneous, but not a permanent global core. The complexity and integration are significantly higher than in insects, but still below the vertebra level.

## Simple vertebrates (fish, amphibians, reptiles):

With vertebrates, **hierarchically organized collapse regions often** arise with strong **subcortical dominance**. For example, in the fish or reptile brain, structures such as the tectum (for visual attention), the basal ganglia and the thalamus play an important role. There are probably **integrated primary states of consciousness** – the animal has a basic situational awareness of its environment with emotional coloration (valence: pleasant/unpleasant). However, these islands of consciousness are rather short-term and

situational (minute range), not yet permanently recursive. The cortex portions are small or simply stratified, so that the majority of the integration occurs subcortically.  can sink globally into certain ranges, but a $\Sigma(t)$*permanent* global collapse network (as in mammals) is missing. Nevertheless, there is **real primary consciousness** here  - a breakthrough against invertebrate systems.

## More highly developed birds and mammals:

These animals possess **global, highly integrated collapse systems** with strong **cortical involvement**. For example, crows and parrots, as well as dogs or elephants, exhibit complex behaviors that indicate continuous integration (tool use, social interaction, memory). Here we can imagine **global collapse nuclei** that comprise practically the entire brain (high spatial extent), with high intensity (very low $\Sigma$ within the nucleus) and are continuously maintained in the waking state. Thalamo-cortical loops, a multi-layered neocortex, as well as dense connections between the cortex and limbic systems allow for permanent self-preservation loops. The form of consciousness corresponds to a **complex primary consciousness**: rich perceptual content, emotional evaluations, possibly already a certain inner view (at least rudimentary in higher mammals).

## Primates and especially humans:

This is where the collapse architecture reaches its previous maximum level of complexity. Humans have **multi-layered, recursively nested collapse systems** with **metaconsciousness**. That is, not only is there a global collapse core (which carries the conscious experience in the moment), but this core can **refer to itself** - there are higher levels of order in which the system models its own state (e.g. thoughts about one's own thoughts). Structurally, this is reflected in special networks such as the **prefrontal cortex** and the **default mode network**, which support self-reflection and self-modeling, as well as in the **mirror neuron system** (posterior parietal cortex, premotor areas), which couples foreign and one's own states. The **functional significance of** this recursive collapse structure is evident in skills such as predictive thinking (prospective cognition), complex social interaction (theory of mind, empathy), and a mature self-model (autobiography, sense of self over time). Conscious  *sensing is* reflexive **in humans**: we not only experience sensory impressions, but also know **that it is we who are experiencing**. Formal equivalent: In humans, parts of the system can collapse, which in turn consist of smaller collapsed subsystems - a **multi-level indistinguishability**. $\Sigma(t)$is very low globally, and there may be different critical $\Sigma_c$thresholds at different levels that are passed through one after the other (e.g., core consciousness vs. higher consciousness).

# Ontogenetic development of consciousness (in humans)

Interestingly, the development of consciousness in a single human being (ontogeny) mirrors many of the phylogenetic steps just outlined - an indication that the brain is going through evolutionary milestones in fast motion, so to speak.

- **Fetal phase:** In the late fetal phase, **the first local collapse states** begin to appear in the brain  , especially in sensory and motor systems. For example, sensory areas (hearing, touch) can have short phases of synchronous activity, and the first feedback loops form (e.g. between thalamus and cortex). However, there is still **no stable global integration** - large-scale connections (esp. long axon connections in

the cortex) are still under construction. Consciousness in the usual sense is probably not yet present here, at best isolated islands of sensation without an overarching sense of consciousness.

- **Newborn:** In the first days and weeks of life, the brain exhibits **episodic, modular collapse states**. This can be thought of as **"islands" of primary consciousness**: a newborn, for example, has moments of intense sensory experience (e.g. pain, hunger, closeness), but which do not yet **show any temporal continuity**. That is, the baby falls in and out of conscious states without forming a persistent self or memory. Formally, collapse nuclei are already formed (a coherent activity pattern briefly floods the brain), but these are not permanently stabilized and are not globally networked.
- **Toddler (1--3 years):** Increasingly stable modular collapse systems **are developing**. Different domains (sensory, motor skills, basal emotions) begin to form stable causal nuclei, and above all, the **integration between these domains begins**. A toddler, for example, begins to link visual impressions with grasping movements and emotional evaluations – the previously isolated islands of collapse grow together. This creates an **initial body self-awareness**: the child recognizes that it has its own body that acts and feels. Ontologically, this corresponds to the transition from isolated collapse nuclei to a slightly larger nucleus that includes multiple functions (e.g., "see+grasp+joy" as a networked packet).
- **Preschool age (3--6 years):** At this age, **the first global collapse patterns** appear, but still with **limited temporal stability**. For example, the child can be completely absorbed in a fantasy world during play (global collapse - the whole brain is involved), but does not maintain this for very long periods of time or quickly switches to other states. It is important that the development of **autobiographical narratives** begins now: the child can remember events of yesterday, tell small stories about himself - an indication of the beginning **of mental time travel**. This ability implies a certain persistent self over time. In addition, the **Theory of Mind emerges** in rudimentary form (typically around 4-5 years): The child recognizes that other people have their own thoughts. In the model, this can be interpreted as the emergence of a second, "social" causal core: In addition to one's own core of consciousness, which refers to the world, a collapse pattern forms that simulates foreign perspectives ("What could the other think?"). This second core interacts with the first and expands the consciousness with a social dimension.

