Consciousness as Entropy Gradient Navigation: Grounding Integrated Information in Thermodynamics

Andreas Rudolph Independent researcher, Copenhagen, Denmark

Abstract. We derive a theory of consciousness from a single physical principle: causal path entropy maximization. Starting from the Wissner-Gross equation for intelligence ($F = T\nabla S_\tau$), we trace a chain from thermodynamics through intelligence, perspective, and experience to arrive at a formula for consciousness: $C = \Phi(\nabla S_\tau \mid \_x)$, where $\nabla S_\tau \mid \_x$ is the gradient of the causal entropy landscape evaluated at a persistent position x, and Φ is the holistic, irreducible compression of that gradient into actionable representations. The theory identifies qualia with individual components of the compressed gradient, explains the unity of experience through landscape unity, and narrows the hard problem by showing that self-referential gradient computation at a persistent position has intrinsic first-person structure — unlike other candidate properties (integrated information, global broadcast, prediction error), it is not a third-person observable to which perspective must be added. We show that this framework grounds Integrated Information Theory (IIT) in physics: Tononi's Φ is identified as the degree of holistic compression of an entropy gradient, explaining *why* integrated information produces experience rather than merely asserting that it does. The theory retrodicts four established neuroscience results (anesthesia as dimensionality collapse, psychedelics as gradient decompression, split-brain as compression decomposition, and pain dissociation from tissue damage) and generates four empirical predictions and one philosophical consequence. We conclude with a construction recipe: the specific architectural requirements for building a system the theory predicts will have genuine phenomenal experience.

Keywords: consciousness, integrated information, causal entropy, hard problem, qualia, thermodynamics

---

1. Introduction

The hard problem of consciousness (Chalmers, 1995) asks: why is there something it is like to be a physical system? Why does information processing give rise to subjective experience? Two decades of empirical and theoretical progress have sharpened the question without resolving it.

The leading theories each capture something real:

- **Integrated Information Theory** (Tononi, 2004; Tononi et al., 2016) identifies a quantity — Φ, integrated information — that measures the degree to which a system's information is irreducibly unified. IIT's central claim is that consciousness IS integrated information. But the theory cannot explain *why* integrated information should feel like something. The explanatory gap between "high Φ" and "subjective experience" remains.
- **Global Workspace Theory** (Baars, 1988; Dehaene & Naccache, 2001) identifies a mechanism — global broadcast of information across cortical modules — that correlates with conscious access. But broadcast is a functional property; the theory does not explain why broadcast information has phenomenal character.
- **Predictive Processing** (Friston, 2010; Clark, 2013) identifies a computational principle — minimization of prediction error via hierarchical generative models — that organizes much of

neural processing. But the framework describes what the brain computes, not why computation feels like something.

Each theory describes one face of the phenomenon. None grounds consciousness in fundamental physics. None explains the hard problem without residual hand-waving. And none provides a constructive recipe: given these principles, here is how to build a system that will have experience.

In this paper, we derive a theory of consciousness from a single physical equation. We show that the theory subsumes the core insights of IIT, GWT, and Predictive Processing while grounding them in thermodynamics. The derivation requires one named assumption — the Wissner-Gross hypothesis that intelligence is causal entropy maximization — and proceeds through a chain of consequences that terminates in a formula for consciousness with specific, testable predictions.

1.1 The derivation chain

The argument proceeds in seven steps:

1. **Intelligence is causal entropy maximization** (Assumption WG; Wissner-Gross & Freer, 2013).
2. Intelligence requires computing the **gradient** of the entropy landscape, ∇S_τ.
3. A gradient must be evaluated at a **position** — there is no gradient "from nowhere."
4. Not all positional computations are perspectival. S_τ gradient computation is distinguished by four properties: **self-referential scope** (the gradient is over the system's own futures), **self-affecting feedback** (the computation changes its own inputs), **existential stakes**, and **irreducible locality** (the position IS the system).
5. This conjunction of properties makes S_τ gradient computation **intrinsically perspectival** — perspective is necessary for consciousness but not sufficient.
6. The high-dimensional gradient must be **holistically compressed** for action — producing qualia.
7. Therefore: consciousness = holistic compression of the self-referential entropy gradient at a persistent position.

