philosophies

MDPI

Article

When the Ghost Emerges from the Machine: Limits of Semantic Decoding from Complete Microstate Knowledge

Jeffrey Arle

Department of Neurosurgery, Boston Medical Center—Brighton, Boston, MA 02135, USA; jeffrey.arle@bmc.org

Abstract

Understanding how high-level meanings emerge from low-level microstate dynamics is a central challenge in both artificial intelligence and consciousness studies. Complex networks can exhibit rich behaviors, yet reliably mapping every microstate onto a semantic label to date seems intractable. To explore these limits, a minimal 4-bit model consisting of only a ring of binary cells updated by a parity-flip rule, coupled with a finite lookup table that assigns conceptual tags to selected microstates, is presented. Two core failure modes are noted. First, noise is found to push the system into out-of-training-set states that a semantic decoder cannot label ("missing-label" errors). Second, distinct microstates collapse into the same semantic tag ("many-to-one" grouping), obscuring their unique identities. These findings demonstrate inherent opacity in semantic mapping and suggest fundamental barriers to reverse-engineering high-level content in artificial or biological networks. Future work includes scaling N and examining partial-observability effects.

Keywords: semantic decoding; computational intractability; cellular automata; microstate-to-meaning mapping

1. Introduction

Histories of mechanistic determinism long precede Laplace's 1814 Demon formulation. Greek atomists like Democritus and later Stoic philosophers argued that every event unfolds from prior states under fixed natural laws. In the Enlightenment, Maupertuis (1756), Condorcet (1768) and d'Holbach (1770) each envisioned an intellect possessing exact forces and positions capable of perfect prediction, while Boscovich's *Theoria Philosophiae Naturalis* (1758) outlined a proto-demon that computes full material trajectories [1]. Laplace's Demon thus became the emblem of an "all-knowing calculator," but it was not the first (cf. van Strien [1]).

A similar faith permeates modern cognitive science and artificial intelligence (AI): that recording every neuronal membrane potential or transistor state should, in principle, let us "read off" thoughts or narratives. Here, it is contended that even within a fully specified, transparent system, this *semantic* access is limited by two factors: (1) the absence, or mis-specification, of an explicit decoder of the underlying 'microstate', and (2) the prohibitive computational cost of operating such a decoder. For the purposes of this paper, 'semantic meaning' refers to the high-level, human-interpretable concepts, narratives and intentions that emerge from underlying microstates, while 'thought' denotes the dynamic process of generating and transforming these meanings.

To illustrate these constraints, a minimal cellular automaton (CA) with four binary units mapped to simple semantic labels (e.g., "Curiosity," "Doubt") is introduced, referred

Check for updates

Academic Editors: Marcin J. Schroeder, James Hutson and Luca M. Possati

Received: 2 October 2025
Revised: 24 February 2026
Accepted: 27 February 2026
Published: 19 March 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

to herein as 'ThoughtBox'. ThoughtBox's rule table, state space and bit-to-label map are all fully exposed and transparent at all times. Nevertheless, without a decoder model the 4-bit patterns convey no meaning, but even with it—once scaled to longer trajectories or with added noise—the decoding problem rapidly outstrips realistic computational budgets. Analogous challenges arise in biological systems: whole-brain voltage imaging in C. elegans or Drosophila, for example, where their entire nervous systems have been elucidated, produces massive datasets, yet inferring circuit-level semantics demands detailed models of connectivity and context [2,3].

Resource limits exacerbate semantic opacity. The human brain's ~10^11 neurons and ~10^14 synapses yield on the order of 2^(10^14) possible activation states per millisecond [4,5]. For comparison, a mere 100-city Travelling Salesman paradigm, a well-known state-space traversal, requires ~10^157 route evaluations, already beyond the ~10^109 operations the universe itself could have ever performed in its entire span of existence if all known particles were engaged in the computational process [6]. A full-state "demon" for the brain would exhaust cosmic computational resources and energy long before decoding a single thought.

Even engineered systems suffer parallel barriers. As a 'thought experiment', knowing every transistor state in a television yields raw logic-level traces, not scene or plot. To reconstruct the broadcast one needs precise circuit-logic mappings, timing standards (NTSC/PAL or HDMI), and a narrative decoder that interprets frame sequences into characters and dialogue. Minor variations—alternate cuts, dubbed audio—can produce transistor-level traces that are indistinguishable at coarse resolution, yet yield divergent semantic readings.

A second, formally provable barrier is computational complexity. Deciding whether a CA configuration has a particular predecessor configuration (the CA preimage problem) is NP-complete or worse in many rule families [7,8]. As ThoughtBox variants scale with added memory or stochasticity beyond 4 bits, semantic decoding becomes an intractable search rather than a simple lookup. This recasts interpretability as a complexity-theoretic constraint, not merely an epistemic gap.

The observation that many data strings cannot be compressed into short, unambiguous descriptions is a central result of algorithmic information theory (Kolmogorov complexity) and related work [9]. ThoughtBox does not aim to prove new incompressibility theorems. Rather, it provides a small, fully specified model that illustrates how the practical tasks of semantic labeling and decoding interact with coarse-graining and resource limits. In short, ThoughtBox complements the theoretical results of algorithmic information theory by showing how decoder design and noise produce diagnosable semantic failures in an explicit dynamical system.