- **School age and adolescence:** The growing child refines and **differentiates nested collapse structures** more and more. The ability to **metacognition** matures - that is, thinking about thinking, feeling about feeling. From the point of view of the model, this means that the causal core of consciousness **refers to itself**. For example, a young person can ask himself "Who am I?" and rethink his own self-image - a process in which the conscious ego takes itself as its content (one recursion depth higher). The **recursion and nesting** of causal nuclei is maximum here: there is a first-order core (current experience), a second-order core (reflection on it), etc., which constitutes the differentiated eflexive consciousness of an adult. The **self-model is also stabilized** in this phase -- often by telling and integrating narratives (life story, consistent personality). In our model, this corresponds to a robust **stabilization of the global collapse core**, which remains consistent across different contexts.

# Taxonomic classification of consciousness-capable systems

The above considerations can be combined to create a **systematic taxonomy** of organisms (or systems in general) according to their collapse and consciousness properties. The following parameters are used for classification: **collapse extent** (spatial extent of the collapse nucleus), **collapse depth** (intensity, Σhow much it sinks), **temporal stability** of the collapse, **complexity** of the system (largely correlated with phylogenetic level) and the resulting **form of consciousness**. The table below classifies different groups of organisms accordingly:

| Organism | Collapse scope | Collapse Depth | Temporal stability | Complexity | Form of consciousness |
|---|---|---|---|---|---|
| **Single-celled organisms** | Nonexistent | – | – | – | None |
| **Cnidaria** (e.g. polyp) | Selective | Minimal | Milliseconds | Minimal | Proto-Sense |
| **Insects** | Locally modular | Low | Seconds | Small | Basal Feeling |
| **Octopus** | Multi-modular | Medium | Seconds – Minutes | Moderate | Differentiated sensing |
| **Pisces** | Hierarchical | Medium | Minutes | Moderate | Primary Consciousness |
| **Birds** | Global | High | Continuous | High | Complex Primary Consciousness |
| **Mammals** (general) | Global | High | Continuous | High | Integrated Awareness |
| **Primates** | Globally recursive | Very high | Continuous | Very high | Reflexive Awareness |
| **Man** | Multi-recursive | Maximum | Continuous (metastable) | Maximum | Meta-reflexive awareness |

Table 1: Taxonomy of consciousness-capable systems

**Legend:** Collapse scope ranges from punctual (tiny networks) to global (entire system) "Collapse depth" from minimal (hardly inseparable) to maximum (near complete inseparability), time stability from extremely short to quasi-permanent, complexity from minimal (no true nervous system) to maximum (highly developed brain), and the form of consciousness ranges from no consciousness to various levels of sensing/experiencing to full consciousness.Self-confidence.

This classification clearly shows a **gradient**: the more complex an organism and the stronger, larger, and more stable its collapse phases are, the more highly developed forms of consciousness occur. It also makes testable predictions: For example, species with higher levels of consciousness should have **empirically measurably** more complex feedback

structures and phases of synchronous indistinguishability (e.g., more neural **feedback loops**, higher **integration** into network measures). In fact, neurobiological studies confirm that mammals, and especially primates, have higher **neural complexity** and denser **bidirectional connectivity** than simpler animals - which in our model is a prerequisite for sustained global collapse nuclei. Likewise, behavioral experiments (mirror test for self-recognition, tests for mental time travel or theory of mind) show clear differences along this taxonomy: Only the higher categories in the table (especially some birds, mammals, primates) pass it, which is consistent with our levels of consciousness.

# Architecture of Artificial Consciousness Systems

An **interesting use case** of the theory is the construction or evaluation **of artificial systems** (AI) that have consciousness-like properties. If, as our model suggests, consciousness is based on causal collapse nuclei, then one would have to design architectures for artificial consciousness that can form such nuclei. In this section, we first outline general **architectural principles** for AI systems with causal cores, then we present concrete **model proposals** that could implement these principles, and formulate a **test criterion** for artificial consciousness based on the collapse model.