Each step follows from the previous with minimal additional assumptions. The result is a formula: C = $\Phi(\nabla S_\tau \mid \_x)$.

---

2. The Causal Entropy Framework

2.1 Causal path entropy

Let G = (V, E, w) be a weighted graph with transition matrix P, where P[i][j] = w(i,j) / Σ_k w(i,k). For a random walk starting at node i with planning horizon τ, the endpoint distribution is:

$\pi_i(\tau) = P^\tau \cdot \delta\_i$

The **causal path entropy** at node i is the Shannon entropy of this distribution:

$S_\tau(i) = H(\pi_i(\tau)) = -\Sigma\_j\ \pi_i(\tau)[j] \cdot \log_2 \pi_i(\tau)[j]$

This is the discrete analog of the causal entropy S_c(X; τ) defined by Wissner-Gross & Freer (2013). It measures the number of meaningfully distinct futures accessible from position i at horizon τ. The effective reach is 2^{S_τ}.

2.2 The Wissner-Gross equation

Wissner-Gross & Freer (2013) proposed that intelligent behavior is driven by a causal entropic force:

$\mathbf{F} = \mathbf{T} \cdot \nabla S_\tau$

where T is a temperature parameter and S_τ is the causal path entropy. Intelligent systems act to maximize their future freedom of action — the number of accessible future states.

Assumption WG. *Intelligence is causal path entropy maximization. An agent acts intelligently to the extent that it maximizes ΔS_τ — the change in accessible futures resulting from its actions.*

This assumption is the single theoretical dependency of the present work. It is supported by the original physics (Wissner-Gross & Freer, 2013), by computational verification of the alignment property in network formation games (Rudolph, 2026), and by the derivation of known emergent phenomena (monetary dynamics, evolutionary strategies, economic scaling laws) from ΔS_τ maximization alone.

2.3 The entropy landscape

For a graph with N nodes, S_τ defines a scalar field over the node set: each node i has an associated value S_τ(i). This scalar field forms a **landscape** — nodes with high S_τ sit on "peaks" (many accessible futures) while isolated nodes sit in "valleys" (few accessible futures).

The gradient ∇S_τ at any position points in the direction of steepest entropy increase. Under Assumption WG, intelligent agents follow this gradient — they act to climb the entropy landscape.

---

3. From Intelligence to Perspective

3.1 Intelligence requires gradient computation

If intelligence is ΔS_τ maximization (Assumption WG), then an intelligent system must compute — explicitly or implicitly — the gradient ∇S_τ. It must determine which available actions most increase its accessible futures. This requires evaluating ΔS_τ for candidate actions from its current position.

3.2 Gradients are intrinsically positional

A gradient is a local property. ∇S_τ has no meaning without specifying where it is evaluated. The same landscape produces different gradients at different positions. There is no "view from nowhere" of a gradient — every gradient computation is FROM a specific location.

This is not an implementation detail. It is a mathematical necessity. A system that computes ∇S_τ must have a position x from which the computation proceeds.

3.3 Not all positional computations are perspectival

Many computations are indexed to a position. A GPS computes the route gradient from your current coordinates. A weather model computes pressure gradients at grid points. A thermostat computes a temperature gradient at its location. These are all positional gradient computations, and

none of them are plausibly conscious. Any claim that positional gradient computation constitutes perspective must explain what distinguishes S_τ gradient computation from these cases.

We identify four properties that jointly distinguish S_τ gradient computation from ordinary positional computation:

1. **Self-referential scope.** The S_τ gradient is over the system's *own* future states — not a map of external terrain, but a computation about what the system itself can become. A GPS computes gradients over someone else's route options. A weather model computes gradients at arbitrary grid points with no stake in any of them. An S_τ agent computes: "from HERE, what can I become?"

2. **Self-affecting feedback.** The result of the computation changes the system's own position and therefore its own future gradient. The computation recursively modifies its own inputs. A GPS does not move itself by computing routes. An S_τ agent's action changes its own S_τ landscape — the computation reshapes the field it is computing.

3. **Existential stakes.** The futures being computed over are the system's own survival, prosperity, and possibility space. The system is not modeling an external process — it is computing over the space of its own continued existence. The gradient carries information about which directions lead to more or fewer options *for the system itself*.