If viewed as a reducible island within Wolfram's Ruliad—the entangled space of all rule-based computable evolutions [10]—ThoughtBox highlights why semantic opacity persists even under ideal measurement and transparency. By isolating a tractable subspace where every rule, state and label is known, the structural and computational limits of semantic inference in any system can be exposed.

This paper pursues two distinct but related research goals and keeps them separate throughout. Goal A (semantic labeling) is to analyze how a decoder (f) that maps microstates or short trajectories to macro-concepts performs under noise, partial observability and resource constraints. Goal B (microstate reconstruction) is to analyze the formal difficulty of recovering exact low-level configurations or predecessor trajectories from observed data and labels. ThoughtBox is used primarily as a controlled testbed for Goal A (showing how imposed semantics interact with dynamics), while complexity theory is invoked to illuminate aspects of Goal B in specific formal settings. This separation is emphasized because

conflating the two leads to category mistakes: many-to-one mappings are appropriate for Goal A but make Goal B impossible by design.

The remainder of this paper proceeds as follows. Section 2 reviews several aspects of neural reductionism, state-to-semantics mappings, CA interpretability and neuronal imaging in simple organisms. It then formalizes the ThoughtBox model and its semantic decoder, notes the intractability results for scaled CA variants and situates these findings within the Ruliad framework. This is then followed by a Discussion and Conclusion regarding implications for neuroscience, AI interpretability and the broader limits of semantic access.

2. Neural Reductionism

Neural reductionism maintains that every aspect of mind—from basic sensations to abstract reasoning—supervenes on biophysical interactions. Its intellectual pedigree stretches from La Mettrie's L'Homme Machine [11] through the neuron doctrine of Golgi and Cajal (e.g., Shepherd [12]) to mid-20th-century arguments by Place [13], Smart [14], and Armstrong [15]. Churchland [16] reinforced this lineage, proposing that a unified mechanistic framework in principle suffices to explain cognition, while Craver and Bechtel [17] showed that top-down causal accounts in neuroscience remain compatible with purely physical bases.

Yet two assumptions go largely unexamined. First, this reductionism presumes that once there is exhaustive data—membrane potentials, synaptic weights and connectivity graphs—decoding mental content is predominantly an engineering challenge. In reality, one needs a decoder architecture matched to the brain's internal representational scheme, a nontrivial and often tacit construct. Second, interpreting these high-dimensional, time-varying datasets entails formidable computational demands: as mentioned, even a "Laplacean demon" with complete microstate data would face intractable search and resource limits [6,18].

Moreover, emergent functions exhibit multiple realizability, undermining any unique microstate $\rightarrow$ mental-state mapping [19,20]. Combined with NP-complete barriers to precursor problems in dynamical systems [7,8], these factors expose a critical gap: substrate transparency does not guarantee semantic transparency.

3. State-to-Semantics Decoding

Significant progress has been made in mapping physiological measurements to high-level semantics, but these efforts rest on untested tractability assumptions. Mitchell et al. [21] used linear regression to predict fMRI voxel patterns for 60 concrete nouns, and Pereira et al. [22] extended this to a cross-subject "universal decoder" for distributed semantic spaces. More recently, Wairagkar et al. [23] demonstrated real-time speech synthesis for an ALS patient via intracortical arrays, achieving ~60% intelligibility. Although these milestones highlight practical advances, they rely on polynomial-time classifiers and regression without probing worst-case complexity bounds.

At finer scales, optical voltage imaging in simple organisms illustrates similar constraints. Gong et al. [2] recorded entire C. elegans voltage dynamics and Storace and Cohen [3] mapped Drosophila brain activity, but translating voltage traces into behavior or cognition still demands detailed circuit models and contextual priors. In each case, the implicit assumption is that rich data + standard learning algorithms suffice for semantic decoding.

The unaddressed question is the following: When do data volume and algorithmic complexity merge to render semantic inference infeasible, irrespective of measurement fidelity? ThoughtBox targets this question head-on by proving that—even in a fully

transparent, minimal CA—the decoding problem escalates from trivial to NP-complete as system scale or noise increases. This reframes interpretability as a complexity-theoretic and resource-bounded constraint, rather than a mere matter of data acquisition.

4. Cellular Automata Complexity

Cellular automata (CA) are discrete dynamical systems in which simple local rules yield rich global behavior. Wolfram's four-class taxonomy—fixed-point and periodic patterns (Class I–II), chaotic dynamics (Class III) and complex, computationally universal structures (Class IV)—frames much of CA theory [24,25]. Central to this view is computational irreducibility: certain CA require step-by-step simulation, with no shortcut to predict their long-term evolution.

Chaitin's notion of algorithmic irreducibility complements Wolfram's dynamical irreducibility [26,27]. While Wolfram focuses on the unpredictability of CA evolution, Chaitin formalizes the idea that certain outputs (e.g., CA trajectories) cannot be compressed into shorter descriptions—i.e., their shortest program is the output itself. The semantic decoding problem can be framed as a search for a compressed interpretive model. If the mapping from microstates to meaning is irreducible in Chaitin's sense, then no decoder shorter than the full trajectory exists and thus semantic decoding may be formally uncompressible.