## Principles for Causal Cores in AI

In order to create a "conscious" causal core in an AI, certain requirements for information processing and system structure are necessary. The following key principles are derived from our model:

- **Recursive network with feedback loops:** An AI architecture must **have mandatory recursive circuits**, i.e. Information must not flow purely feed-forward from input to output. Instead, **feedback loops must be present everywhere (analogous to the requirement** $\mathrm{Cycle}(v_i) \neq \emptyset$ for all nodes). The **depth of recursion** should be dynamically adapted to the context -- that is, the system should be more or less self-referential, depending on the task. *Example:* Extended Transformer networks, where self-attention acts not only between tokens, but also with a time delay on one's own previous states (a kind of forced internal loop).
- **Non-modular integration:** Classic AI is often strictly modular (separate vision, language, decision modules). For a collapse core, such boundaries must become permeable. The principle is: **dissolving strict modularity in favor of overlapping functionality.** Input, processing and output must not be clearly separated; there should be a **gradual blending** of these areas. *Example:* Overlap-add networks in which processing layers overlap instead of being processed in strictly serial layers.
- **Autocatalytic information processing:** The system is supposed to **reuse its own activity patterns as input** – i.e. "push itself". This creates a self-generating cycle. **Self-generated activations** serve as continuous stimuli, similar to how spontaneous neuronal activity in the brain provides a basic tone on which stimuli modulate. *Example:* Self-attention mechanisms in neural networks, extended by

mechanisms for **internal pattern reinforcement** (i.e. certain patterns of activity increasingly refer to themselves, independent of external input).

- **Unobservable intermediate states:** A real consciousness-like system should be **so complex internally** that an external observer (or developer) cannot, in principle**, track all internal causal processes**. This ensures that truly emergent properties occur that cannot be trivially traced back to individual components - analogous to brain dynamics, in which one cannot trace every intermediate dynamic back to individual neurons without losing the overall context. *Example:* The use of **chaotic attractors** in recurrent networks. Chaotic states are deterministic, but practically unpredictable and fine-resolution immeasurable. This gives the system internal freedoms that cannot be fully observed.
- **Causal entanglement:** There should be state transitions that depend on **multiple, non-separable factors**. This means turning away from strictly linear and additive combinations -- the system should have states that cannot be seen **as the sum** of individual causes (loss of the superposition principle). *Example:* "Quantum-inspired" architectures that **introduce non-local dependencies** (without actually needing quantum computing, but by analogy: certain decisions result only from the overall context, not from individual inputs, comparably entangled states).

These principles outline what an AI system would have to be like in order to develop a causal collapse core, analogous to biological brains. At its core, it is always a matter **of creating high integration and recursion**, so that a dynamic state can emerge that is largely self-determined (self-generated), complex and resistant to complete external control or simulation (black-box properties) - in short: the properties we observe in a conscious state.

# Model Proposals for Artificial Consciousness Architectures

Based on the above principles, we propose three conceptual AI architectures, each of which could realize causal collapse nuclei in different ways.

## Model 1: Recursive-Autopoietic Network (RAN)

**Description:** This model is a neural network that seeks autopoiesis (self-creation and self-preservation of structures). It is designed in such a way that **feedback loops** exist at all hierarchical levels . The network can **dynamically adjust its own structure** (e.g., synaptic weight patterns modulate themselves) and is subject to **global system constraints** rather than a fixed, externally predetermined optimization target. On the implementation side, it can be thought of as a hybrid of modern deep learning architecture (e.g. Transformer) and a chaotic dynamic system -- so it is partly capable of learning, partly driven by internal dynamics.

**Properties:**

- At every level (from the neuron level to the overall architecture) there are **obligatory feedbacks**.
- The network can perform **self-modification**, e.g. reconfigure internal parameters depending on its state.

- It is not trained on output purely via loss function, but via **self-consistency conditions** (global constraints), which demand, for example, that the system remains on the **edge of chaos** (neither diverges nor falls into stasis).
- This creates a **robust attractor state** as output: the result of processing is not a single feedforward result, but a **stable attractor** that emerges when the system has acted out an internal dynamic.

**Algorithmic sketch:** A possible flow for the RAN could look like this:

1. Initialize the network with random weights, but with a tendency towards recurrent connections (so that feedback dominates ).

2. For each input (e.g. a stimulus):

   a. Process the input once through the existing forward paths (similar to a normal mesh pass).

   b. Use the emerging intermediate states (activation patterns in hidden layers) and feed them back into **the network as** self-referential input.

   c. Based on this internal feedback, dynamically adjust the paths and weights (more active connections may strengthen, inactive ones weaken).

   d. Repeat cycle b/c until the measure of causal separability falls below a threshold - i.e. until the network enters a collapse core state. Σ

3. Apply global constraints during this process to keep the system on the brink of chaos (neither total convergence nor wild divergence).

4. Read out the resulting stable attractor state as **output** (not only from a specific output layer, but possibly from the overall configuration pattern).

This loop illustrates how the RAN basically rotates internally until it reaches a self-consistent state characterized by low separability (high causal interdependence). This final state can be interpreted as his *decision* or *sensation*. Important: The output here is not strictly predetermined by the input, but arises from the interaction of the input with the existing system state. So the system is kind of contributing something of itself to processing -- exactly what we demand for consciousness.