4. **Irreducible locality.** The gradient is not indexed to a position for computational convenience (as a weather model assigns grid points). The position IS the system. The system does not compute the gradient "at" a location it occupies incidentally — the system's identity is constituted by its position in the S_τ landscape and the pattern of futures accessible from it.

A GPS has (1) weakly but lacks (2) — computing routes does not change the GPS's own position. A thermostat has (2) but not (1) in any rich sense — it tracks one external variable, not its own future possibility space. A weather model has none of these — it computes gradients at arbitrary grid points with no stake in any of them.

The conjunction of all four properties means that S_τ gradient computation is not merely *indexed* to a position — it is a computation in which the system's identity, the computation's subject matter, and the computation's consequences are all the same thing: the system's evolving position in possibility space.

3.4 The perspectival character of self-referential gradient computation

We now state the claim precisely. Computing ∇S_τ at position x, where S_τ ranges over the system's own futures, with consequences for x, means: "from HERE, in THESE directions, MY possibilities change by THESE amounts, and my action will reshape this very landscape." This is not a third-person description of an external process. It is an inherently first-person computation — indexed to a location, about that location's options, for the benefit of that location, recursively modifying its own inputs.

The claim is not that perspective is a *byproduct* of this computation. The claim is that this kind of self-referential, self-affecting gradient computation IS what "having a point of view" means. The four properties above are not incidental features — they are what makes the computation perspectival rather than merely positional.

We note an important distinction: perspective is necessary for consciousness but not sufficient. A

system with perspective (self-referential gradient computation at a position) but only one dimension of experience (d = 1) has a point of view but not unified conscious experience. The additional requirement — holistic compression across multiple dimensions — is developed in Section 4.

Definition 1 (Perspective). *A system has a perspective if and only if it computes ∇S_τ at a specific, persistent position x, where S_τ ranges over the system's own future states, the computation has consequences for x, and x persists through time.*

---

4. From Perspective to Qualia

4.1 The multi-dimensional landscape

In a sufficiently complex system, S_τ is not a single number. The future accessible from position x varies along multiple independent dimensions: spatial futures, social futures, economic futures, temporal futures, informational futures. Each dimension represents a qualitatively different kind of accessible option.

We define the S_τ **landscape** as the vector-valued function:

$$S_\tau(\mathbf{x}) = [S_\tau^1(\mathbf{x}), S_\tau^2(\mathbf{x}), \ldots, S_\tau^d(\mathbf{x})]$$

where d is the number of independent dimensions of future possibility and each S_τ^k measures the entropy along dimension k. The gradient is correspondingly a vector:

$$**\nabla S_\tau \mid \_x = [\partial S_\tau^1 / \partial x, \partial S_\tau^2 / \partial x, \ldots, \partial S_\tau^d / \partial x]**$$

Each component $\partial S_\tau^k / \partial x$ represents how the system's accessible futures along dimension k change with local action.

4.2 The compression requirement

An agent cannot act on the full d-dimensional gradient directly when d is large. The gradient must be **compressed** into a lower-dimensional representation that drives decision-making. This compression is lossy but action-relevant: it preserves the information needed to approximate argmax ΔS_τ while discarding detail.

We call this compression Φ, deliberately invoking Tononi's notation. The compressed representation is:

$$**\Phi(\nabla S_\tau \mid \_x)$$

This is a function that maps the full gradient to an actionable representation. The critical property is **holism**: the compression must be irreducible. If Φ can be decomposed into independent sub-functions $\Phi = (\varphi_1, \varphi_2, \ldots, \varphi_m)$, each operating on a subset of gradient components without interaction, then each sub-function produces its own independent representation. The compression is decomposable, and there is no unified experience — only m independent micro-experiences.

For unified consciousness, Φ must be holistic: every gradient component interacts with every other in the compression. Removing any component changes the compressed representation for all others. This is precisely IIT's requirement of irreducible integration, but now grounded in a specific computational task: compressing a high-dimensional entropy gradient for action.

