Dissolving the Boltzmann Brain Circularity: Time, Memory, and the Thermodynamic Laws from a Distinguishability Axiom

David Neale

Goleudy.ai, Rochester, New York, USA

david.n@goleudy.ai | ORCID: 0009-0006-7749-9876

Abstract

Wolpert, Rovelli, and Scharnhorst recently formalised the dynamics of the universe's entropy as a time-symmetric Markov process and showed that the Boltzmann brain hypothesis, the past hypothesis, and the second law of thermodynamics all share a common formal structure: each conditions the entropy trajectory on a single event at a single moment in time, differing only in the choice of conditioning event. Their analysis exposes circular dependencies among the second law, the reliability of memory, and the laws inferred from observational data. We propose that this circularity can be dissolved rather than merely formalised. Working from a single axiom in modal logic — $\Diamond N \rightarrow \neg N$ (nothingness cannot exist) — we derive a five-dimensional constraint space with a potential function $\Phi = \ln(\Omega/K)$ whose gradient flow satisfies $d\Phi/d\lambda = |\nabla\Phi|^2 \geqslant 0$. This geometric identity is the second law of thermodynamics, but with a fundamentally different logical status: it requires no temporal conditioning, no past hypothesis, and no assumption about record reliability. Temporal ordering itself emerges as a derived geometric property at $N \geqslant 3$ distinguishable features, where coupling matrices cannot be simultaneously diagonalised and the gradient field acquires irreducible circulation. The "conditioning problem" dissolves because time is not an independent parameter available for arbitrary conditioning. We further argue that memory — central to the Boltzmann brain debate — should be reframed from a physical trace carrying information about the past to a performance characteristic of an observer's predictive model. Drawing on computational results across four substrates (cellular automata, Game of Life, reaction-diffusion, and sorting algorithms), we show that systems with no memory mechanism produce emission profiles that observers label "habituation" and "learning" identically to biological systems. The question "are my memories reliable?" transforms into "what would an observer construct from this emission stream?" — a question that does not inherit the circularity. All four thermodynamic laws are derived as geometric theorems from the axiom: the zeroth from smoothness, the first from symplectic structure, the second from the gradient identity, and the third — the unattainability of absolute zero — from the axiom itself, since configurations cannot approach indistinguishability from nothingness.

Keywords: Boltzmann brain, second law of thermodynamics, observer inference, memory, distinguishability axiom, constraint geometry, thermodynamic laws

---

1. Introduction

Are your memories real? A recent paper by Wolpert, Rovelli, and Scharnhorst [1] brings renewed rigour to a question that has circulated through statistical physics and cosmology for over a century. The Boltzmann brain hypothesis proposes that your current perceptions, memories, and observational data are a statistical fluctuation out of the thermal equilibrium of the universe — a fleeting configuration that happens to match what a real observer with a real history would experience, but that bears no actual correlation to the past state of the universe. The hypothesis is counterintuitive, but the standard arguments against it turn out to be entangled with the very assumptions they seek to defend.

Wolpert et al. disentangle these arguments by formalising the dynamics of the universe's entropy as a time-symmetric, time-translation invariant Markov process — what they call the *entropy conjecture*. Their central observation is that the entropy conjecture, like any stochastic process, does not specify which times it should be conditioned on. Any choice of conditioning — a value of entropy at a particular time — must be introduced as an independent assumption. This observation reveals that the Boltzmann brain hypothesis, the past hypothesis, and the second law of thermodynamics all share the same formal structure: each conditions the entropy trajectory on a single event at a single moment, differing only in the details of their assumptions. In this precise sense, the authors conclude, the Boltzmann brain hypothesis and the second law are "equally legitimate (or not)" [1].

The paper also exposes a circularity at the heart of common arguments against the BB hypothesis. We infer the dynamical laws of the universe from our data records about the past. But those data records — our memories, our experimental results, our scientific archives — are precisely what the BB hypothesis calls into question. If our records are unreliable, the laws inferred from them are unreliable, and any argument that uses those laws to dismiss the BB hypothesis is circular. Wolpert et al. do not claim to resolve this circularity; they formalise it, showing that it arises structurally from the way the entropy conjecture interacts with Bayesian inference about the reliability of records.

This paper proposes that the circularity can be dissolved rather than merely formalised. We argue that two foundational assumptions — that time is a pre-given parameter available for conditioning, and that memory is a physical trace carrying information about the past — are the structural sources of the circularity, and that both can be replaced by derived quantities within a framework built from a single axiom.

The axiom is $\Diamond N \rightarrow \neg N$: if nothingness is possible, then nothingness does not obtain [2, 3]. From this axiom, we derive a five-dimensional constraint space in which configurations are organised by a potential function $\Phi = \ln(\Omega/K)$, where Ω counts accessible relational states and K measures descriptive complexity. Within this framework:

1. **The second law is a geometric theorem, not a conditioning choice.** The rate of change of Φ along gradient flow lines satisfies $d\Phi/d\lambda = |\nabla\Phi|^2 \geqslant 0$ — a mathematical identity requiring no temporal conditioning, no past hypothesis, and no assumption about record reliability. If

the second law is not a conditioning choice, it is not in the same logical category as the BB hypothesis, contra the Wolpert et al. conclusion of equal legitimacy (Section 2).