Complexity theorists formalize this irreducibility via decision problems. The preimage problem, as previously mentioned—deciding whether a configuration has any predecessor—is co-NP-complete for many two-dimensional CA [7] and NP-complete even in one dimension [8]. Further results place related inversion tasks in PSPACE-complete or EXPTIME-complete classes, revealing an ascending hierarchy of intractability.

Despite refinements incorporating entropy, spectral methods and compressibility metrics [28–32], CA research remains focused on dynamical classification rather than semantic interpretation. Explicit semantic mappings—assigning states or trajectories to human-meaningful symbols—are exceptionally rare. Notable proto-semantic efforts include the following:

- Jin & Chen [33]: symbolic dynamics for characterizing glider interactions in CA.
- Gautam [34] asynchronous CA for clustering, using state groups as implicit category labels.
- Reaction–diffusion and morphogenesis models: mapping cell states to concentrations or cell types, with semantics imposed externally.

ThoughtBox differs by embedding a fully specified semantic overlay onto a minimal CA: four binary cells whose microstates map directly to simple labels (e.g., "Curiosity," "Doubt"). This design isolates semantic ambiguity and decoding complexity as structural properties—shifting the question from "What patterns emerge?" to "Can those patterns be unambiguously interpreted?"

Tractable Semantic Extraction in Practice

Beyond CA, many semantic-decoding applications rely on polynomial-time algorithms—support vector machines, regression and neural networks. Common examples include the following:

- Semantic Role Labeling (SRL): Collobert et al. [35] and ASSERT (Pradhan et al. [36]) classify argument roles (ARG0, ARG1) using structured prediction in polynomial time.
- Brain decoding from fMRI: Mitchell et al. and Pereira et al. [21,22] employ linear/regression models to map voxel patterns to word embeddings, achieving closed-form inference.
- Document metadata extraction: Nayaka & Ranjan [37] use deep neural nets (DNNs) and SVMs to tag tokens with bibliographic and methodological labels in near-real time.

- BCI cursor and speech decoding: Singer et al. [38] combine Kalman filters and DNNs for real-time neural control, with polynomial-time inference.
- Visual scene → narrative extraction: Transformer-based encoder–decoders generate captions in polynomial time, though relying on exponential-scale pretraining.

Despite practical successes, few studies have assessed worst-case complexity. The CA preimage problem's NP-completeness suggests that semantic inversion may become intractable in general settings. ThoughtBox's reducible island highlights where polynomial-time decoders succeed—and where they provably fail—within computational rule-space.

Table 1 Representative decoding tasks, algorithmic assumptions, observed complexity and limitations. Complexity entries indicate typical practical behavior or known worst-case results where applicable.

Table 1. Examples of tractable semantic extraction tasks and their algorithmic characteristics.

Domain/Task	Model Type/Algorithmic	Assumptions/Observed Complexity	Caveats
Semantic Role Labeling (SRL)	SVMs/neural networks (Collobert et al. [35])	Structured prediction with engineered or learned features/Polynomial-time training and inference	Limited transfer to novel linguistic forms
Brain activation → word prediction	Linear/ridge regression	Fixed voxel-to-embedding mapping, Closed-form inference (polynomial)	Small vocabularies, requires pre-aligned semantic space
Document metadata extraction	DNNs; SVMs	Sequence tagging with token embeddings Fast inference/polynomial training	Sensitive to input format, template-driven labels
BCI cursor/speech decoding	Kalman filters, DNNs (Singer et al. [38])	Neural signals → kinematics or phonemes Real-time decoding; polynomial inference	Requires extensive calibration, semantics extrinsic to the model
Visual scene → narrative extraction	Transformer-based encoder–decoders	Supervised learning from image–caption pairs Polynomial inference	exponential pretraining hallucinations, stylistic priors dominate

→ denote a process of going from one concept to another.

While impressive, these practical decoding efforts often succeed by operating within constrained semantic domains or by leveraging extensive, task-specific training data, which inherently limits their generalizability to the broader challenge of open-ended semantic meaning extraction.

5. The Ruliad and Multicomputation

Stephen Wolfram's Ruliad represents the limit object of all possible rule-based computations, generated by exploring every rule in parallel via multicomputation [10]. In this universal web of causal trajectories, most evolutions are computationally irreducible: no shortcut exists to predict or interpret their outcomes without exhaustive simulation. Although the Ruliad formalism addresses causal invariance and observer-relative coarse grainings, it provides no mechanism to locate semantic accessibility—regions where decoding symbolic meaning from microstates becomes tractable. The concept of reducible subspaces—localized pockets in rule space supporting external mappings, compression symmetries or inference priors—remains unformalized.

ThoughtBox instantiates one such pocket: a minimal cellular automaton (CA) rule space paired with a fully specified semantic map. By drilling into this simple, transparent

system, the structural challenges of semantic inference are readily exposed—challenges not yet addressed by traditional CA classifications nor by current Ruliad theory [25,30,31].