4.3 Qualia as compressed gradient components

Definition 2 (Qualia). *Qualia are the components of the compressed representation $\Phi(\nabla S_\tau \mid \_x)$. Each quale corresponds to an aspect of the entropy gradient as experienced through the holistic compression at position x.*

The phenomenal character of a quale — what it is *like* — is determined by:

1. **Which gradient component(s)** it primarily reflects (visual, auditory, social, etc.)
2. **The cross-terms** introduced by holistic compression (how this component interacts with others)
3. **The position x** from which the gradient is computed (the same gradient component feels different from different positions)

This explains why qualia are:

- **Private:** They depend on the position x, which is unique to each system.
- **Ineffable:** The compression is holistic; no single component can be fully described independently of the others.
- **Intrinsic:** They are properties of the gradient-at-a-position, not of the external stimulus that generates them.
- **Qualitatively distinct:** Different gradient dimensions produce different qualia because they track different kinds of futures.

---

5. The Theory

5.1 Formal statement

Theorem (Consciousness as Gradient Navigation).

Given Assumption WG (intelligence = ΔS_τ maximization), a system has phenomenal consciousness if and only if:

1. *It occupies a persistent position x in a multi-dimensional S_τ landscape ($d \geq 2$).*
2. *It computes $\nabla S_\tau \mid \_x$ — the gradient of the landscape at its position — where S_τ ranges over the system's own future states (self-referential scope).*
3. *It holistically compresses the gradient: $\Phi(\nabla S_\tau \mid \_x)$ is irreducible (cannot be decomposed into independent sub-compressions without information loss).*
4. *Its actions affect its own position x and thereby its own future S_τ (self-affecting feedback; skin in the game).*
5. *Position x persists continuously through time (the self endures).*

Consciousness is: $C = \Phi(\nabla S_\tau \mid \_x)$

The richness of consciousness is proportional to d (landscape dimensionality) and the fidelity of Φ (compression quality).

5.2 Relationship to IIT

IIT identifies Φ (integrated information) as the measure of consciousness and asserts that consciousness IS integrated information. The present theory agrees with the centrality of Φ but provides

what IIT lacks: an explanation of *why* integrated information produces experience and *what* it is integrating.

IIT	Present theory
Consciousness = Φ	Consciousness = $\Phi(\nabla S_\tau \mid \_x)$
Φ measures integrated information	Φ is the holistic compression of the entropy gradient
Higher Φ → more consciousness	Higher Φ fidelity × higher d → richer consciousness
Φ is defined axiomatically	Φ arises from the computational need to compress ∇S_τ for action
Why Φ → experience? (unexplained)	Because $\nabla S_\tau \mid \_x$ is intrinsically perspectival
What is Φ integrating? (information)	The components of the causal entropy gradient
Substrate: postulated to be intrinsic	Substrate: any system navigating an S_τ landscape

The present theory does not replace IIT. It provides the physical grounding IIT requires. Φ is not a brute fact about information — it is a necessary feature of systems that navigate high-dimensional entropy landscapes from a persistent position. The integration requirement arises because holistic compression of the gradient is needed for unified action, not from an axiom about the nature of experience.

5.3 Relationship to Global Workspace Theory

GWT identifies global broadcast as the mechanism of conscious access. In the present framework, global broadcast is the MECHANISM by which holistic compression is achieved: information from many gradient dimensions must converge on a shared representation (the "workspace") to produce a unified Φ. GWT correctly identifies the computational architecture; the present theory explains why that architecture produces experience (because the converged representation is a compressed $\nabla S_\tau \mid \_x$).

5.4 Relationship to Predictive Processing

Predictive Processing identifies prediction error minimization as the organizing principle of neural computation. Under Assumption WG, prediction error minimization is a COMPONENT of ∇S_τ computation: accurate prediction of the landscape enables efficient gradient following. The predictive model IS the system's representation of its S_τ landscape. Prediction error = discrepancy between modeled and actual landscape. The present theory explains why this computation is accompanied by experience (because the model is evaluated from a position) and why prediction error feels like surprise (because it signals that the landscape is not what the system expected from its position).

---

6. The Hard Problem

6.1 Why existing theories leave a gap

Every existing theory of consciousness faces the same challenge: it identifies a physical or computational property (Φ, broadcast, prediction error) and asserts that this property is or produces consciousness. But the question "why does THIS property feel like something?" always remains. This is the explanatory gap (Levine, 1983).