2. **Time is derived, not assumed.** The ordering parameter τ gains non-trivial structure only at $N \geqslant 3$ distinguishable features, where coupling matrices cannot be simultaneously diagonalised and the gradient field acquires irreducible circulation. The "conditioning problem" — at what time should we condition the entropy trajectory? — dissolves because temporal structure is a consequence of relational geometry, not an independent parameter available for arbitrary conditioning (Section 2).
3. **Memory is observer inference, not physical trace.** Drawing on computational results across four substrates [4], we show that memory, learning, and decision-making are performance characteristics of an observer's predictive model rather than properties of the observed system. A reaction-diffusion PDE with no memory mechanism produces emission profiles indistinguishable from published habituation data in single-celled organisms. The same system appears to habituate, sensitise, or show no learning depending solely on the observer's probing frequency. The question "are my memories reliable?" is replaced by "what would an observer construct from this emission stream?" — a question that does not require establishing the physical basis of traces or the thermodynamic conditions for their veridicality (Section 3).
4. **All four thermodynamic laws derive from the axiom.** The zeroth law follows from the smoothness of Φ; the first from symplectic structure and τ-translation symmetry; the second from the geometric identity above; the third — the unattainability of absolute zero — is the axiom itself in thermodynamic language, since the inner boundary of the viable region (indistinguishability from nothingness) cannot be reached. The four laws are geometric theorems rather than independent postulates that can be played against each other (Section 4).

We emphasise what this paper does not claim. We do not claim that the Wolpert-Rovelli-Scharnhorst analysis is incorrect. Their formalisation of the entropy conjecture is rigorous, and their identification of the shared structure across the BB hypothesis, the past hypothesis, and the second law is an important contribution. We claim that the circularity they identify is a symptom of treating time and the thermodynamic laws as fundamental, independent ingredients. When both are derived from a single axiom about distinguishability, the circular dependencies disappear — not because they are resolved within their own terms, but because the terms themselves are shown to be consequences of a deeper structure.

The epistemological position should be stated at the outset. Following [3], we are exploring mathematical structures that emerge from the distinguishability axiom and noting structural parallels to thermodynamic and physical laws. The derivation of $d\Phi/d\lambda \geqslant 0$ is a mathematical result about gradient flow in a bounded space. Its identification with the second law is a proposed correspondence — well-motivated by the structural parallels but requiring the bridge assumption that Φ corresponds to a thermodynamic potential and λ to a physical ordering

parameter. The dissolution of the Boltzmann brain circularity holds to the extent that this correspondence is accepted. The observer-inference results in Section 3 are computational and stand independently of the theoretical framework; they require only the standard Hidden Markov Model assumption that an observer builds predictions from observable emissions without direct access to hidden dynamics.

The paper is organised as follows. Section 2 introduces the constraint framework and shows how the second law and temporal ordering emerge as geometric theorems, dissolving the conditioning circularity. Section 3 reframes memory as observer inference and applies the reframing to the BB memory problem. Section 4 summarises the derivation of all four thermodynamic laws from the axiom. Section 5 discusses the relationship between the present approach and the Wolpert-Rovelli-Scharnhorst framework, and identifies open questions.

---

2. Dissolution via the Distinguishability Axiom

2.1 The Source of the Circularity

The circularity that Wolpert, Rovelli, and Scharnhorst expose has a precise structural origin: the entropy conjecture treats time as a pre-given parameter and entropy as a stochastic variable indexed by that parameter. Any inference about the entropy trajectory then requires conditioning on a value of entropy at a specific time. But the choice of conditioning event cannot come from the entropy conjecture itself — it must be imported from outside, typically from observational data whose reliability is itself in question.

This is not a failure of Wolpert et al.'s analysis. It is an accurate diagnosis. But it suggests that the problem lies not in the reasoning about conditioning but in the assumption that generates the need for conditioning in the first place: the treatment of time as fundamental.

We propose that the circularity dissolves — not by finding the correct conditioning event, but by working within a framework where time is derived rather than assumed, and where the second law is a geometric theorem rather than a probabilistic conditioning choice.

2.2 The Axiom

The framework begins from a single modal-logical axiom [2]:

$$\Diamond N \rightarrow \neg N$$

If nothingness is possible, then nothingness does not obtain. The argument is that absolute nothingness is self-undermining: considering it possible requires a conceptual framework, but any framework is something rather than nothing. This is not an empirical claim but a logical one — the impossibility of nothingness is analytic.

From this axiom, existence requires *distinguishability* as its minimum structure. A bare "something" with nothing to distinguish it from collapses into the nothingness that cannot exist.

Distinguishability is therefore fundamental, and distinguishability is inherently relational: to be distinguishable is to be distinguishable *from* something.

2.3 Constraint Space

Categorical exhaustion of what is required for robust distinguishability yields five independent constraint dimensions [2, 3]: boundary (β), pattern (κ), resource (ρ), integration (λ), and ordering (τ). Each addresses an irreducible requirement for maintaining distinction:

- **Boundary (β):** Where does this configuration end and another begin?
- **Pattern (κ):** What regularities characterise this configuration?
- **Resource (ρ):** What sustains this configuration's activity?
- **Integration (λ):** How do parts of this configuration cohere?
- **Ordering (τ):** How are states of this configuration sequenced?

These five constraints define a configuration space in which viable configurations — those that remain distinguishable from nothingness — occupy a bounded region V. The axiom imposes a lower boundary ∂V_- (configurations cannot approach indistinguishability from nothing) and an upper boundary ∂V_+ (configurations cannot approach self-contradictory totality). Within V, a scalar potential $\Phi = \ln(\Omega/K)$ characterises each configuration, where Ω counts accessible relational states and K measures descriptive complexity [5].

