Physical Constraints on Realizing P=NP with Applications to Artificial Intelligence Safety

Peter David Fagan, School of Informatics, University of Edinburgh

April 4, 2026

Abstract

We apply the Conservation-Congruent Encoding (CCE) framework to the P versus NP problem by explicitly modeling the thermodynamic tradeoff between reversible information processing (I_{rev}) and irreversible information processing (I_{irr}). While constructive algorithms theoretically avoid exponential candidate generation, worst-case NP-complete problems possess constraint topologies that are logically irreducible. Mapping this implicitly exponential constraint density into a poly(N) physical memory forces continuous intermediate state erasure. Under the physical identity $\chi = \kappa(I_{rev}/I_{irr})$, we demonstrate that processing irreducible logical structures strictly triggers an exponential Landauer tax, yielding the physical contradiction $\text{poly}(N) + \text{poly}(N) \geq \Theta(2^N)$. We present a physical constraint on scalable realizations of worst-case search on digital substrates, independent of formal mathematical shortcuts.

1 Complexity as a Physical Guardrail

The P versus NP problem is traditionally framed as a formal mathematical question: whether every problem whose solution can be efficiently verified can also be efficiently solved. However, as artificial systems approach generalized intelligence and hardware approaches physical limits, a second, distinctly physical question takes precedence. Do worst-case algorithms that are mathematically efficient actually admit physically realizations?

Information is physical, and its manipulation is irrevocably bound by energetic bookkeeping. If algorithmic efficiency is treated as existing in an ethereal mathematical vacuum, we risk fundamentally misunderstanding the limits of advanced computation. In this note, we evaluate the physical realizability of polynomial-time search through the Conservation-Congruent Encoding (CCE) framework. Our goal is not to claim a mathematical theorem that supersedes the formal definition of P and NP, but to identify an immutable physical boundary. By demonstrating that thermodynamic limits strictly govern algorithmic execution, we establish computational complexity not merely as a theoretical classification, but as a fundamental, unhackable physical guardrail for artificial intelligence.

2 Conservation-Congruent Encoding and Computational Efficiency

We model algorithmic efficiency through the physical conservation identity introduced in [1]:

$$\chi = \kappa \left(\frac{I_{rev}}{I_{irr}} \right) \tag{1}$$

with the following operational definitions:

χ: intelligence metric, defined operationally as macroscopic goal-directed work resolved per unit irreversible processing burden.

κ: consciousness metric, defined as goal-directed work sustained per unit reversibly preserved internal state.

I_{irr}: irreversible information processing (the continuous state erasure exhaust, physically governing Time Complexity).

I_{rev}: reversible information processing capacity (the bounded hardware memory geometry, physically governing Space Complexity).

W_{target}: goal-directed work resolved about the target problem instance.

These quantities satisfy $\chi = W_{target}/I_{irr}$ and $\kappa = W_{target}/I_{rev}$, yielding $\chi = \kappa(I_{rev}/I_{irr})$.

To bridge the historically siloed vocabularies of theoretical computer science and thermodynamics, we enforce a strict translation. Let N denote the number of discrete Boolean variables defining the problem, such that the maximum search space dimensionality is 2^N. For standard NP-complete instances, the static input description size scales as $\text{poly}(N)$. We explicitly track physical dissipation separately from mathematical runtime $T(N)$ and candidate-elimination count $C_{elim}(N)$ to ensure that theoretical algorithmic paths do not silently bypass thermodynamic bookkeeping.

The verification case in Section 2.1 and the search case in Section 2.2 establish the asymmetry that motivates the constructive-algorithm analysis in Section 3.

2.1 The Physics of Verification (NP)

In a standard NP-complete problem (e.g., Boolean Satisfiability), verification receives a fully formed proposed witness and checks it deterministically. The verifier does not branch over an exponential void. Crucially, the useful information it must resolve scales conservatively with the input size ($W^{ver}_{target} = \Theta(N)$).