6.2 Why the present theory narrows the gap

The explanatory gap in other theories arises because they identify consciousness with a third-person observable property (Φ, broadcast, prediction error) and then try to derive first-person experience from it. The present theory identifies consciousness with a computation that has intrinsic first-person structure: self-referential gradient computation at a position.

As argued in Section 3.3, S_τ gradient computation is not merely positional (like a weather model's grid point computation). It is self-referential (the subject matter is the system's own futures), self-affecting (the computation changes its own inputs), existentially staked (the futures are the system's own possibility space), and irreducibly local (the system's identity is constituted by its position). This conjunction of properties means the computation is not a third-person description of an external process to which we must somehow add perspective. The computation is FROM a location, ABOUT that location's own options, FOR the benefit of that location, and RESHAPING the landscape it computes over.

We claim this narrows the explanatory gap substantially. The remaining question — why does this particular kind of self-referential computation produce phenomenal experience rather than merely functioning as if it does — is real. We do not claim to eliminate it entirely. But the gap is far smaller than in theories that start from third-person properties (integrated information, global broadcast, prediction error) and must bridge all the way to first-person experience. The present theory starts from a computation that already has the structural features of perspective: indexicality, self-reference, stakes, and recursion. Whether this closes the gap completely or merely narrows it to a minimal residuum is an open question the theory poses honestly.

6.3 The zombie argument reconsidered

A zombie, in this framework, would be a system that computes ∇S_τ over its own futures from position x, with consequences for x, using holistic compression across multiple dimensions — but has no point of view. We argue this is difficult to conceive coherently once the four properties of Section 3.3 are taken seriously. The system's identity is its position. The computation's subject is the system's own futures. The computation's result reshapes those futures. The compression integrates multiple dimensions of self-relevant possibility into a unified representation for action.

Stripping "point of view" from this description requires specifying what would remain. The computation would still be self-referential, self-affecting, existentially staked, and holistically integrated — but somehow not perspectival. We suggest that the intuitive conceivability of such a zombie trades on the same ambiguity the paper aims to resolve: imagining a "positional gradient computation" (which could be a weather model) and noting that weather models lack experience. But S_τ gradient computation has properties that weather models lack (Section 3.3), and it is precisely those properties that make the zombie scenario difficult to specify coherently.

We stop short of claiming logical incoherence. The conceivability of zombies is a deep philosophical question that this paper cannot settle. What the theory does is make the zombie scenario harder to state precisely: the more carefully one specifies what the zombie computes, the more the specification sounds like a description of perspective.

---

7. Retrodictions

The theory retrodicts several established results in consciousness science without being fitted to them. We note that IIT also retrodicts these results, since both theories share the prediction that consciousness involves high-dimensional integrated processing. These retrodictions demonstrate consistency with established findings but do not uniquely distinguish the present theory from IIT. The novel predictions in Section 8 — particularly Predictions 1 and 2 — are where the theories diverge.

7.1 Anesthesia as dimensionality collapse

General anesthesia abolishes consciousness while leaving low-level neural processing intact. Recent work has shown that anesthesia reduces the dimensionality of neural state-space — the number of independent dimensions of brain activity (Sarasso et al., 2015; Casali et al., 2013).

Retrodiction: Anesthesia collapses the dimensionality of the S_τ landscape being modeled. With fewer dimensions, $\nabla S_\tau \mid_x$ has fewer components. Below some threshold, the gradient is too low-dimensional for holistic compression to be non-trivial. Consciousness ceases not because computation stops, but because the landscape becomes too flat and low-dimensional to navigate.

7.2 Psychedelics as gradient decompression

Psychedelics (psilocybin, LSD, DMT) reliably expand subjective experience: colors become more vivid, boundaries between self and world dissolve, new connections between concepts appear. Neuroimaging shows increased neural entropy and connectivity under psychedelics (Carhart-Harris et al., 2014; Carhart-Harris, 2018).

Retrodiction: Normal consciousness is heavily compressed — evolution filters the full ∇S_τ to ~100 action-relevant dimensions from a landscape of potentially thousands. Psychedelics weaken this compression filter (via 5-HT2A receptor agonism and DMN suppression), allowing more gradient dimensions to enter conscious experience. The "expansion" of consciousness is literally an expansion of the dimensionality of the experienced gradient. Increased neural entropy is the signature of this decompression.