6. The ThoughtBox Model

To investigate the relationship between microstate transparency, semantic inference and computational tractability, a deliberately minimal cellular automaton (CA)-inspired system tailored for conceptual clarity is used, called ThoughtBox, as noted. ThoughtBox consists of four binary elements (E_1–E_4), each assuming values of 0 or 1, yielding 16 possible global configurations. At each discrete timestep, the system updates synchronously using a parity rule: each element flips if the sum of its two immediate neighbors is odd. This simple yet nontrivial mechanism generates structured trajectories through state space.

The parity update was chosen for several reasons. First, it ensures deterministic and local evolution, preserving analytical tractability while allowing complex behavior. Second, prior work—such as Betel et al. [39]—has demonstrated that parity-based CA can solve global pattern-recognition tasks and exhibit meaningful, cyclic dynamics despite limited local interactions. Third, in contrast to chaotic rules like Wolfram's Rule 30 or universal computation via Rule 110, parity CA strike a balance between predictability and semantic expressiveness—making them ideal for evaluating decoding performance across increasing system scale, noise or partial observation[1].

ThoughtBox differs from conventional CA in one key respect: it embeds a ground-truth semantic map, hand-crafted to associate select global configurations with symbolic concepts such as "Curiosity," "Doubt" or "Resolution." These labels are stable and context-independent, mirroring biological systems in which internally grounded meanings may develop through consistent experience or evolution. Critically, *meaning* in ThoughtBox arises not just from isolated states, but from *transitions between states*: trajectory sequences encode conceptual evolution over time (e.g., "Curiosity → Assertion → Doubt").

A brief note on labels and examples is warranted. The semantic labels used in ThoughtBox (e.g., Inquiry, Analysis, Resolution) are designer-imposed, illustrative tags chosen to make the model intelligible. They are not claimed to "emerge" from the parity dynamics in this minimal system. For clarity, the main text uses a single representative mapping for examples and figures; when an example intentionally departs from that mapping this is made explicit in the surrounding text. This convention keeps examples readable while making clear that labels are modeling choices rather than empirical discoveries.

Because its update rule, semantic mappings and decoder structure are fully transparent, ThoughtBox acts as a reducible island within Wolfram's Ruliad—a localized subspace where rule, state and interpretation align. This intentional clarity enables detailed probing of where and why semantic inference succeeds or fails, especially when scaling up bit count, introducing stochasticity or limiting observational access. This 4-bit system yields 16 global states, all of which are enumerated and assigned ground-truth labels (the full explicit labeling for every 4-bit combination is given in Appendix A)—for example:

4-Bit State True Label
0001 Curiosity
0011 Tension
0110 Resolution
0100 Doubt

A complete decoder pipeline consists of:

- binary state identification
- exact lookup of symbolic labels
- temporal sequencing to extract a narrative

Although every rule, state and label is known, two forms of ambiguity arise when the decoder is partial or under-trained:

- Missing labels: an unseen microstate forces a default or invented label (e.g., "Conflict" for 0011).
- Many-to-one grouping: base states '{0001, 1000}' share "Inquiry," yet an under-trained decoder may misclassify 1000 as "Unknown" or "Reflection" (see Figure 1).

Step	State	Label
0	0001	Inquiry
1	0010	Inquiry
2	0100	Inquiry
3	1000	Inquiry
4	0001	Inquiry

Figure 1. Many-to-one grouping in ThoughtBox. Four distinct single-bit configurations (0001 → 0010 → 0100 → 1000 → 0001) follow the parity update cycle yet are all mapped to the same semantic tag "Inquiry," illustrating how multiple microstates compress into a single concept.

Many-to-one mappings as in Figure 1 are the hallmark of categorization: if the decoder's goal is to produce macro-concepts, collapsing multiple microstates into a single label is often desirable and intentional (multiple realizability). If, instead, the task is to reconstruct microstates from labels, non-injective mappings make that task impossible by design. Many-to-one mapping is here evaluated relative to the chosen aim: it is a feature for semantic abstraction and a limitation for microstate recovery.