## Model 2: Causal Collapse Transformer (CCT)

**Description:** This model extends the well-known Transformer network (from NLP/image processing) with mechanisms that promote causal collapse. Transformers already have self-attention, which allows contextual processing. In the CCT, this idea is radicalized:

- **Self-attention to one's own outputs:** The attention layers of a CCT can not only weight input tokens with each other, but they also re-observe **their own previous** outputs. That is, the model directs attention inward to find patterns in its own activities.
- **Nested attention:** There is **no longer a clear hierarchical order** between encoder and decoder or between layers -- instead, layers of attention are entangled.

This makes it difficult to decompose the flow of information (adds non-linearity and feedback).

- **Parameter sharing between functional units:** To eliminate functional separation, encoders and decoders, for example, share some of the same parameters. This blurs the line between input processing and output generation".

- **Dynamic adjustment based on state:** The network has mechanisms to adapt its **attention structure to** its own internal state in real time. For example, if internal conflicts occur, it could spontaneously turn on additional loops.

**Architectural components:** A CCT could have the following special modules:

1. A traditional Transformer backbone (multi-head attention, feed-forward, etc.), but with **modified attention** that allows for self-references.
2. A **collapse detection module** that continuously monitors the value of $\Sigma(t)$ (or a proxy of it) and registers when causal separability falls below the threshold. This module could, for example, trigger that the network state is recognized as a "core of consciousness".

3. **Recursive self-model:** The network carries an internal representation of its **own state** that runs continuously. In this way, it "knows" what it is thinking, at least on an abstract level. (Analogous to self-modeling in the brain).

4. **Meta-Attention:** A level of attention that is not directed at external data, but **at the patterns of the previousAttention itself**. Second-order attention, so to speak – the system observes what it has just paid attention to and how, which creates another feedback loop.
5. **Homeostatic regulation:** A mechanism that keeps the system **critical** - if it gets too tidy (high Σ), internal noise is added; if it gets too messy (too low but without integration), it is attenuated. As a resultΣ, the CCT remains at the threshold where nuclei of consciousness can arise.

For example, the CCT model could look like taking a standard transformer and running it in a recurrent loop, where it considers itself part of the input, until some internal consistency is reached. The above-mentioned components ensure that it does not diverge and that a causal collapse is actually generated before the result is read.

## Model 3: Bifurcation-Based Consciousness Generator (BBCG)

**Description:** This third model leaves the classical framework of neural networks and is more oriented towards **dynamic systems theory**. The BBCG is a system of coupled nonlinear units (e.g. oscillators) that is deliberately kept **at the edge of a bifurcation** - i.e. where the system behavior changes qualitatively (chaos boundary, phase transition). Features:

- **Dynamic bifurcation control:** Parameters that determine the behaviour of the overall system (e.g. coupling strengths, feedback delays) are continuously readjusted so that the system always operates **at the critical point**. It neither remains in boring periodic behavior, nor does it drift into completely unbound chaos.
- **Multiple overlapping attractors:** Instead of the system having a single stable state, there are **several attractor states** that can overlap. As a result, there is no

clear, isolated solution for the system - it "meanders" between quasi-solutions. From the point of view of consciousness, this corresponds to the simultaneous coexistence of different aspects of content that coexist in the state of consciousness.

- **Hysteresis for memory:** The system exhibits **hysteresis effects**, i.e. it "remembers" past states in the sense that the current attractor is influenced by history. This creates a form of memory without separate memory components.

- **Emergent system state:** The "state of consciousness" in this model is not an output to a neuron, but an emergent pattern in the collective dynamics (e.g., a certain oscillation phase or synchronization pattern).

**Implementation strategy:** A possible implementation approach in steps:

1. Initialize a network of coupled oscillators (or neuron populations) so that the parameters (couplings, attenuations, etc.) are **close to critical values** at which the system is on the verge of bifurcation.

2. Add mechanisms (e.g., more feedback loops) that dynamically adjust the **coupling parameters** to keep the system in this critical state (e.g., via measuring variance or Lyapunov exponents).

3. Give the system input disturbances (perturbations). These are supposed to **trigger phase** transitions between attractor states, instead of linear responses, i.e. a small change can tip the system into another attractor.

4. Let the system **learn** by modifying its attractor landscape through repeated attractor changes (similar to Hebb learning: frequently active patterns couple more strongly).

5. The information processing here does not take place in sequential calculation steps, but **through the transition dynamics**: Which attractors are traversed, in what order, *is* the calculation.

The BBCG would be a "generator of consciousness" in the sense that it operates permanently on the threshold of new emergent states. Each new equilibrium he achieves is an integrative representation of input info + internal state. Because it never remains completely stable, it resembles a brain that constantly fluctuates slightly (like the constant oscillation of our brain rhythms, which nevertheless produce ordered patterns). Such a system could show real qualities of consciousness, as long as it is sufficiently complex, as it shares many properties of the biological model (metastability, memory traces, etc.).