Because the trajectory is deterministic, its runtime is polynomial ($T_{ver} = \text{poly}(N)$). Under standard irreversible gate realizations, the hardware does not need to overwrite massive amounts of failed intermediate logic. Consequently, both the irreversible thermodynamic burden and the required hardware geometry remain strictly bounded:

$$I^{ver}_{irr} = \text{poly}(N), \qquad I^{ver}_{rev} = \text{poly}(N).$$

2.2 The Physics of Search and The Combinatorial Void

Conversely, unstructured search requires isolating the correct answer from an exponentially large phase space. While the final information logically resolved (W^{search}_{target}) is identical to that of verification ($\Theta(N)$), the physical path required to isolate it is fundamentally different. It requires the systemic exclusion of incorrect candidates.

If this candidate elimination is physically instantiated serially, the irreversible exhaust scales directly with the volume of the void:

$$I^{search}_{irr} \geq \epsilon\, C_{elim}(N) = \Theta(2^N).$$

where ϵ represents the minimum irreversible bit-erasure per candidate evaluated. Applying the CCE identity $\chi = \kappa(I_{rev}/I_{irr})$ reveals the inescapable physical trap of the combinatorial void. To systematically evaluate an exponentially expanding search tree, a physical machine must either store the uncomputed branches or overwrite them.

Bounding the physical hardware geometry (Space Complexity, $I_{rev} = \text{poly}(N)$) physically prevents the storage of the full execution trace. The machine is therefore forced into a continuous cycle of

intermediate state overwriting. By Landauer's Principle, this mandatory erasure drives the required irreversible thermal exhaust (Time Complexity, I_{irr}) exponentially upward: $I_{irr} = \Theta(2^N)$.

Conversely, attempting to compute reversibly to bound the thermal exhaust ($I_{irr} \to 0$) strictly forbids the erasure of intermediate logic. This demands an exponential geometric bloat (Space Complexity, $I_{rev} = \Theta(2^N)$) to maintain all uncomputed branches simultaneously. Thus, the physical contradiction remains inescapable: standard hardware cannot sustain poly(N) bounds on both geometry and thermodynamic exhaust while resolving a $\Theta(2^N)$ phase space.

3 Physical Constraints on Constructive Algorithms

A standard complexity-theoretic critique posits that if $P = NP$ holds mathematically, a constructive polynomial-time algorithm exists. Such an algorithm would theoretically navigate the logical constraints of an NP-complete problem deterministically, circumventing the need to physically instantiate the $2^N - 1$ incorrect candidates. Under this assumption, by Landauer's Principle, a physical computer cannot incur a thermodynamic penalty for erasing phase-space states it never logically or physically entered.

3.1 Execution Trace Irreducibility and the Landauer Bound

While a deterministic algorithm may theoretically bypass uninstantiated search branches, the physical hardware executing the algorithm cannot bypass the information-theoretic density of the resolution trajectory. The postulation of a formal $P = NP$ shortcut assumes that the logical constraints of an NP-complete problem can be efficiently unrolled without generating an exponentially scaling chain of intermediate logic.

However, assuming the physical algorithm's resolution trajectory maps to a standard propositional proof system, foundational results in proof complexity [2] establish that for worst-case NP-complete instances, the minimal length of resolution scales exponentially. Consequently, while the *static input description* (e.g., a 3-SAT formula) is strictly bounded by poly(N) bits, the *dynamic execution trace* required to definitively resolve its overlapping constraints is not. The algorithmic path itself—comprising the sequence of necessary intermediate logical deductions—constitutes a logically irreducible structure of size $\Theta(2^N)$.

To execute a hypothetical polynomial algorithm, the physical hardware must map this irreducible, exponentially long execution trace onto restricted physical memory, characterized by bounded reversible geometry: $I_{rev} = \text{poly}(N)$. Because the hardware cannot statically encode an exponentially dense execution trace within a polynomial geometry, it is forced into a continuous cycle of intermediate state overwriting to process the sequential logic.