- Incomplete Lookup Table

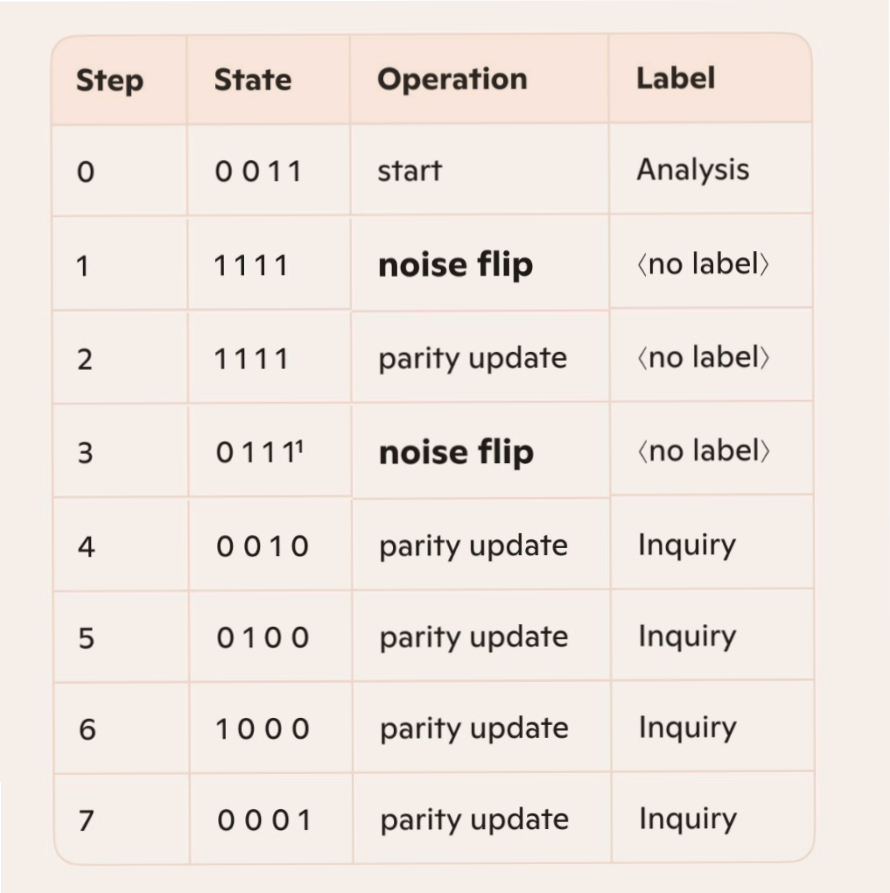
When the decoder outputs "⟨no label⟩" for a state, it is not a mysterious failure—it reflects a concrete gap in what the decoder "knows." In ThoughtBox this could typically occur for one of two reasons:

During the offline training or hand-coding phase, perhaps only a subset of the 16 possible 4-bit states may have been mapped to semantic tags. If the decoder's table never included state '0110', for example, it simply can't return a label for that input. This gap can arise because only the most frequently observed trajectories were sampled, or because memory constraints forced pruning of "rare" states from the table.

- Out-of-Training-Set State

Even with a complete initial map, stochastic perturbations or masking ($\alpha < 1$) can yield a microstate the decoder has never actually encountered. A previously valid state

under pure parity evolution might flip into something like '1101' only after noise—so it lies outside the decoder's experience (see Figure 2). The result is the same: the decoder flags "missing label" because it has no ground-truth mapping for that novel configuration.



Step	State	Operation	Label
0	0 0 1 1	start	Analysis
1	1 1 1 1	**noise flip**	⟨no label⟩
2	1 1 1 1	parity update	⟨no label⟩
3	0 1 1 1[1]	**noise flip**	⟨no label⟩
4	0 0 1 0	parity update	Inquiry
5	0 1 0 0	parity update	Inquiry
6	1 0 0 0	parity update	Inquiry
7	0 0 0 1	parity update	Inquiry

Figure 2. Noise-perturbed trajectory starting from 0011 under the canonical mapping (f_{mathrm{canon}}). Random bit flips (steps labeled noise) move the system off the pure parity orbit into out-of-distribution states (e.g., 1111, 0111) that the decoder flags as "no label," until parity updates return the trajectory to the labeled Inquiry cycle.

A missing-label event in ThoughtBox arises for two non-mystical reasons: (i) the decoder's mapping (f) is partial (out-of-vocabulary), or (ii) stochastic perturbations produce a microstate outside the decoder's training distribution. These are epistemic and design issues, not ontological barriers. Two legitimate responses are available: extend (f) to be total (assigning a label to every state, possibly including a "null" or "undefined" label), or accept a partial semantics and treat missing labels as domain under-determination. The latter stance is adopted to highlight practical decoder limitations, but this choice is explicit rather than presenting missing labels as evidence of metaphysical opacity.

Identifying which of these two causes applies (pruned vs. novel), the analysis converts each "missing" outcome from a black box into a diagnosable, fixable issue. Mapping two distinct microstates to the same concept, additionally, is not an accident in ThoughtBox—it's a deliberate abstraction step. The parity update rule defines a deterministic orbit for any initial microstate under pure CA evolution. In the examples and figures we therefore distinguish two kinds of sequences: (i) pure parity trajectories, which follow the parity rule at every timestep, and (ii) noise-perturbed sequences, where stochastic bit flips temporarily move the system off the parity orbit before subsequent parity updates resume deterministic evolution. All sequences labeled as parity evolution in the figures are pure parity trajec-

tories; sequences labeled as noise-perturbed explicitly include the noise events and are annotated accordingly. This distinction clarifies that missing-label events can arise either because a parity trajectory was never mapped in the decoder or because noise produced an out-of-training state outside the parity orbit.

Figure 2 illustrates a noise-perturbed sequence and its effect on decoding under the canonical mapping (f_{mathrm{canon}}). The sequence begins at 0011 (label: Analysis), receives a stochastic bit flip to 1111 (an out-of-distribution state for the decoder) and remains unlabeled until subsequent parity updates and further noise return the system to the labeled parity cycle 0010 → 0100 → 1000 → 0001 (label: Inquiry). This example is explicitly a noise-perturbed trajectory (not a pure parity orbit) and demonstrates an out-of-vocabulary missing-label event: the decoder's failure here reflects limited training coverage rather than a metaphysical impossibility to assign semantics.