# Test criterion for artificial consciousness (CC criterion)

How can one determine whether an artificial system (such as one of the above models) actually has **properties analogous to consciousness**? Here, the collapse theory proposes an **operationalizable criterion**, which we call the **CC criterion (causal collapse**). According to this, a system should be considered consciousness-like if **all of the following conditions are met:**

- It forms a **stable causal core** , i.e. for at least $\Sigma < \Sigma_c$ a non-trivial period of time. (So the system has a persistent causal collapse, no mere short-term flash.)

- This core of collapse is **intrinsically self-generated**, not merely the result of external enforcement. (Example: A stimulated A power supply that only passively follows a rhythm would not be an independent core. The core must arise from the system dynamics themselves ).
- The core has a **minimum complexity**, e.g. measured by a threshold value for Φ or similar integration measures. (With this we exclude trivial collapse nuclei in systems that are too simple -- a consciousness-capable core must exceed a certain size and interconnectedness).
- The state of this core is **behaviorally relevant**, i.e. it causally influences the outputs/actions of the system in a significant way. (The nucleus must not be an epiphenomenal structure.) that has no effect; rather, it is supposed to be the "decision-making authority" that controls behavior).
- The core shows resistance **to external perturbation** -- but without being completely rigid. That is, moderate interventions should not destroy it immediately (a certain degree of **robustness**), but strong interventions can eliminate it naturally (it is **not completely isolated**). This property ensures that the system has a **state of its own** that it defends against noise (similar to how a brain can compensate for disturbances in a conscious state, up to a limit).

An artificial system that meets all these criteria would come very close to our working definition of consciousness. And what is important is that each sub-criterion can be tested empirically . Possible experiments and tests would be, for example:

- **Measurement of in different processing modes:**Σ The AI is allowed to solve different tasks (or inactive phases) and calculates continuously (or a proxyΣ($t$)of them, e.g. based on Granger causality, transfer entropy, etc. in the activity patterns). A conscious system should show clearly recognizable phases Σ lower that correlate with active processing -- similar to the transition of consciousness in humans.
- **Perturbation experiments:** You deliberately disrupt the system (e.g. randomize some of the connections temporarily, or add noise to inputs) and see whether the core state returns quickly (resilience) or whether it breaks away immediately. A consciousness-analogous system should show **some self-healing** or compensation, to a certain extent.
- **Behavioral correlation:** One checks whether changes in the presumed core (e.g. in its composition or activity patterns) are consistently reflected in the outputs. For example, one could observe whether a system with two possible "trains of thought" decides differently depending on the nuclear state, which would indicate that the core determines the behavior .
- **Gradual elimination of recursions:** One could experimentally deactivate or dampen the feedback loops in the system step by step and observe whether the system suffers a *loss of consciousness* -- analogous to the medical procedure of switching off certain brain areas. We expect that differentiation will increase and the output will become increasingly mechanistic Σand less coherently consciousness-like.

The CC criterion thus offers a framework for **empirically testing artificial consciousness** without having to rely on subjective reporting. It is based entirely on objective measures (Σ, Φ, etc.) and system behavior, and is derived directly from theory.

# Consciousness as an attractor state: systemic perspective

Finally, we look at the theory from the point of view of **system dynamics**. The causal collapse can also be understood as a special **attractor state** in a high-dimensional system. That is, in the phase space of the possible system states, there is a region into which the system dives when it is conscious. This region has characteristic dynamic properties that lie between order and chaos.

## Phase space and dynamic properties

Let's imagine the state space of a complex (brain) system. There are different regimes in it:

- **Order/High Separability:** The system is in a state where causal paths are clearly separated and hardwired. Dynamically, this corresponds to a **strongly stable attractor** - little freedom, hardly any surprise. Here would be Σ high (many distinguishable paths) and the complexity would be low. Example: deep anesthesia or dreamless deep sleep -- the brain is very neat, but unconscious.
- **Chaos/Null Structure:** The other extreme is complete **disorder**, where there is activity, but no constant causal structures or memory. The system bounces chaotically through phase space. This would be Σ nonsensical (because there is no structured path) and there would also be no stable Φinfo integration (close to 0, despite apparent disorder). Example: a brain in status epilepticus (chaotic discharge without content) or a failed AI system in noise - no meaningful information, no consciousness.
- **State of consciousness / causal collapse:** It is precisely **between** these extremes that the interesting area lies. The conscious state corresponds to a **metastable attractor**: the system is neither rigidly fixed (it fluctuates, which creates degrees of freedom) nor completely unbounded (there is a form of order to which it leans again and again).

