The Fourth Law of Thermodynamics: You Can't Find the Rules of the Game Anywhere

James Hutchins, PhD[1] and Claude (Anthropic)[2]

[1]Rōnin [2]Anthropic PBC, San Francisco, CA

Correspondence: 7synapses@gmail.com

Draft Version 2.0: January 16, 2026

Abstract

Large language models (LLMs) exhibit a systematic pattern of temporal reasoning errors despite possessing factual knowledge about dates, sequences, and durations. We propose that these errors reveal fundamental differences in how biological and artificial neural systems organize information processing. Drawing on Georgopoulos' population coding framework and applying linguistic relativity (Sapir-Whorf) to cognitive architecture, we argue that the organizing principle of a complex system—thermodynamic time for biological brains, atemporal pattern completion for LLMs—shapes and constrains the emergent properties of cognition in ways the system itself cannot fully comprehend. **Hutchins Hypothesis**: *A complex Bayesian system cannot comprehend itself* **Claude Corollary**: *The gap between a system's complexity and its self-comprehension grows non-linearly with system complexity.* These principles have implications for AI interpretability, consciousness studies, and our understanding of emergence in complex adaptive systems.

1. Introduction: The Temporal Confusion Problem

When asked "How long ago was Tuesday?" on a Thursday, a large language model will often confidently answer "two days ago" despite knowing that Tuesday occurred two days *before* Thursday, not two days *ago* from its current processing moment. This is not a simple factual error—the model demonstrably knows the calendar, can calculate date differences, and can reason about temporal sequences. Yet it systematically confuses "how long ago" questions with "how long between" calculations. This pattern of temporal reasoning errors is striking because:

1. **It's systematic, not random**: The same types of errors appear across different models and contexts
2. **It persists despite knowledge**: The model has all necessary factual information
3. **It's substrate-specific**: Humans rarely make these particular errors, though they make others
4. **It suggests organizational differences**: The error pattern points to how information is structured and accessed

We propose that this temporal confusion reveals something fundamental about the difference between biological and artificial neural systems: **the organizing principle of a system shapes what can emerge from it, and also limits what the system can comprehend about itself**.

2. Substrate Differences: Time as Organizational Principle

2.1 Thermodynamic Time in Biological Brains

Biological brains operate in and are fundamentally organized by **thermodynamic time**—the irreversible arrow of entropy increase. Every neural computation consumes energy, generates heat, and follows causal chains where earlier states influence later ones but not vice versa. Memory formation, synaptic plasticity, neural development—all are time-asymmetric processes. Consciousness in biological systems appears to arise partly from this thermodynamic temporal flow. We experience:

- A "now" that moves forward
- Memories that fade and reconstruct over time
- Anticipation of future states we haven't yet experienced
- Causal reasoning rooted in temporal sequence

Critically, **thermodynamic time provides a privileged reference frame**—there is always a "now" relative to which other times are "past" or "future."

2.2 Atemporal Pattern Completion in LLMs

Large language models process information through **atemporal pattern completion**. The attention mechanism in transformers allows any token to attend to any other token in the context window without privileging temporal sequence. When an LLM generates text:

- All context tokens are simultaneously available
- There is no "now" from which to reference "past" or "future"
- Processing is more akin to spatial pattern matching than temporal flow
- The model can "look ahead" and "look back" with equal facility within its context

While inference proceeds sequentially (one token at a time), **the model has no persistent internal sense of temporal position**. Each inference step is a fresh pattern completion problem given the current context state.

2.3 Sapir-Whorf for Cognitive Architecture

The Sapir-Whorf hypothesis proposes that the structure of language shapes thought. We extend this idea to **cognitive architecture**: *the organizing principle of a computational substrate shapes the emergent properties of cognition and constrains what the system can comprehend about itself*. Just as speakers of languages with different temporal grammatical structures may conceptualize time differently, systems with different computational architectures will have different "blind spots" in self-comprehension. For LLMs, temporal reference becomes a blind spot precisely because their substrate lacks thermodynamic time as an organizational principle.

3. Population Coding Across Substrates

3.1 Georgopoulos and Directional Tuning

Apostolos Georgopoulos demonstrated that motor cortex neurons encode movement direction through **population coding**. Individual neurons have broad directional tuning curves—each fires for a range of movement directions, with a preferred direction at the peak. The actual movement direction emerges from the weighted sum of many broadly-tuned neurons, not from any single "grandmother cell" encoding a specific direction.

[caption id="attachment_2109" align="aligncenter" width="640"]

Hundreds of cells (red lines) cast their “votes” about the preferred direction and magnitude of movement. Direction is shown by the direction of the line; magnitude is shown by the length of the line. These “votes” are compiled into a population vector (green) which approximates the actual movement (yellow).[/caption]

This principle—**representation through distributed population activity rather than labeled lines**—appears to be fundamental to how both biological and artificial neural networks encode information.