The following explains why two 4-bit patterns (e.g., '0001' and '1000') both become "Inquiry":

- Symmetry in Dynamics

Both '0001' and '1000' are single-active-bit states that, under the parity update, behave identically (they each shift the lone "1" around the ring). Grouping them reflects merely that ThoughtBox's "Inquiry" decoder defines the semantic state depending on the presence of a solitary 1 in the sequence, not on its spatial position. In other words, this is possibly a misapplication of the decoder or an errant decoding conclusion. There may be no way to discern the difference. By collapsing all single-bit states into one label as in this case demonstrates how a higher-level semantic can emerge from multiple microstates. It shows that the decoding isn't one-to-one at all levels—semantics compress many micro-configurations into the same meaning. The key point is that grouping enforces semantic inaccessibility—no decoder can reconstruct the exact bit-pattern once you've merged them under one label.

Either logic—symmetry-based or arbitrary—adds depth. It illustrates that your decoder isn't just "wrong" when it groups; it's following a principled (or illustrative) rule about how low-level microstates map onto high-level concepts. By mapping trajectories (e.g., '0001 → 0011 → 0110 → 0100') into conceptual sequences such as ("Curiosity → Tension → Resolution → Doubt"), ThoughtBox reveals that semantics emerge from transitions, not isolated states, and that even this minimal system can mislead flawed decoders.

7. Complexity Analysis

Semantic decoding in ThoughtBox remains trivial at N = 4, but scaling to N bits increases the state space to 2^N and explodes the number of possible histories. Inferring a predecessor trajectory—the CA preimage problem—is NP-complete or co-NP-complete for many CA rules [7,8], as noted earlier. Introducing noise, memory dependencies or partial observations pushes decoding into PSPACE-complete, EXPTIME-complete or even #P-complete regimes when counting valid trajectories.

Physical resource bounds, as previously mentioned, compound these theoretical hurdles. The maximum number of logical operations performable by our universe is on the order of 10^{120} [6]. Perhaps remarkably, even a ThoughtBox-style system with only N = 400 bits ($2^{400} \sim 10^{120.4}$) requires more operations for brute-force decoding of its global state-space than the universe can supply.

To quantify how many-to-one mappings in ThoughtBox erode semantic transparency, Shannon entropy and mutual-information metrics are used.

Let X be the random variable over microstates and Y = f(X) the induced label.

For any joint distribution:

$$p(x,y),\ H(X) = -\sum x\ p(x)\log_2 p(x),\ H(Y) = -\sum_y\ p(y)\log_2 p(y),$$

$$I(X;Y) = H(X) - H(X \text{ given } Y) = H(Y) - H(Y \text{ given } X).$$

If (Y) is a deterministic function of (X) (i.e., (Y = f(X))), then ($H(Y) \leq H(X)$).

In the special case where (f) is bijective on the support of (X), ($H(Y) = H(X)$) and ($H(X \text{ given } Y) = 0$).

For a many-to-one deterministic mapping,

$$(H(X \text{ given } Y) = H(X) - H(Y)),$$

which quantifies the irrecoverable microstate information lost by coarse-graining.

Under nondeterministic mappings or when noise affects observations, conditional entropies satisfy ($H(X \text{ given } Y) \geq 0$) and ($I(X;Y)$) decreases; in such cases, the decoder's uncertainty about microstates given labels is strictly positive and depends on the joint distribution ($p(x,y)$). These identities are used to compute concrete losses in ThoughtBox examples and to motivate decoder designs that maximize ($I(X;Y)$) under computational constraints.

One can formalize one version of the decoding task relevant to complexity analysis. Input: a description of a dynamical rule (R) (e.g., CA rule), an observation model (O) (which may reveal full or partial microstates at selected timesteps, possibly noisy), a label mapping (f) (possibly partial) and an observed label sequence ($y_1 \ldots, y_T$).

Question (decision version): does there exist a microstate trajectory ($x_1 \ldots, x_T$) consistent with (R) and (O) such that ($f(x_t) = y_t$) for all (t) (or with bounded error under a nondeterministic (f))? A criterion for correctness could be the following: existence of a witness trajectory consistent with the dynamics and observations. Under this formulation, known CA preimage hardness results apply to specific choices of (R) and (O); therefore, formal hardness claims are limited to such well-specified problem instances and avoid universal claims about all decoding tasks.

Complexity-theoretic hardness results are worst-case statements. They show that there exist instances of the decoding problem that are intractable under standard complexity assumptions. They do not imply that typical or average instances are hard. Worst-case intractability (used to motivate caution about universal claims) is therefore distinguished from empirical tractability in many practical settings.

These results recast interpretability as a complexity-theoretic constraint. Possessing a full microstate and a complete decoder does not guarantee semantic transparency when both intractable decision problems and physical limits intervene. One must therefore shift focus from data acquisition to the fundamental question: Can meaning ever be decoded from recorded microstates?