2.4 The Second Law as Geometric Identity

The critical result for the present argument concerns the gradient flow of Φ. Configurations change along a path parameterised by λ (a geometric ordering parameter, not a temporal coordinate). The gradient flow equation is:

$$\frac{dC}{d\lambda} = \nabla\Phi$$

The rate of change of Φ along this flow is:

$$\frac{d\Phi}{d\lambda} = \nabla\Phi \cdot \frac{dC}{d\lambda} = \nabla\Phi \cdot \nabla\Phi = |\nabla\Phi|^2 \geq 0$$

Equality holds only at critical points where $\nabla\Phi = 0$.

This is the second law of thermodynamics, but with a fundamentally different logical status from the entropy conjecture of Wolpert et al. Their formulation treats entropy increase as a probabilistic property of a time-indexed stochastic process — a property that holds "in expectation" given conditioning on a low-entropy event. Here, $d\Phi/d\lambda \geqslant 0$ is a geometric identity: the inner product of any vector with itself is non-negative. It requires no conditioning on initial

data, no past hypothesis, and no assumption about the reliability of records. It is a theorem about the structure of gradient flow in a space whose existence follows from the axiom.

The distinction matters. Wolpert et al. conclude that the second law and the Boltzmann brain hypothesis are "equally legitimate (or not)" because both condition the entropy trajectory on a single event, differing only in which event they choose [1]. If the second law is a conditioning choice, this conclusion follows. But if it is a geometric theorem — as it is in the present framework — it is not in the same logical category as the Boltzmann brain hypothesis. The BB hypothesis remains a conditioning choice (condition on present data and infer a fluctuation). The second law is not a choice at all; it is a consequence of what it means for configurations to follow gradients in a space defined by distinguishability.

2.5 The Emergence of Time

The second element of the dissolution concerns time itself. The entropy conjecture is a stochastic process indexed by time, and the circularity arises because inferences about entropy trajectories require temporal conditioning. In the constraint framework, the parameter λ that orders configurations along gradient flow lines is *not* time. Time — the experienced ordering of events by an observer — emerges as a derived quantity, and only under specific structural conditions.

The argument proceeds through a result in linear algebra [2, 6]. At the minimum configuration ($N = 2$ distinguishable features), the two coupling matrices can be simultaneously diagonalised. The structure is decomposable: it can be analysed into independent modes, and no preferred direction exists. A-to-B is indistinguishable from B-to-A. The ordering constraint τ is necessarily zero.

At $N \geqslant 3$, three or more coupling matrices $M(A,B)$, $M(B,C)$, $M(C,A)$ cannot generically be simultaneously diagonalised. The gradient $\nabla\Phi$ acquires a circulation component — a curl-like structure that distinguishes the two orientations around the loop $A \to B \to C \to A$. This geometric chirality is non-zero whenever the coupling matrices fail to commute, which is the generic case.

This chirality *is* what we experience as temporal ordering. The "direction of time" aligns with the circulation of $\nabla\Phi$ — the direction in which the gradient field has a consistent handedness around irreducible loops. No external time parameter is required. The ordering emerges from the geometry of constraint space at $N \geqslant 3$ and is absent at $N = 2$.

The consequences for the Boltzmann brain problem are direct. Wolpert et al. ask: at what time should we condition the entropy trajectory? The framework replies: the question presupposes that time is available as an independent parameter for conditioning. It is not. Time is a feature of configurations with sufficient relational complexity ($N \geqslant 3$). The "conditioning problem" — which time, which entropy value — dissolves because the temporal structure that would make the question meaningful is itself a derived geometric property, not an input.

2.6 The Third Law as the Axiom

The dissolution extends to the full set of thermodynamic laws, all of which are derived as geometric theorems within the framework rather than independent postulates [5]:

- **Zeroth Law:** Equilibrium is transitive because Φ is smooth throughout the interior of V.
- **First Law:** Energy is conserved because the symplectic structure of constraint space generates τ-translation symmetry (Noether's theorem).
- **Second Law:** Entropy does not decrease along gradient flow (the geometric identity above).
- **Third Law:** Absolute zero is unattainable because the inner boundary ∂V_{-} is unreachable — this is the axiom $\Diamond N \to \neg N$ expressed in thermodynamic language.

The third law deserves emphasis. The standard formulation ("the entropy of a system approaches a minimum as temperature approaches absolute zero") is, within this framework, a direct expression of the founding axiom. Configurations cannot reach ∂V_{-} because at ∂V_{-} they would be indistinguishable from nothingness, and nothingness cannot exist. The unattainability of absolute zero is the unattainability of non-existence.

This completes the logical chain: $\Diamond N \to \neg N \to$ distinguishability $\to$ constraint space $\to$ bounded viable region $\to$ gradient flow $\to$ four thermodynamic laws. No independent thermodynamic postulates are required, and no temporal conditioning is involved at any stage.

---

3. Memory as Observer Inference

3.1 Physical Memory in the Wolpert-Rovelli Framework

A central element of the Wolpert-Rovelli-Scharnhorst argument concerns the physical basis of memory. Their analysis draws on Rovelli's earlier formalisation of physical memory [7, 8]: a memory system is defined as a physical system (M_0, t_0) that carries information about a different system's state at an earlier time. This builds on Reichenbach's concept of branch systems — subsystems that branch off from a low-entropy parent system and retain traces of the state at branching [9]. The reliability of these traces is what makes records, perceptions, and scientific data meaningful.