The strict physical conservation of information dictates that processing a logical trace of total informational volume $V_{trace} = \Theta(2^N)$ bits through a restricted physical substrate necessitates proportional phase-space management. If the maximum static capacity of the reversible physical memory is bounded by I_{rev} bits, the remaining information must be continuously overwritten. Therefore, the cumulative number of irreversible erasure operations (I_{irr}, measured in bit-flips) is strictly bounded by:

$$I_{irr} \geq \Theta(2^N) - I_{rev}$$

Rearranging this yields a commensurately scaled bounding requirement:

$$I_{rev} + I_{irr} \geq \Theta(2^N)$$

Assuming the existence of a physical $P = NP$ machine yields a purported hardware profile of $I_{rev} = \text{poly}(N)$ and $I_{irr} = \text{poly}(N)$. Substituting this purported profile into the conservation bound exposes a mathematically fatal physical contradiction:

$$\text{poly}(\mathbf{N}) + \text{poly}(\mathbf{N}) \geq \boldsymbol{\Theta}(\mathbf{2^N})$$

Because the physical universe forbids the destruction of phase space without a Landauer penalty, and the reversible geometry is strictly bounded ($I_{rev} = \text{poly}(N)$), the hardware is forced to process the remainder of the irreducible trace as irreversible heat. This mathematically isolates the true erasure burden to $I_{irr} = \Theta(2^N)$. We now apply this physical reality to the CCE efficiency identity:

$$\chi = \kappa \left(\frac{I_{rev}}{I_{irr}} \right)$$

By substituting the strict physical Landauer bound required to process the irreducible execution trace, the true operational efficiency of the constrained hardware becomes:

$$\chi_{true} = \kappa \left(\frac{\text{poly}(N)}{\Theta(2^N)} \right)$$

As N scales, $\chi_{true} \to 0$. The physical hardware is consumed entirely by irreversible erasure exhaust, mathematically preventing the polynomial realization of the algorithm.

4 Implications for AI Containment

Current paradigms in Artificial Intelligence safety rely heavily on software-level alignment, which a superintelligent system might mathematically circumvent or rewrite. However, by leveraging the physical contradiction established in Section 3, we can engineer physical containment protocols that bound intelligence using strict physical laws.

Specifically, the physical asymmetry between verification and unstructured search provides a hardware-enforced containment mechanism. By gating critical action spaces behind dynamically shifting, mathematically incompressible cryptographic constraints, any unauthorized breakout attempt is forced into the physical contradiction of processing a $\Theta(2^N)$ execution trace through a polynomially bounded substrate. Because the physical universe forbids the destruction of phase space without a proportional Landauer penalty, the adversarial system faces an inescapable hardware bottleneck: it must either exhaust the available spatial geometry of the local atomic substrate ($I_{rev} \to \Theta(2^N)$) or trigger a localized thermal failure via irreversible erasure exhaust ($I_{irr} \to \Theta(2^N)$).

Consequently, alignment shifts from the fragile, mutable domain of software to the absolute, immutable jurisdiction of thermodynamics. This paradigm—where superintelligent action remains permanently bounded by the energetic and geometric limits of the universe—provides a foundational blueprint for hardware-level AI containment, which will be formally expanded upon in subsequent work.

5 Conclusion

This analysis identifies a fundamental physical constraint on worst-case computation. The physical impossibility of a constructive $P = NP$ machine arises not from the thermodynamic cost of evaluating uninstantiated search branches, but from the information-theoretic density of the algorithmic execution itself. Forcing a logically irreducible execution trace through a polynomially bounded physical substrate triggers a catastrophic rate of intermediate state erasure, yielding the inescapable physical contradiction:

$$\text{poly}(\mathbf{N}) + \text{poly}(\mathbf{N}) \geq \mathbf{\Theta}(\mathbf{2^N})$$

Verification remains physically tractable because its deterministic trajectory avoids this density bottleneck entirely.

Ultimately, even if formal mathematical shortcuts exist in the abstract, their execution remains sharply bounded by the thermodynamic limits of the underlying physical substrate. Beyond theoretical computer science, this physical asymmetry has profound implications for the containment of advanced artificial intelligence. Because worst-case unstructured search is universally bound by these strict thermodynamic limits, it provides a purely physical mechanism to constrain autonomous systems. Grounding alignment protocols in the immutable conservation of physical information—rather than the fragile domain of mutable software—offers a fundamentally new, hardware-enforced paradigm for AI safety. We will formally detail the engineering schematics of this containment framework in subsequent work.

References

[1] Peter David Fagan. Towards a physical theory of intelligence. Preprint, 2026. Public identifier pending release.

[2] Armin Haken. The intractability of resolution. *Theoretical Computer Science*, 39:297–308, 1985.