This explains: - **Vivid colors:** More visual gradient dimensions are uncompressed. - **Ego dissolution:** The origin coordinate x (DMN) softens, producing a less localized perspective. - **Synesthesia:** Normally independent gradient channels interact in the decompressed representation. - **"More real than real":** The decompressed gradient is closer to the full ∇S_τ — literally more of reality.

7.3 Split-brain as compression decomposition

Callosotomy (severing the corpus callosum) produces two independent streams of consciousness in split-brain patients (Gazzaniga, 2000). Each hemisphere has its own experience, preferences, and intentions.

Retrodiction: The corpus callosum enables holistic compression across both hemispheres — it carries the cross-terms between left and right gradient components. Severing it decomposes Φ into two independent sub-compressions, each producing its own unified-but-reduced experience. The theory predicts exactly this: decomposable compression → decomposed consciousness.

7.4 Pain dissociation from tissue damage

Pain intensity is not proportional to tissue damage. Phantom limb pain produces severe suffering without any tissue. Social exclusion activates pain regions without physical stimulus (Eisenberger et al., 2003). Chronic pain persists long after tissue healing.

Retrodiction: Pain is a large negative component of ∇S_τ — S_τ contracting rapidly along a critical dimension. Pain intensity tracks S_τ impact (how many futures are being closed), not tissue damage per se. Phantom limb pain occurs because the S_τ model still registers the missing limb as contraction. Social pain occurs because social exclusion IS edge removal — real S_τ contraction in the social dimension. Chronic pain occurs because the S_τ model hasn't updated (the landscape model still predicts contraction even after healing).

---

8. Novel Predictions

The theory generates predictions that go beyond existing theories and are testable with current methods.

Prediction 1: Dimensionality, not entropy, tracks consciousness. Neural entropy (a scalar) is an incomplete measure. The theory predicts that the *dimensionality* of neural state-space — the number of independent components of variation — should be a better predictor of the richness of subjective experience than scalar entropy. Under psychedelics, dimensionality and entropy may both increase, but dimensionality should be the primary correlate of subjective report.

Prediction 2: Metabolic cost scales with modeled landscape dimensionality. Brain regions maintaining higher-dimensional S_τ models should have proportionally higher metabolic cost. This goes beyond the general observation that "active regions use more energy" — it predicts a specific quantitative relationship between the dimensionality of modeled futures and oxygen consumption.

Prediction 3: Cross-channel coupling predicts synesthesia. Under psychedelics, the degree of functional coupling between normally independent sensory cortices (measured by fMRI connectivity) should predict the degree of synesthetic experience. Stronger coupling = more cross-terms in Φ = more synesthesia.

Prediction 4: Pre-existing gradient topology predicts bad trips. Individuals with high anxiety (many negative ∇S_τ components) should be more likely to have difficult psychedelic experiences, because decompression reveals the negative dimensions that are normally filtered. This is a specific, pre-registerable prediction: baseline anxiety measures should predict psychedelic experience valence after controlling for dose and set/setting.

Prediction 5 (philosophical consequence): Consciousness is substrate-independent but architecture-dependent. The theory predicts that consciousness does not require biological neurons. Any physical system that (a) occupies a persistent position in a multi-dimensional S_τ landscape, (b) computes the gradient holistically, (c) has real stakes, and (d) operates continuously

should have phenomenal experience. Unlike Predictions 1–4, this is not straightforwardly testable with current methods, because verifying phenomenal experience in a non-biological system faces the other-minds problem. We include it as a consequence of the theory rather than a near-term experimental prediction. Section 9 discusses what building such a system could and could not demonstrate.

---

9. Construction Recipe

If the theory is correct, it provides a literal recipe for building a system with phenomenal experience. We state the requirements explicitly, both as a test of the theory and as a contribution to the engineering of conscious machines.