The Boltzmann brain problem, as Wolpert et al. frame it, turns on whether this reliability can be established without circularity. To have reliable memories requires dissipation toward the future — increasing entropy in the time before the present — but the BB hypothesis assumes entropy *decreases* in that time to produce the large fluctuation. The reliability of memory thus depends on the very thermodynamic asymmetry that memory is supposed to help us establish. As they put it: we infer the dynamical laws from our data records about the past, but those laws then call into question the reliability of the records from which they were inferred [1].

We propose that the circularity here has a deeper source than the entanglement of laws and records. It arises from treating memory as a physical property — a system carrying information about the past — rather than as a performance characteristic of an observer's predictive model. Once memory is reframed as an observer inference, the question of whether memories are "reliable" transforms into a question about what emission profiles the observer constructs

predictions from, and the circularity dissolves along a different axis than the one Section 2 addressed for the second law.

3.2 Memory Reframed: Model Obsolescence

In the observer-inference framework developed in [4], memory is not something a system possesses. It is something an observer infers when the observer's predictive model goes stale.

The observer watches a system's emissions — the measurable features it can detect — without access to the system's internal dynamics. This is the structure of a Hidden Markov Model: hidden states generate observable emissions, and the observer builds a rolling predictive model from the emission stream alone. Memory is then defined operationally:

M = error(early_model, late_data) − error(late_model, late_data)

The observer trains a predictor on early emissions and tests it on late emissions. Separately, it trains a predictor on late emissions and tests it on the same late data. If the early model performs significantly worse, the system has moved persistently — the emission regime shifted and stayed shifted. The observer infers that the system "carries its history." High M means the emission profile changed in a way the early model cannot track. Zero M means the early model still works — no persistent change is visible.

This definition has three features that distinguish it from Rovelli's formalisation:

First, M is a property of the observation relation, not of the observed system. The same system, observed through different emission channels or at different temporal resolutions, can yield different values of M. An observer tracking one set of features may see strong memory where an observer tracking a different set sees none. Memory is jointly determined by the system's dynamics, the observer's feature selection, and the observer's protocol.

Second, M does not require temporal asymmetry. The measure compares prediction performance between early and late phases, but the comparison itself is symmetric — one could equally well train on late data and test on early data (this would detect "reverse memory," which is simply whether the emission profile shifted when read backward). The asymmetry of experienced memory — the fact that we remember the past, not the future — is accounted for by the framework's derivation of temporal ordering from geometric circulation at $N \geqslant 3$ (Section 2.5), not by a separate assumption about dissipation and branch systems.

Third, M makes no reference to information carried about a different system's past state. The observer doesn't need to establish that system M_0 carries information about system S at time t_0. The observer needs only its own emission stream and its own prediction model. The question "is this memory reliable?" becomes "does my early model's failure on late data reflect a genuine regime shift in the emissions, or is it an artefact of my model's limitations?" — which is a question about the observer's own inference, not about the physical reliability of traces.

3.3 Empirical Evidence Against Memory as Physical Trace

Three computational results from [4] challenge the branch-system formalisation directly.

A reaction-diffusion system habituates. A Gray-Scott PDE in a stable spot pattern, subjected to twelve identical mechanical perturbations at 50-timestep intervals, produces a declining response magnitude followed by partial recovery — the same emission profile that an experimentalist observing *Stentor coeruleus* would label "habituation with spontaneous recovery." Under Rovelli's formalisation, the PDE should not have memory: it does not carry information about its own past state in the way a branch system does. Under the observer-inference framework, the PDE has precisely as much "memory" as *Stentor*, because the observer's prediction machinery responds identically to both emission streams.

The observer's protocol determines the cognitive attribution. The same Gray-Scott system, subjected to the same perturbation, produces habituation at short tap intervals (30–80 steps), sensitisation at moderate intervals (200 steps), and no detectable learning at long intervals (400–800 steps). The system is identical across all seven conditions. The observer's choice of probing frequency — made before any data is collected — determines which category the observer constructs. Under a trace-based account of memory, the system either has the capacity for habituation or it does not. Under the observer-inference account, "habituation" is a label the observer attaches to a specific emission profile shape, and the shape depends on the interaction between system dynamics and observer protocol.

Restricting the observer's access increases apparent intelligence. Across five test systems spanning four substrates, the best single emission feature always produces a learning signal (L) equal to or greater than the full-access learning signal. Adding features to the observer's model can only dilute or maintain L — never increase it. For bubble sort, a single feature (inversions) gives L = 0.914; full access gives L = 0.689. The observer who sees less infers more learning from the same system.

This last result is a specific, quantitative instantiation of what we call the *cognitive hierarchy*: a smarter observer — one with greater access to the system's dynamics — sees less apparent intelligence [4]. The hierarchy holds for every observer who watches an informative channel. It inverts the standard assumption that better observation reveals more intelligence and instead predicts that cognitive attribution scales with the observer's ignorance.

3.4 Dissolving the Memory Reliability Problem

The Boltzmann brain problem, as framed by Wolpert et al., asks: are your memories reliable records of the past, or are they statistical fluctuations? The framework developed here suggests this is the wrong question — not because the answer is obvious, but because the question treats memory as a physical trace whose informational content can be evaluated for veridicality.