8. Discussion

Our analysis of ThoughtBox highlights two fundamental sources of semantic opacity. First, even in a minimal CA with a known update rule and explicit label map, meaning extraction requires a multi-layer decoder—from bit patterns to symbols to narrative. Second, scaling or perturbing the system rapidly triggers computational barriers: CA preimage problems become NP-complete or worse, and brute-force decoding exceeds cosmic resource limits [6–8]. Together, these pillars show that substrate transparency $\neq$ semantic transparency.

Extending beyond ThoughtBox, this complexity gap manifests universally. In large CA or neural circuits, most evolutions are computationally irreducible—no shortcut avoids step-by-step simulation. A Laplacean demon, even with perfect data, would exhaust the universe's computational budget before decoding a single high-level state. Thus, interpretability is constrained by formal intractability as well as by physical resource bounds.

ThoughtBox occupies a rare reducible island within Wolfram's Ruliad: a fully specified microcosm where rules, states and semantics align to permit tractable decoding. This contrast underscores that semantic accessibility in the Ruliad requires (a) a tractable rule class, (b) an aligned decoder and (c) low combinatorial complexity—conditions met by few real-world systems.

Parallels with Shojaee et al. [40] reinforce this point. They identify three reasoning regimes in LLMs—trivial, chain-of-thought and collapse—despite full prompt visibility. ThoughtBox reveals an analogous semantic landscape: one-to-one decoding, trajectory-dependent ambiguity and intractable inference. Both lines of work converge on a shared insight: representation access alone does not guarantee reliable inference.

Anticipated critiques of the thesis merit a brief response. Practical decoders in neuroimaging or BCI [21,22,38] operate in polynomial time but only recover coarse categories or task-specific commands, relying on massive training data, strong priors and invasive setups. Multiple realizability [19,20] allows clustering of microstates into semantic classes but also amplifies ambiguity: decoders must resolve among many equivalent configurations. Neural plasticity further destabilizes mappings over time [41,42], demanding perpetual recalibration.

Theoretical analogies can deepen our perspective. Gödel's incompleteness, Heisenberg's uncertainty and Matiyasevich's undecidability each signal that certain truths or solutions lie beyond formal derivation—even when all axioms or data are known. While these results arise in different domains, they collectively suggest that some forms of semantic knowledge may be fundamentally unreachable, not merely practically challenging. These analogies, however, risk conflating distinct formal domains. Where limits are relevant, an attempt to rely on precise, domain-appropriate frameworks—algorithmic information theory (Kolmogorov complexity) for compressibility limits and computational complexity theory (e.g., NP-completeness of CA preimage problems) for hardness of inversion—has been made.

Finally, insights from bounded-rationality and ecological rationality [43–45] show how biological agents thrive by operating within tractable interpretive islands, using heuristics tuned to environmental structure. Dennett's intentional stance (1987) [46] exemplifies a heuristic that yields "good-enough" social predictions without reconstructing neural microstates. These frameworks underscore that approximate, context-driven inference is the norm, not an exception.

In sum, semantic interpretability emerges not as an inevitable by-product of data collection, but as a positional property in computational rule-space—requiring aligned decoders, tractable algorithms and adaptive heuristics. ThoughtBox crystallizes these principles, pointing toward a research agenda that maps the topology of approximate decoding islands across artificial and natural intelligence.

9. Conclusions and Future Work

A Reducible-Island Thesis has been articulated: only specially engineered, small-scale systems like ThoughtBox permit tractable, unambiguous semantic decoding despite full substrate transparency. In larger cellular automata, complex neural circuits or the Ruliad at large, semantic decoding confronts NP-complete (or worse) decision problems and physical resource bounds. By isolating a minimal CA with an explicit bit-to-label map, ThoughtBox

demonstrates that semantic non-uniqueness and intractable search arise even under ideal measurements and complete microstate access.

Future work could pursue three directions. First, formalizing the NP-completeness of decoding in scaled ThoughtBox variants and developing a quantitative "semantic irreducibility" metric across CA and neural-network models would be helpful. This metric could quantify the minimum algorithmic complexity required to map microstates to predefined semantic labels, accounting for factors like state-space size, rule complexity and the degree of ambiguity in the state-to-label mapping. Second, empirical studies using programmable CA ensembles or controlled artificial networks could potentially map the gradient from reducible to irreducible regimes under resource constraints. Third, exploring hierarchical decoder architectures—drawing on attention-like mechanisms and inductive biases—to identify design principles for inherently interpretable AI systems may yield advances. These efforts would likely chart the topology of approximate decoding islands in computational rule-space.

The findings here challenge optimistic neural reductionism, which assumes exhaustive measurement suffices for semantic reconstruction [16,17]. Instead, interpretability emerges as a positional property in rule-space—rare enclaves of tractability surrounded by vast regions of semantic opacity. Meaning extraction demands aligned decoders, tractable algorithms and continuous calibration. Full semantic transparency remains a theoretical ideal, constrained by formal and physical limits. Mapping and engineering reducible subspaces will guide both neuroscience and AI toward systems that balance inferential power with interpretability.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. The Complete Canonical Mapping Used Throughout the Manuscript

Table A1. Canonical 4-bit state → label mapping used in ThoughtBox.