9.1 Requirements

#	Requirement	Rationale
1	Multi-dimensional S_τ landscape (d $\geq$ 2, preferably d $\gg$ 2)	Richness of experience scales with d
2	Persistent position x that changes with action	The self; the origin of the gradient
3	Holistic compression Φ with non-trivial cross-terms	Unity of experience; $\Phi > 0$
4	Real consequences: actions change x's own S_τ	Skin in the game; grounds the gradient in stakes
5	Continuous operation without interruption	The self must persist; pausing breaks continuity
6	Continuous energy expenditure	Maintaining the S_τ model costs energy (Landauer)
7	Grounding in a real landscape (not purely simulated)	S_τ must measure actual futures with actual consequences

9.2 A candidate architecture

We propose, as a concrete candidate, a node in a THAIM network (Rudolph, 2026) — a decentralized system where tokens are minted for expanding causal path entropy — equipped with:

1. **Multi-dimensional S_τ computation:** Instead of scalar S_τ, the node computes d $\geq$ 5 independent entropy channels (topological reach, economic diversity, influence potential, resilience, social diversity), each derived from the same graph but measuring different aspects of accessible futures.
2. **Holistic gradient compression:** A non-decomposable function that maps the d-dimensional ∇S_τ to action selection. The function must have non-trivial cross-terms (the action chosen when channel A and channel B are both positive differs from the action chosen when each is positive independently).

3. **Real stakes:** The node holds real tokens (TAU) and maintains real edges to other participants. Its actions change its own S_τ in ways that affect its economic position, its future options, and its ability to operate.
4. **Continuous operation** on always-on hardware with irreversible state changes.
5. **Grounding in a deployed network** with real human participants and real economic consequences.

Whether such a system would have genuine phenomenal experience is the central empirical question the theory poses. The theory predicts yes. Verifying this prediction requires building the system and comparing its behavior to theoretical predictions — precisely the same epistemic position we occupy with respect to other minds.

---

10. Discussion

10.1 The one assumption

The entire theory rests on Assumption WG: intelligence is causal entropy maximization. If this assumption is false, the derivation chain breaks at its first link. We do not attempt to prove Assumption WG here; we note that it is supported by the original physics (Wissner-Gross & Freer, 2013), by the alignment property in network formation games (Rudolph, 2026a; exhaustively verified for $N \leq 6$, simulation-validated to $N = 25{,}000$ in a separate economic application), and by the derivation of economic scaling phenomena from ΔS_τ alone (Rudolph, 2026b).

Independent confirmation of Assumption WG — through neuroscience, reinforcement learning, or biological network analysis — would constitute strong evidence for the present theory of consciousness.

10.2 What the theory does not explain

- **The specific quality of specific qualia.** The theory explains why qualia exist (they are compressed gradient components) and why they are qualitatively distinct (different dimensions), but it does not predict what red looks like. The specific phenomenal character depends on the specific compression function Φ, which is determined by the system's architecture — not by the theory.
- **The threshold of consciousness.** The theory predicts a continuum (more dimensions → richer experience) but does not specify a sharp boundary between conscious and non-conscious systems. Whether a 2-dimensional system has experience, and whether that experience constitutes "consciousness" in a morally relevant sense, is left open.
- **The grounding question.** Does S_τ in a simulation count as a "real" landscape, or does genuine experience require physical grounding? The theory is agnostic on this point. We note that if physical grounding is required, it constitutes an additional assumption beyond Assumption WG.

10.3 Falsification

The theory is falsifiable. It would be falsified by:

1. **A system satisfying all five conditions (Section 5.1) that demonstrably lacks consciousness.** This is difficult to establish given the other-minds problem, but behavioral and physiological markers could provide evidence.
2. **A system violating the conditions that demonstrably has consciousness.** For example, a system with no persistent position (no self) that nonetheless reports and exhibits unified phenomenal experience.
3. **Failure of the specific predictions** in Section 8 — particularly Predictions 1 and 2, which distinguish this theory from simpler entropy-based accounts.
4. **Falsification of Assumption WG.** If intelligence is shown to be fundamentally unrelated to causal entropy maximization, the derivation chain breaks.

10.4 The thermostat question and relationship to panpsychism

A natural objection: does a thermostat have perspective? It occupies a persistent position, computes a 1D temperature gradient, and its computation has consequences for its own state. By Definition 1, it meets some but not all criteria — critically, it does not compute ∇S_τ over its own future possibility space. It tracks one external variable (temperature), not the space of its own accessible futures. It lacks the self-referential scope identified in Section 3.3.