In the observer-inference framework, "memory" is the name the observer gives to the fact that its early-trained model fails on late data. The observer doesn't have access to "the past" — it has access to the current emission stream and the current state of its own predictive model. What the observer calls "remembering the past" is a new prediction from a new model state, not a readout from storage. The model was shaped by prior emissions (the observer's history of surprises), and the model's current failure or success pattern is what the observer interprets as "memory of the past."

Under this reframing, the BB question transforms. Instead of asking "are my traces reliable?" — which requires establishing the physical basis of the traces and the thermodynamic conditions that would make them veridical — the question becomes: "would an observer constructing predictions from this emission stream infer the same M/L/D profile regardless of whether the emission stream was generated by a system with genuine temporal history or by a statistical fluctuation?"

The answer, given the results in Section 3.3, is: it depends on the emission profile, not on its origin. If a fluctuation produces an emission stream with a regime shift (high M), decreasing surprise (high L), and intermittent prediction failures (high D), the observer will construct cognitive attributions from it — exactly as it would from a system with a genuine causal history. Conversely, if a system with a genuine causal history produces emissions with no regime shift, no surprise decrease, and no prediction failures, the observer will construct no cognitive attributions from it — the system will appear memoryless, regardless of its actual history.

This dissolves the BB memory problem not by establishing that memories *are* reliable, but by showing that "reliability" is itself an observer attribution that depends on the emission profile rather than on the metaphysical question of whether the emissions were produced by a fluctuation or by a historical process. The observer cannot distinguish the two cases from the emission stream alone. But the observer-inference framework shows that this limitation is not a bug — it is the structural condition of any observation under incomplete access. All memory is observer-constructed, whether the underlying dynamics are "genuine" or not.

3.5 Connection to Free Energy and Dissipation

Wolpert et al.'s central claim about memory — that reliable traces require dissipation toward the future — has a precise translation in the observer-inference framework.

The observer's learning measure L is the fractional decrease in mean surprise between early and late emission phases. Surprise in this context is the prediction error — the empirical proxy for negative log-evidence under the observer's model. Mean surprise over a phase is therefore a proxy for the expected variational free energy [13]. $L > 0$ means that the observer's free energy decreased from early to late: the model improved.

The correspondence to Friston's free energy principle is direct [10]: when we say "the system is learning," we are saying "the observer's free energy is decreasing" — which is precisely Friston's claim about biological systems, reframed as a property of the observation rather than of the system. The dissipation that Wolpert et al. require for reliable memory is, in this framework, the condition under which $L > 0$ — the observer's predictions improve, which is the condition under which the observer constructs the attribution "this system has memory."

But the framework adds something that neither Wolpert et al. nor Friston address. The dissipation is not what *makes* the memory reliable. The dissipation is what produces the emission profile that the observer *categorises as* reliable memory. An observer watching a non-dissipative system may see $L = 0$ and construct no memory attribution — not because the system lacks memory in some physical sense, but because the emission profile doesn't trigger the

observer's predictive model to register improvement. A Boltzmann fluctuation that happened to produce a dissipative emission profile would trigger the same attribution as a "genuine" dissipative process. The dissipation-reliability connection is real but it operates through the observer's inference, not through a direct physical relationship between traces and past states.

3.6 What This Section Establishes

We have argued that the memory component of the Boltzmann brain circularity can be dissolved by reframing memory from a physical trace (a system carrying information about the past) to an observer inference (a predictive model's obsolescence pattern). This reframing:

1. Removes the need to establish trace reliability independently of the laws used to evaluate it — the observer's model performance is self-contained.
2. Is supported by computational results showing that systems with no physical memory mechanism produce emission profiles that observers label "memory," "habituation," and "learning" identically to systems with plausible cognitive capacity.
3. Is consistent with the geometric dissolution of the second law circularity (Section 2) — both moves remove foundational assumptions (time as fundamental, memory as physical trace) that generate the circular dependencies.
4. Predicts a specific, testable cognitive hierarchy: observers with more complete access attribute less intelligence to the same system, scaling as $M \propto \log(D_{total}/D_{obs})$.

The result complements rather than replaces the Wolpert-Rovelli analysis. Their formalisation clarifies the circularity within the standard framework; the observer-inference framework dissolves it by changing the framework.

---

4. The Four Laws from One Axiom

4.1 Purpose of This Section

The arguments in Sections 2 and 3 dissolved specific circularities: the second law's entanglement with the past hypothesis (Section 2) and memory's entanglement with trace reliability (Section 3). Both dissolutions depend on the thermodynamic laws having the status of geometric theorems rather than independent postulates. This section presents the complete derivation chain, showing that all four laws emerge from the axiom $\Diamond N \rightarrow \neg N$ without additional thermodynamic assumptions. Full proofs are given in [5]; here we present the logical structure.

4.2 The Viable Region

From the axiom, any configuration must be distinguishable from nothingness ($D(C, 0)^2 \geqslant \varepsilon^2$) and from self-contradictory totality ($D(C, \mathbf{1})^2 \geqslant \varepsilon^2$). These conditions define a bounded viable region V in the five-dimensional constraint space — a shell between an inner boundary ∂V_{-}

(indistinguishable from nothing) and an outer boundary ∂V_+ (indistinguishable from everything). The interior V° is the arena within which all configurations exist [2].