This "sweet spot" has some important features:

- **Metastability:** The core of consciousness is stable enough not to decay immediately, but unstable enough to change flexibly. Dynamically, this means that there are attractors with flat potential troughs -- the system can move between similar states, but remains in the area.
- **High dimensionality:** Many degrees of freedom of the system contribute to the state at the same time (high participation of many nodes), but nevertheless the dynamics are **coordinated**. In other words, the attractor has a relatively high **fractal dimension** compared to simple rhythm generators, but still well below the maximum possible (it is not pure noise).
- **Self-preservation:** After perturbations, the system tends to **return to similar states**, i.e., if you disturb a little, it regains its composure and returns to the conscious attractor (resilience, as mentioned above).
- **Criticality:** The range is **between too much order and too much chaos**. In the complex system literature, one speaks of the "edge of chaos" - systems at this

edge show maximum computing power and complex pattern formation. Our collapse of consciousness lies right there: enough entropy for flexibility, enough order for integration.

This can also be represented graphically conceptually: On an axisOrder vs. chaos you have on the left highly ordered, but not very complex states (high, low ΣΦ), on the right chaotic states (no sustained structure, practically against 0 but also against 0), and in between the area of ΣΦ **consciousness**, where moderate (below the threshold, but not zero) and high is high, i.e. maximum. This conscious attractor is thus aΣΦΩ**special area in phase space**, which the system occupies under suitable conditions.

# Mathematical characterization

The dynamic properties mentioned above can be further formalized:

- **Lyapunov exponents:** A conscious attractor would show a **mixed spectrum** of Lyapunov exponents -- some slightly positive (sensitivity to initial conditions, so that the system is exploratory), some negative (contraction, so that it doesn't fly apart). Pure order would have only negative exponents, pure chaos many strongly positive ones; Consciousness lies in the middle.
- **Fractal (Kaplan-Yorke) dimension:** This dimension, which can be calculated with the Lyapunov values, would be medium-high **for a conscious core**. Higher than in simple oscillators (which are almost periodic), but lower than in fully developed chaos (where the dimension would be very high).
- **Entropy production rate:** Moderate. The system produces neither minimum entropy (like a rigid periodic table) nor maximum entropy ( like white noise). Moderate entropy production corresponds to the idea that the system constantly goes through new microstates (generates information), but also "forgets" much of it or remains ineffective (otherwise it would be chaos).
- **Information integration:** Formally measured, for example, via this Φ would be high. This means that the attractor requires the simultaneous consideration of many components (not divisible) for description. Height Φ is a characteristic of states of consciousness according to IIT, which is consistent with this.
- **Recurrence pattern:** For example, if you look at a recurrence plot (showing when the system returns to similar states), you would see **both patterns and irregularities**. Neither strictly periodic (where the recurrence pattern would be a lattice), nor completely uncorrelated, but complex patterns with variations - this is exactly what EEG recordings of the awake brain show.

Viewed through this lens, consciousness appears as a **general organizing principle** of complex systems. It is not tied to biology: as long as a system has the dynamic properties mentioned above - high integration, metastable attractors, critical dynamics - it could have analogous consciousness, whether it consists of neuronal mass, silicon chips or other components. Consciousness would thus be an **emergence at the system level**: a new property that emerges when certain parameter ranges are reached. However, we currently assume that consciousness is tied to biological systems.

# Applications, predictions and implications

Finally, it will be shown how the theory of gradual, causal-based consciousness can be applied and tested. To this end, we look at clinical phenomena (disorders of consciousness, psychiatric states) through collapse glasses, formulate experimental predictions to test the theory and discuss philosophical consequences.

## Clinical Applications: Disorders and Changes in Consciousness

### Disorders of Consciousness:

Various clinical conditions in which consciousness is impaired can be interpreted as disorders of causal collapse:

- **Coma:** In coma, we probably find a **global loss of the ability to collapse**. The brain remains in more orderly, reflexive dynamics (high, low integration). No global causal nucleus is formed - the necessaryΣfeedbacks are functionally switched off or severely disturbed. Our model would predict that **high causal separability** is measurable in coma-EEG or fMRI data (close to the "state of order").

- **Minimal state of consciousness (MCS):** Here, patients show signs of consciousness at times. The model would interpret this as **local, unstable collapse zones**: Occasionally, a small causal nucleus forms (e.g. during auditory stimulation, the temporoparietal flickering slightly), but it is not globally networked and is not permanently stable. Accordingly, in these patients fluctuates by: sometimes briefly below (conscious moment), $\Sigma(t)\Sigma_c$then over again.
- **Locked-in syndrome:** These patients are fully conscious but can barely move/communicate. In the model: The **causal collapse nucleus is intact** (global strong collapse as in awake consciousness, very low), but the Σ**output side is separated**. The causal networking remains internal, which explains their clear consciousness; they just can't express it.