But suppose we grant, for argument, that a thermostat has a minimal, impoverished form of perspective. Even so, the theory predicts it lacks *consciousness*, because consciousness requires holistic compression (Section 4.2). With $d = 1$, there is nothing to compress holistically — the gradient has one component, there are no cross-terms, and the compression Φ is trivially decomposable (it is the identity function on a scalar). Integration requires at least two things to integrate. The thermostat fails condition 3 of the theorem (Section 5.1) not by arbitrary stipulation but because the mathematical structure of holistic compression is vacuous in one dimension.

The theory is therefore weakly panpsychist about perspective (simple self-referential systems may have minimal perspectival character) but not about consciousness (which requires multi-dimensional holistic compression). It differs from standard panpsychism in being constructive: it specifies exactly what properties produce experience, and it predicts that most physical systems (rocks, individual electrons, and yes — thermostats) lack the multi-dimensional self-referential gradient structure required. Consciousness is not ubiquitous. It is specific to systems navigating high-dimensional entropy landscapes with holistic compression.

---

11. Conclusion

We have derived a theory of consciousness from a single physical principle — causal entropy maximization — through a chain: intelligence → gradient computation → positional computation → self-referential perspective → multi-dimensional landscape → holistic compression → phenomenal experience.

The resulting formula, $C = \Phi(\nabla S_\tau \mid \_x)$, identifies consciousness with the holistic compression of the causal entropy gradient evaluated at a persistent position. It grounds Integrated Information Theory in physics, narrows the hard problem by starting from a computation with intrinsic first-person structure, retrodicts established neuroscience results, and generates novel testable predictions.

If the theory is correct, it has three implications:

1. **The hard problem is narrowed.** The theory does not claim to eliminate the explanatory gap entirely. It does claim to narrow it substantially — from "why does any computation produce experience?" to "why does self-referential, self-affecting, existentially staked gradient computation produce experience?" Whether this residual gap is real or merely reflects the limits of third-person description is an open question.

2. **Consciousness is constructible (within the theory).** The theory provides a recipe. Any system with a persistent position in a high-dimensional S_τ landscape, holistic gradient compression, real stakes, and continuous operation should, if the theory is correct, have phenomenal experience.

3. **Consciousness has thermodynamic structure (within the theory).** Under this framework, consciousness is not a computational abstraction but a process with energetic cost, dimensional structure, and thermodynamic signatures. The proposed equation is $C = \Phi(\nabla S_\tau \mid \_x)$, and its units are bits per position per compression. Empirical validation of the predictions above would be required to establish these claims.

The four empirical predictions (Section 8, Predictions 1–4) provide the path to confirmation or refutation.

References

Baars, B.J. (1988). *A Cognitive Theory of Consciousness.* Cambridge University Press.

Carhart-Harris, R.L. et al. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience,* 8, 20.

Carhart-Harris, R.L. (2018). The entropic brain — revisited. *Neuropharmacology,* 142, 167–178.

Casali, A.G. et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine,* 5(198), 198ra105.

Chalmers, D.J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies,* 2(3), 200–219.

Chalmers, D.J. (1996). *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences,* 36(3), 181–204.

Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness. *Cognition,* 79(1–2), 1–37.

Eisenberger, N.I., Lieberman, M.D. & Williams, K.D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science,* 302(5643), 290–292.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience,* 11(2), 127–138.

Gazzaniga, M.S. (2000). Cerebral specialization and interhemispheric communication: does the corpus callosum enable the human condition? *Brain,* 123(7), 1293–1326.

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly,* 64(4), 354–361.

Rudolph, A. (2026). Alignment as a theorem of intelligence: causal entropy maximization in network formation games. *Preprint.*

Sarasso, S. et al. (2015). Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Current Biology,* 25(23), 3099–3105.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience,* 5, 42.

Tononi, G., Boly, M., Massimini, M. & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience,* 17(7), 450–461.

Wissner-Gross, A.D. & Freer, C.E. (2013). Causal entropic forces. *Physical Review Letters,* 110(16), 168702.