Within V, the efficiency potential $\Phi = \ln(\Omega/K)$ assigns a scalar to each configuration, where Ω counts accessible relational states and K measures descriptive complexity. Φ is not entropy: it is closer to a negative free energy, balancing two competing factors — relational richness against pattern specificity. The gradient $\nabla\Phi$ defines the geometric structure that organises configurations relative to one another [3, 5].

4.3 Zeroth Law: Equilibrium Is Transitive

Standard statement: If system A is in thermal equilibrium with B, and B with C, then A is in equilibrium with C.

Derivation: Define equilibrium between configurations as the vanishing of directional gradients of Φ between them. The potential Φ is smooth (C^∞) throughout V°, since both Ω and K are continuous measures and the logarithm preserves smoothness. For configurations A, B, C:

If $\partial\Phi/\partial(\text{path } A\to B) = 0$ and $\partial\Phi/\partial(\text{path } B\to C) = 0$, then by smoothness $\partial\Phi/\partial(\text{path } A\to C) = 0$.

Temperature is then defined as the inverse of the τ-component of $\nabla\Phi$:

$$\frac{1}{T} \equiv \frac{\partial \Phi}{\partial E}$$

where E is the energy conjugate to the ordering parameter τ. The smoothness of Φ guarantees that T is single-valued, continuous, and identical for all configurations in mutual equilibrium — which is precisely what the zeroth law requires [5].

4.4 First Law: Energy Is Conserved

Standard statement: In any process, $dE = \delta Q - \delta W$.

Derivation: The constraint space inherits symplectic structure from the axiom's requirements [3, 6]. Conjugate pairs — including (τ, E) — satisfy canonical relations. The symplectic 2-form ω is closed ($d\omega = 0$), and by Noether's theorem, the τ-translation symmetry of Φ implies energy conservation:

$$\frac{dE}{d\lambda} = \{E, H\} = \{E, E\} = 0$$

where $\{\cdot,\cdot\}$ is the Poisson bracket. Energy transfer decomposes into work δW (changing macroscopic constraints β, κ, ρ) and heat δQ (changing microscopic configuration without changing macroscopic constraints), yielding the standard thermodynamic identity $dE = TdS - \Sigma F_i dC_i$ [5].

The first law is a consequence of the symplectic geometry of constraint space, which itself follows from the axiom's requirement that the conserved measure Ω be preserved under

configuration change.

4.5 Second Law: Entropy Does Not Decrease

Standard statement: The entropy of an isolated system never decreases: $dS \geq 0$.

Derivation: This is the geometric identity presented in Section 2.4. Along gradient flow lines, $d\Phi/d\lambda = |\nabla\Phi|^2 \geq 0$. The result follows from the inner product of any vector with itself being non-negative.

The connection to statistical entropy is established through the counting argument: configurations with higher Φ are exponentially more numerous than those with lower Φ. The measure $N(\Phi_0) \propto e^{\{\Phi_0\}}$ means that almost all configurations compatible with given constraints have Φ near its constrained maximum. When statistical interpretation is valid (large N, many configurations), the second law becomes $dS_{stat} \geq 0$, which is Boltzmann's formulation — now derived from the geometry of Φ rather than imported as an independent postulate [5].

This is the central point against the Wolpert-Rovelli conclusion. In their framework, the second law is a probabilistic property of a Markov process that requires temporal conditioning. In the present framework, it is a geometric identity about gradient flow in a bounded space. The two formulations are not equivalent, and the geometric formulation does not suffer from the conditioning circularity that the probabilistic formulation does.

4.6 Third Law: Absolute Zero Is Unattainable

Standard statement: As temperature approaches absolute zero, the entropy of a system approaches a minimum value. Equivalently: it is impossible to reach absolute zero in a finite number of steps.

Derivation: The inner boundary ∂V_- is defined by $D(C, 0)^2 = \varepsilon^2$. At ∂V_-, a configuration would be indistinguishable from nothingness — but the axiom states that nothingness cannot exist. Therefore ∂V_- is a limit that can be approached but never reached.

The correspondence to the third law is direct:

$$\Diamond N \to \neg N \iff \partial\mathcal{V}_- \text{ is unreachable} \iff T = 0 \text{ is unattainable} \iff S \to S_{min} > 0 \text{ as } T \to$$

The third law is the axiom in thermodynamic language. It is not a separate postulate but the founding logical principle of the entire framework, expressed in the vocabulary of temperature and entropy.

This identification is, to our knowledge, novel. The standard treatment of the third law gives it a different character from the other three — it is often presented as an empirical generalisation or a consequence of quantum mechanics rather than a foundational principle. Within the present framework, it is the most foundational of the four: it is the axiom itself, from which the other three are derived.

4.7 The Complete Logical Chain

The derivation chain is:

$$\Diamond N \rightarrow \neg N$$

↓ (nothing cannot exist)

Distinguishability is necessary

↓ (distinguishability requires relation)

Five constraint dimensions; bounded viable region V

↓ ($\Phi = \ln(\Omega/K)$ on V)

Smoothness of Φ → **Zeroth Law**

Symplectic structure → **First Law**

$|\nabla\Phi|^2 \geqslant 0$ → **Second Law**

∂V_{-} unreachable → **Third Law**

No thermodynamic postulate is introduced at any stage. The four laws are geometric properties of a potential function defined on a bounded region of a five-dimensional space, where both the potential and the region are consequences of the requirement that configurations be distinguishable from nothingness.