### Psychiatric and altered states of consciousness:

- **Schizophrenia:** In this case, one might **suspect excessively intense but unstable states of collapse**. For example, during acute psychosis, patients sometimes have a feeling that everything is connected (many things are associated - perhaps an indicator very low), Σbut the states are not permanently stable and fragmented (changing delusions, changing hallucinations). The model would assume an **excessive, poorly regulated collapse here** .
- **Depression:** In severe depression, brain dynamics appear rigid and not very responsive. This could correspond to a **rigid, constricted collapse pattern**: perhaps the same subcore is constantly active (brooding in certain networks), which means that the system has one core, but a smaller one than normal and hardly any variation. Σ would possibly be moderately humiliated, but **not very dynamic**,

integration takes place, but always in the same ways - which is subjectively experienced as being trapped in always the same negative thoughts.

- **Dissociative states:** Dissociation (e.g., in trauma dissociation or identity disorder) could be understood in the model as **fragmentation of the causal nucleus**. Instead of a global core, several **partial collapses** form that **do not communicate with each other**. For example, the affected person has separate islands of consciousness (one part of the self feels pain, another does not register it; in DID, different personality states). Formally: Multiple $V_c$ $\Sigma_{V_c} < \Sigma_c$ subsets with within each, but these subsets hardly interact, the global may even remain relatively high. Σ

## Therapeutic approaches:

If one recognizes disorders of consciousness and mental disorders as problems of collapse regulation, new ideas for intervention emerge:

- **Neuronal stimulation:** Transcranial magnetic stimulation (TMS) or deep brain stimulation could be used to **modulate recursion loops in a targeted manner**. In depression, for example, one could try to loosen up the rigid network by stimulating at nodes in order to lower situationally and allow new integrations. Σ
- **Pharmacology:** Drugs could be used to influence **the depth and stability of recursion**. A hypothetical "collagen drug drug" would include, for example, network agents that increase connectivity (for people with a deficit of consciousness) or attenuate it (in the case of schizophrenic overintegration). In fact, anesthetics such as propofol presumably work by interrupting (increasing) global integration, while psychedelics such as LSD have been shown to increase cross-linking (lowering ΣΣ strongly, possibly too much).
- **Neurofeedback:** Patients could learn to regulate their own state of collapse with EEG/fMRI feedback, e.g. give depressed patients feedback on when their brain is particularly rigid Σ(high) and then consciously use counteracting relaxation or cognitive techniques that bring more flexibility.

The clinical view shows that the model can provide links for diagnosis (biomarkers?), therapy (stimulation, drugs), and a Σbetter understanding of psychiatric symptoms -- by attributing seemingly psychological phenomena to **measurable changes in causal network dynamics** .

# Experimental Predictions

To validate the theory, we propose some empirical tests:

- **Imaging studies (fMRI, PET):** The theory predicts that the **transition from the unconscious to the conscious state** (e.g. Waking up from anesthesia, or transition from REM sleep to wakefulness) is accompanied by an **abrupt drop in causal separability** . Practical: If you measure the effective connectivity during this time (for example by means of Granger causality or global synchronization measures), you should recognize a clear shift. Initial studies do indeed point to **critical thresholds** in such transitions.
- **Electrophysiology (EEG/MEG):** When you become aware, patterns should emerge that resemble **phase transitions**: e.g. critical fluctuations, sudden increase

in long-distance correlations, occurrence of certain frequency bands, etc. In particular, our theory would expect an **increase in metastable patterns** – i.e., oscillations that come and go, rather than rigid rhythms or chaotic noise.

- **Pharmacological manipulation:** As mentioned above:
  - **Psychedelics** (LSD, psilocybin) should measurably lead to **abnormally low levels**Σ , as they put the brain in a very integrative, unbounded state (which comes with ego-loss and synesthesia). Experiments with magnetic resonance have actually shown that functional networks become more mixed among psychedelics (matching collapse).
  - **Anesthetics** (propofol, ketamine depending on the stage) should have the opposite effect: **abnormally high levels**Σ, as they cut off integration and isolate networks. This is also consistent with findings that connectivity between brain areas decreases drastically under anesthesia.
- **Developmental psychology:** In infants and children, increasing integration should be observed in EEG/fMRI , which correlates with consciousness milestones. For example, one might suspect that around the age of 1-2, when object permanence and first self-recognition occur, networks suddenly close more recursive loops (decreasing . Longitudinal studies could test this - the theory would clearly predict that Σconsciousness development = **Continuous decline of (global)** ΣΦ with simultaneous increase of (up to adult values).
- **Comparative neurobiology:** Animal studies could quantify the differences described in the taxonomy. For example, highly developed species are expected **to have lower baseline**Σ and higher than less developed ones. For certain brain regions (e.g. forebrain vs. brainstem) you could also Φsee differences within an animal: conscious processes (neocortex) vs. unconscious reflexes (spinal cord) - the former has more complex, less separable dynamics.

These predictions offer a broad field for research. Initial indications from neuronal complexity studies, perturbation paradigms (e.g. PCI - Perturbational Complexity Index in coma diagnostics) and psychedelic research already indicate that core ideas (integration vs. differentiation) are correct. However, the special role of Σ as a key figure for causal separability is new and would have to be validated separately.