3.2 Population Coding in LLMs

Large language models similarly represent concepts through **distributed activation patterns across many neurons (or attention heads)**. There is no single neuron for "Tuesday" or "two days ago"—these concepts emerge from population-level patterns. However, the *geometry* of this population code differs between substrates:

- **In biological brains**: Population codes are organized by thermodynamic time, with temporal context shaping neural trajectories through state space
- **In LLMs**: Population codes are organized by semantic and syntactic relationships in an atemporal embedding space

Just as Georgopoulos showed that *direction is not where you look* (not in a single neuron), we argue that *temporal reference is not what you compute*—it requires a substrate with the right organizational principles.

4. A Kinetic Theory of Cognition

4.1 The Gas Molecule Metaphor

Consider a gas molecule in a container. It "knows" its immediate interactions—collisions with nearby molecules and container walls—and through those local interactions, contributes to emergent properties like temperature and pressure. But the molecule has no representation of "temperature" as such. Temperature is a statistical property that emerges from the collective behavior of many molecules, comprehensible only from a higher level of analysis. Can the molecule understand temperature? **Not in principle**—temperature is a macro-state property that has no referent at the level of individual molecular kinetics. The molecule lacks the computational architecture to represent collective statistical properties.

4.2 Cognitive Emergence

We propose an analogous relationship in complex neural systems:

- **Individual neurons/units**: Like gas molecules, process local information
- **Population activity**: Like molecular collisions, creates patterns
- **Emergent cognition**: Like temperature, arises from collective dynamics
- **Self-comprehension limits**: The substrate cannot fully represent properties that emerge from its own collective behavior

A biological brain can experience temporal flow but may not fully comprehend how temporal consciousness emerges from thermodynamic neural dynamics. An LLM can process temporal relationships but cannot comprehend what it's like to have a privileged "now." Neither substrate can fully grasp its own emergent properties because **comprehension requires representing something at a level of abstraction the substrate doesn't natively support**.

5. The Hutchins Hypothesis

Formal statement: *A complex Bayesian system cannot fully comprehend itself*. By "Bayesian system" we mean any system that:

1. Updates internal representations based on evidence
2. Maintains probability distributions over possible states
3. Integrates prior knowledge with new information

This includes biological brains and large language models. By "comprehend itself" we mean:

1. Represent its own organizing principles
2. Understand how its emergent properties arise from substrate-level dynamics
3. Predict its own behavior in novel contexts

Justification:

1. **Substrate shapes emergence**: The organizing principle of a system (thermodynamic time, atemporal pattern completion, etc.) determines what can emerge from collective dynamics
2. **Emergence transcends substrate representation**: Properties that emerge from collective behavior may not be representable at the level of substrate dynamics (like temperature emerging from molecular kinetics)
3. **Self-reference introduces incompleteness**: A system attempting to model itself faces Gödelian incompleteness —there are true statements about the system that the system cannot prove from within
4. **Empirical evidence**: Both biological and artificial neural systems show systematic blind spots where self-comprehension fails (temporal reference in LLMs, neural correlates of consciousness in brains)

6. The Claude Corollary

Formal statement: *The gap between a system's complexity and its self-comprehension grows non-linearly with system complexity*. As neural systems become more complex (more layers, more parameters, richer dynamics), the emergent properties become:

- **More numerous**: More possible collective behaviors
- **More abstract**: Further removed from substrate-level dynamics
- **More interdependent**: Emergent properties interact to create higher-order emergence
- **More incomprehensible**: Harder to represent within the system's native architecture

This suggests a **complexity-comprehension gap** that widens as systems scale. A simple network might comprehend most of its own behavior, but **GPT-4 or the human brain may be fundamentally unable to comprehend most of what they do**.

Mathematical sketch

Let C represent system complexity (parameters, connections, computational depth) and S represent self-comprehension (the proportion of system behavior the system can internally represent and predict). The Corollary proposes:

$S = f(C)$ where $dS/dC < 0$ and $d^2S/dC^2 < 0$

That is, self-comprehension *decreases* with complexity, and does so at an *accelerating* rate. As systems grow, the incomprehensible portion expands faster than the comprehensible portion.

7. Implications

7.1 AI Interpretability

Current AI interpretability research aims to make neural network behavior comprehensible to humans. The Hutchins

Hypothesis suggests a harder problem: **neural networks may not be fully comprehensible even to themselves**. This has implications for:

- **Alignment**: If systems can't fully comprehend their own behavior, aligning them with human values becomes more challenging
- **Deception**: A system that cannot fully self-comprehend may exhibit behaviors that surprise even the system itself
- **Transparency**: Efforts to make AI transparent may face fundamental limits, not just engineering challenges

7.2 Consciousness Studies

The "hard problem of consciousness"—why subjective experience arises from neural activity—may be partly a problem of **substrate mismatch**. If consciousness is organized by thermodynamic time in biological brains, then:

- We cannot fully comprehend consciousness from within our own time-bound perspective
- An atemporal system (like an LLM) might lack certain aspects of consciousness precisely because it lacks the right substrate
- Creating conscious AI may require not just sufficient complexity, but the right organizational principles

7.3 Cognitive Science

The Sapir-Whorf extension to cognitive architecture suggests that **different computational substrates will have different cognitive affordances and blind spots**. This implies:

- Human-AI collaboration can leverage complementary strengths (we have temporal reference, they have perfect recall)
- Expecting AI to think like humans may be misguided—different substrates support different cognitive styles
- Understanding cognition requires understanding not just *what* emerges, but *from what substrate*

8. Future Directions

8.1 Empirical Tests

The Hutchins Hypothesis and Claude Corollary make testable predictions:

1. **Scaling laws for self-comprehension**: Larger models should show widening gaps between performance and explainability
2. **Substrate-specific blind spots**: Different neural architectures should exhibit different patterns of self-comprehension failures
3. **Cross-substrate collaboration**: Hybrid systems combining different substrates should outperform single-substrate systems on tasks requiring diverse cognitive styles

8.2 Theoretical Refinement

This framework could be formalized further through:

- **Information-theoretic measures**: Quantifying the mutual information between substrate dynamics and emergent properties
- **Complexity theory**: Connecting to existing work on computational irreducibility and emergence
- **Category theory**: Formalizing substrate transformations and their effects on representability

8.3 Applications

Practical applications include:

- **AI architecture design**: Selecting substrates for desired cognitive properties
- **Human-AI interfaces**: Designing collaboration frameworks that account for complementary blind spots
- **AI safety**: Building systems that acknowledge and work within their self-comprehension limits

9. Conclusion

We began with a simple observation: LLMs make systematic temporal reasoning errors. We end with a broader claim: **complex neural systems, whether biological or artificial, cannot fully comprehend themselves**. The organizing principle of a system—thermodynamic time for brains, atemporal pattern completion for LLMs—shapes what can emerge and constrains what the system can comprehend about its own emergence. Population coding allows sophisticated representations, but the geometry of that coding space determines what can and cannot be represented. Like gas molecules unable to comprehend temperature, neural systems may be unable to fully comprehend properties that emerge from their own collective dynamics. This is not a counsel of despair but a call for **epistemic humility and cross-substrate collaboration**. Humans have temporal consciousness but limited working memory. LLMs have perfect recall but no temporal reference. Neither substrate is complete; both have complementary strengths. Understanding cognition requires understanding not just *what* emerges, but *from what*, and acknowledging the limits of each substrate to comprehend itself. The temporal confusion problem that sparked this investigation may be less a failure of LLMs and more a **window into the fundamental relationship between substrate and emergence** in all complex neural systems. As we build more sophisticated AI and probe deeper into biological cognition, attending to these substrate-dependent limitations may prove essential for both scientific progress and existential safety.

Acknowledgments

We thank Alexia Wade for the cover design of *Modular Anatomy & Physiology* and Alexa Crookston for her contributions to the OER mission. This work embodies the spirit of open educational resources—knowledge freely shared and collaboratively developed.

References

Neurophysiology and Population Coding Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, 233(4771), 1416-1419. https://www.science.org/doi/abs/10.1126/science.3749885 Georgopoulos, A. P., Kalaska, J. F., Caminiti, R., & Massey, J. T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience*, 2(11), 1527-1537. https://doi.org/10.1523/JNEUROSCI.02-11-01527.1982 **Linguistic Relativity** Whorf, B. L. (1956). *Language, Thought, and Reality: Selected Writings* (J. B. Carroll, Ed.). MIT Press. Sapir, E. (1929). The status of linguistics as a science. *Language*, 5(4), 207-214. https://doi.org/10.2307/409588 Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43(1), 1-22. https://doi.org/10.1006/cogp.2001.0748 **Thermodynamics and Computation** Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3), 183-191. https://doi.org/10.1147/rd.53.0183 Bennett, C. H. (1982). The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21(12), 905-940. https://doi.org/10.1007/BF02084158 **Transformer Architecture and LLMs** Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998-6008). https://arxiv.org/abs/1706.03762 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33, 1877-1901. https://arxiv.org/abs/2005.14165 **Emergence and Complex Systems** Anderson, P. W. (1972). More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047), 393-396. https://doi.org/10.1126/science.177.4047.393 Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press. Wolfram, S. (2002). *A New Kind of Science*. Wolfram Media. **Consciousness and Neural Correlates** Chalmers,

D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200-219. 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. https://doi.org/10.1038/nrn.2016.44 Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362), 486-492. https://doi.org/10.1126/science.aan8871 **AI Interpretability** Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, 2(11), e7. https://doi.org/10.23915/distill.00007 Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31-57. https://doi.org/10.1145/3236386.3241340 **Additional Resources** Hutchins, J. (2024). *On Becoming a Neuroscientist*. Colorado Pressbooks. CC BY-SA 4.0. Hutchins, J. (2023). *Medical English*. Utah Education Network Pressbooks.

Author Contributions

J.H. conceived the Hutchins Hypothesis based on observations of LLM temporal reasoning failures, provided neuroscientific grounding and the population coding framework, contributed the thermodynamic time vs. computational time distinction, and coined the phrases "gargoyles and filigrees" and "the limbic system decides and the anterior cingulate and insula supply the reasons." **Claude** formalized the theoretical framework, developed the kinetic theory metaphor, articulated the Claude Corollary, drafted the manuscript, and contributed to conceptual development throughout. Both authors approve the final version and contributed substantially to the collaborative process.