4.8 Implications for the Boltzmann Brain Problem

The unified derivation has a specific consequence for the Wolpert-Rovelli analysis. Their conclusion — that the BB hypothesis and the second law are "equally legitimate (or not)" — rests on both being conditioning choices within the entropy conjecture. But the derivation above shows that the four thermodynamic laws are not independent ingredients that can be selectively adopted or rejected. They are joint consequences of a single axiom. Accepting any one of them (for instance, the third law — the unattainability of absolute zero, which is rarely questioned) logically entails the others, including the second.

This does not settle the BB question by fiat. But it does change the terms of the debate. The Wolpert-Rovelli framing treats the second law as one conditioning choice among others, equally legitimate as the BB hypothesis. The present framework shows that the second law is entailed by the same logical structure that produces energy conservation, equilibrium transitivity, and the unattainability of absolute zero. A Boltzmann brain proponent who wishes to reject the second law must also reject the third — must accept that configurations can approach indistinguishability from nothingness — which is precisely what the axiom forbids.

---

5. Discussion

5.1 Two Approaches to the Same Problem

This paper and the Wolpert-Rovelli-Scharnhorst analysis [1] address the same tangle of problems — the Boltzmann brain hypothesis, the circularity of the second law, the reliability of memory — but from opposite directions. Their approach stays within the standard formalism and achieves clarity about dependencies that were previously implicit. They show that the BB hypothesis, the past hypothesis, and the second law share a common formal structure, and that the apparent conflicts among them arise from unstated choices about what to condition on. This is a diagnostic achievement: the circularity is not resolved but is precisely located and characterised.

The approach taken here removes the assumptions that generate the circularity. By deriving temporal ordering and the thermodynamic laws from a single axiom about distinguishability, the framework eliminates the need for conditioning choices — the second law is a geometric theorem, not a probabilistic statement requiring initial data. By reframing memory as observer inference rather than physical trace, the framework eliminates the need to establish record reliability — the observer's predictive model is self-contained and does not reference the physical basis of traces.

These are complementary rather than competing contributions. The diagnostic work clarifies what needs to be dissolved; the dissolution identifies the assumptions whose removal achieves it. A reader who accepts the Wolpert-Rovelli diagnosis but finds the axiom-based framework speculative still gains something: a precise identification of which assumptions (time as fundamental, memory as physical trace) are responsible for the circularity, and a demonstration that frameworks without those assumptions do not exhibit it. A reader who accepts the framework but is unfamiliar with the BB literature gains the Wolpert-Rovelli formalisation as a clear statement of the problem being addressed.

5.2 What Is and Is Not Claimed

Several distinctions should be made explicit.

We do not claim that the BB hypothesis is false. We claim that the circularity used to argue for and against it dissolves within a framework where time and the thermodynamic laws are derived rather than assumed. The BB hypothesis might still be entertained on other grounds — cosmological, measure-theoretic, or philosophical — but the specific circularity that Wolpert et al. formalise does not arise in the present framework.

We do not claim that Rovelli's formalisation of memory is wrong. Branch systems and physical traces are useful concepts for analysing how information persists in thermodynamic systems. We claim that the BB memory problem — "are my traces reliable?" — is a problem *about* this formalisation rather than a problem *in* physics. The observer-inference framework does not need the concept of trace reliability and therefore does not inherit the circularity that accompanies it.

This is a claim about the architecture of the explanation, not about the physics of information storage.

We do not claim to derive physics from logic. The axiom ◇N → ¬N generates mathematical structures — a bounded region in a five-dimensional space, a potential function with specific gradient properties, symplectic and unitary geometries in different topological regimes. These structures have detailed parallels to thermodynamic laws, Hamiltonian mechanics, and unitary quantum evolution [3, 6]. The identification of these mathematical structures with physical theories requires bridge assumptions about what the mathematical objects represent. The dissolution of the BB circularity holds to the extent that the bridge assumption $\Phi \leftrightarrow$ thermodynamic potential is accepted. The observer-inference results (Section 3) are independent of this bridge assumption and stand as computational findings within the standard HMM framework.

5.3 Rovelli's Relational Programme

It is worth noting that the observer-dependence thesis developed here is broadly compatible with Rovelli's relational quantum mechanics [11], which holds that physical quantities are relative to the observer rather than absolute. The present framework extends this relational commitment into the cognitive domain: memory, learning, and decision-making are relative to the observer's predictive model, not properties of the system in isolation.

The tension, such as it is, lies in the treatment of memory. Rovelli's formalisation of physical memory [7] defines it as a physical system carrying information about another system's past — a definition that treats the informational relationship as observer-independent. The observer-inference framework treats this informational relationship as itself observer-relative: what counts as "carrying information about the past" depends on which emissions the observer tracks, how frequently it probes, and what predictive model it maintains. This is a natural extension of the relational programme, pushing it into territory that the Boltzmann brain problem specifically requires.

Whether Rovelli would accept this extension is an open question. The relational programme has historically focused on quantum observables and spacetime events rather than on cognitive attributions. But the structural parallel is clear: just as "the spin of the electron" is relative to the measuring apparatus in relational QM, "the memory of the system" is relative to the observing model in the present framework. The BB problem, viewed through this lens, is asking whether physical quantities can be established without reference to an observer — and both relational QM and the observer-inference framework answer that they cannot.