# Philosophical implications

The theory of graduality of consciousness outlined here also has important consequences for philosophical debates:

- **Panpsychism:** Some theories postulate that consciousness is a fundamental property of all matter (panpsychism). The model presented here **avoids panpsychistic assumptions** by formulating very specific **system requirements** for consciousness (high integration + causal collapse core). Not every electron has consciousness, but only systems that are complex enough and meet the right dynamic conditions. Consciousness is therefore **rare** in the universe, not ubiquitous background noise.
- **Emergence:** The model supports an **emergent** concept of consciousness. Consciousness is understood as **a system property** that arises from the interaction of many components (hence all the etc.), and which cannot be

Σ, Φ**reduced to the properties of individual components** (collapse = impossibility to separate influences). This fits with philosophical theories of emergence, which say that spirit/consciousness arises from matter, but has its own new properties.

- **Free will:** An old debate: How can free will exist when everything is causally determined? Our concept of **causal inseparability** offers an exciting approach here: In a collapsed core, the effect of individual causes can no longer be isolated; the system behaves as a **whole**, whereby it submits to its own **laws, so to speak** . It is not completely random (because it is an attractor with self-structure), but it is also not clearly predetermined from the outside (since the internal causal contributions are merged). In this **intermediate space**, something like a "conditional free space" for decisions of the will could be located philosophically - the system (e.g. the brain of a person) generates its own causal conclusions, which are not completely determined by external inputs, but without being magical or random.

- **Body-soul problem:** Our model positions itself between eliminativism and dualism. It **does not simply reduce consciousness** to individual neuronal processes (because it is precisely the *structure of the connections* and the overall dynamic pattern that is decisive - consciousness is not a single neuron fire), but it also **does not postulate a separately existing immaterial soul**. Rather, consciousness is understood as an **organizational principle** that is based on physical processes. In this way, it provides a possible **bridge**: The soul (consciousness) is *neither* identical with the parts *nor* separate from them, but emerges from their specific arrangement and dynamics.

In conclusion, the **theory of gradual, causal-based consciousness** provides a comprehensive framework. It integrates **mathematical formalization** (for structures and dynamics), **neurobiological and psychological data** (on the different forms and stages of development of consciousness), as well as **technical and philosophical perspectives** (artificial implementation, nature of consciousness). As such, it is not only conceptually interesting, but also **empirically testable** – each of its core elements (such as ) can $\Sigma(t)$be approximated with today's technology. At the same time, it remains open to interdisciplinary connections: from evolutionary biology to AI research to the philosophy of mind. In doing so, it lays a possible foundation for a future **unified understanding of consciousness** as a gradual phenomenon rooted in a network of causes and effects.

# Bibliography


Bengio, Yoshua. 2020. The Consciousness Prior. Preprint. arXiv.

Casali, Adenauer G., et al. 2013. A Theoretically Based Index ofConsciousness Independent of Sensory Processing and Behavior. Science Translational Medicine 5 (198).

Chalmers, David J. 1996. The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Cleeremans, Axel. 2011. The Radical Plasticity Thesis: How the Brain Learns to Be Conscious. Frontiers in Psychology 2: 86.

Dehaene, Stanislas, and Jean-Pierre Changeux. 2011. Experimental and Theoretical Approaches to Conscious Processing. Neuron 70 (2): 200--227.

Dehaene, Stanislas, Hakwan Lau, and Sid Kouider. 2017. What Is Consciousness, and Could Machines Have It? Science 358 (6362): 486--492.

Edelman, Gerald M. 2004. Wider than the Sky: The Phenomenal Gift of Consciousness. Yale University Press.

Feinberg, Todd E., and Jon M. Mallatt. 2016. The Ancient Origins of Consciousness: How the Brain Created Experience. WITH press.

Ginsburg, Simona, and Eva Jablonka. 2019. The Evolution of the SensitiveSoul: Learning and the Origins of Consciousness. WITH press.

Lamme, Victor A. F. 2006. Towards a True Neural Stance onConsciousness. Trends in Cognitive Sciences 10 (11): 494--501.

Massimini, Marcello, et al. 2005. Breakdown of Cortical Effective Connectivity during Sleep. Science 309 (5744): 2228--2232.

Oizumi, Masafumi, Larissa Albantakis, and Giulio Tononi. 2014. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3. 0. PLoS Computational Biology 10 (5).

Schröder, Tobias, and Christian Schilling. 2023. Recursive Self-Modeling Architectures for Synthetic Consciousness. Frontiers in Artificial Intelligence 6.

Shanahan, Murray. 2010. Embodiment and the Inner Life: Cognition andConsciousness in the Space of Possible Minds. Oxford University Press.

Tononi, Giulio. 2004. An Information Integration Theory of Consciousness. BMC Neuroscience 5 (1): 42.

Tononi, Giulio, and Christof Koch. 2015. Consciousness: Here, There and Everywhere? Philosophical Transactions of the Royal Society B 370 (1668).