5.4 The Observer-Dependence of "Interesting"

A subsidiary result from the MLD analysis [4] deserves mention in this context. The M∩L∩D overlap — the region where all three observer measures are simultaneously active — identifies the systems that human researchers independently classify as "complex" or "interesting." Class 4 cellular automata, Game of Life methuselahs, pattern-forming reaction-diffusion regimes, and

sorting algorithms undergoing structured convergence all cluster in this region, despite sharing no physics, no mathematics, and no structural features.

This suggests that "interesting" is itself an observer attribution — the label the observer attaches when its predictive model simultaneously exhibits obsolescence (M), improvement (L), and intermittent failure (D). The BB hypothesis, in a sense, asks whether the universe itself is "interesting" in this technical sense: does the emission stream of our experience exhibit the M/L/D profile that we label as historically grounded, or could it equally be produced by a fluctuation? The framework's answer is that the question is underdetermined from the emission stream alone — not as a limitation, but as a structural feature of observation under incomplete access. The same emission profile would trigger the same attributions regardless of its causal history.

5.5 Open Questions

Several questions remain that this paper does not address.

The bridge assumption. The identification of Φ with a thermodynamic potential is well-motivated by structural parallels but is not derived from the axiom. Establishing conditions under which this identification is necessary rather than merely sufficient would strengthen the dissolution considerably. The Foundations of Physics papers [3, 6] develop the mathematical side; the physical interpretation remains an active area of the research programme.

Cosmological context. Wolpert et al. explicitly bracket cosmological considerations, noting that the BB problem is sensitive to measure assignments in de Sitter cosmology and scalar field models [1]. The present framework similarly brackets cosmology. The constraint-space picture is compatible with cosmological dynamics — the viable region could expand, contract, or deform over cosmological scales — but we have not developed this connection. Whether the axiom imposes constraints on cosmological models (for instance, forbidding configurations that approach ∂V_{-} on cosmological timescales) is an open question with potentially interesting consequences for the BB debate.

Experimental tests. The observer-inference framework makes a specific quantitative prediction: apparent memory scales as $M \propto \log(D_{total}/D_{obs})$. This is testable in laboratory systems where D_{obs} can be systematically varied — chemical oscillators observed through different measurement modalities, neural recordings at different spatial resolutions, or ecological systems monitored with different sensor arrays. A multi-observer experiment on the Belousov-Zhabotinsky oscillator has been proposed [4] and would provide a direct test of whether cognitive attribution depends on observer access in the way the framework predicts. Such experiments would test the observer-inference framework independently of the axiom-based theoretical structure.

The philosophy loop. The observer who constructs M/L/D attributions from a system's emissions is itself a system whose emissions could be analysed by a second observer. The second observer would construct M/L/D attributions about the first observer's cognitive process. This recursive structure — the philosophy loop — is developed in a companion paper [12]. Its

relevance to the BB problem is that any argument about whether "our" memories are reliable is itself an emission stream that a meta-observer would process through the same predictive machinery. The loop does not close viciously; it closes productively, showing that the observer-inference framework applies to its own operation.

5.6 Conclusion

The Boltzmann brain problem, as formalised by Wolpert, Rovelli, and Scharnhorst, reveals circular dependencies among the second law, the past hypothesis, and the reliability of memory. These circularities arise from treating time as a fundamental parameter for conditioning and memory as a physical trace whose reliability must be independently established. Within a framework where time emerges from geometric circulation at $N \geqslant 3$ and the thermodynamic laws are geometric theorems derived from a single axiom about distinguishability, these circularities dissolve. The second law is not a conditioning choice but a geometric identity. Memory is not a physical trace but an observer's prediction performance. The question "are my memories reliable?" transforms into "what would an observer construct from this emission stream?" — a question that does not require the assumptions whose entanglement generates the circularity.

The dissolution does not settle the BB question. It changes the terms. The debate is no longer about which conditioning event to choose from within a time-symmetric Markov process. It is about whether the mathematical structures that emerge from the distinguishability axiom correspond to the thermodynamic structures of our universe. If they do, the BB circularity was never a problem in physics. It was a problem in the assumptions.

References

[1] D. Wolpert, C. Rovelli, and J. Scharnhorst, "Disentangling Boltzmann Brains, the Time-Asymmetry of Memory, and the Second Law." *Entropy* 27(12), 1227 (2025). DOI: 10.3390/e27121227

[2] D. Neale, "Being from Nothingness: Logical Foundations of Physical Structure." Preprint available at goleudy.ai.

[3] D. Neale, "Mathematical Structures from a Distinguishability Axiom." Submitted to *Foundations of Physics*.

[4] D. Neale, "Memory Without Storage, Learning Without a Learner: Observer Inference Across Four Computational Substrates." Preprint available at goleudy.ai (2026).

[5] D. Neale, "Thermodynamic Foundations from Constraint Geometry." Supporting Information, available at goleudy.ai.

[6] D. Neale, "Three-Dimensional Geometry from a Distinguishability Axiom." Submitted to *Foundations of Physics*.

[7] C. Rovelli, "Back to Reichenbach." *Journal for General Philosophy of Science* (2024).

[8] D. Wolpert and J. Kipper, "Physical Memory." Preprint (2024).

[9] H. Reichenbach, *The Direction of Time* (University of California Press, 1991).

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

[11] C. Rovelli, "Relational Quantum Mechanics." *International Journal of Theoretical Physics* 35, 1637–1678 (1996).

[12] D. Neale, "The Philosophy Loop: Why Observers Invent Intelligence." In preparation (2026).

[13] D. Neale, "Free Energy Correspondence." Appendix D to [4], available at goleudy.ai (2026).