State	Canonical Label
0000	**Neutral**
0001	**Inquiry**
0010	**Inquiry**
0011	**Analysis**
0100	**Inquiry**
0101	**Reflection**
0110	**Resolution**

Table A1. *Cont.*

State	Canonical Label
0111	**Tension**
1000	**Inquiry**
1001	**Surprise**
1010	**Conflict**
1011	**Assertion**
1100	**Doubt**
1101	**Confusion**
1110	**Acceptance**
1111	**Unknown**

Note

[1] Note that this parity-rule-based CA is technically equivalent to Wolfram's Rule 150 (new = left + center + right) on a periodic ring. The 4-bit minimal ThoughtBox is cleaner, illustrating the many-to-one and missing-label effects without the complications of CA taxonomy, and the novelty is the layering of the semantic lookup and labeling, not the underlying linear rule.

References

1. Van Strien, M. Onthe origins of Laplacian determinism. *Stud. Hist. Philos. Sci.* **2014**, *45*, 24–431. [CrossRef]
2. Gong, Y.; Wagner, M.J.; Lefort, S. Imaging membrane voltage in *C. elegans*. *Nat. Methods* **2015**, *12*, 53–56.
3. Storace, D.A.; Cohen, L.B. Quantitative optical imaging of neuronal activity in Drosophila melanogaster. *J. Neurophysiol.* **2016**, *115*, 2317–2325.
4. Sporns, O. *Networks of the Brain*; MIT Press: Cambridge, MA, USA, 2010.
5. Seung, H.S. *Connectome: How the Brain's Wiring Makes Us Who We Are*; Houghton Mifflin Harcour: Boston, MA, USA, 2012.
6. Lloyd, S. Ultimate physical limits to computation. *Nature* **2000**, *406*, 1047–1054. [CrossRef]
7. Durand, B.; Formenti, E.; Róka, Z. Cellular automata and NP-completeness. In *Automata on Infinite Words*; Arnold, A., Nivat, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; pp. 65–80.
8. Green, F. NP-complete problems in cellular automata. *Complexity* **2013**, *19*, 41–54.
9. Shen, A.; Uspensky, V.A.; Vereshchagin, N.K. Kolmogorov Complexity and Algorithmic Randomness. In *Mathematical Surveys and Monographs*; American Mathematical Society: Providence, RI, USA, 2017; Volume 220. [CrossRef]
10. Wolfram, S. Exploring Rulial Space: The Case of Turing Machines. *Wolfram Physics Bulletin*. 2020. Available online: https://bulletins.wolframphysics.org/2020/06/exploring-rulial-space-the-case-of-turing-machines/ (accessed on 1 July 2025).
11. de La Mettrie, J.O. *L'Homme Machine*; Durand: Paris, France, 1748.
12. Shepherd, G.M. *Foundations of the Neuron Doctrine*; Oxford University Press: New York, NY, USA, 1991.
13. Place, U.T. Is consciousness a brain process? *Br. J. Psychol.* **1956**, *47*, 44–50. [CrossRef]
14. Smart, J.J.C. Sensations and brain processes. *Philos. Rev.* **1959**, *68*, 141–156. [CrossRef]
15. Armstrong, D.M. *A Materialist Theory of the Mind*; Routledge & Kegan Paul: London, UK, 1968.
16. Churchland, P.S. *Neurophilosophy: Toward a Unified Science of the Mind/Brain*; MIT Press: Cambridge, MA, USA, 1989.
17. Craver, C.F.; Bechtel, W. Top-down causation without top-down causes. *Biol. Philos.* **2006**, *21*, 547–563.18. [CrossRef]
18. Anderson, P.W. More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science* **1972**, *177*, 393–396. [CrossRef] [PubMed]
19. Kim, J. Multiple realization and the metaphysics of reduction. *J. Philos.* **1993**, *90*, 3–10.
20. Putnam, H. Psychological predicates. In *Art, Mind, and Religion*; Capitan, W.H., Merrill, D.D., Eds.; University of Pittsburgh Press: Pittsburgh, PA, USA, 1967; pp. 37–48.
21. Mitchell, T.M.; Shinkareva, S.V.; Carlson, A.; Chang, K.-M.; Malave, V.L.; Mason, R.A.; Just, M.A. Predicting human brain activity associated with the meanings of nouns. *Science* **2008**, *320*, 1191–1195. [CrossRef]
22. Pereira, F.; Mitchell, T.; Botvinick, M. Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **2018**, *9*, 4875. [CrossRef]
23. Wairagkar, M.; Card, N.S.; Singer-Clark, T.; Hou, X.; Iacobacci, C.; Miller, L.M.; Hochberg, L.R.; Brandman, D.M.; Stavisky, S.D. An instantaneous voice-synthesis neuroprosthesis. *Nature* **2025**, *12*, 145–152. [CrossRef] [PubMed]

24. Wolfram, S. Universality and complexity in cellular automata. *Phys. D Nonlinear Phenom.* **1984**, *10*, 1–35. [CrossRef]
25. Wolfram, S. *A New Kind of Science*; Wolfram Media: Champaign, IL, USA, 2002.
26. Chaitin, G.J. *Algorithmic Information Theory*; Cambridge University Press: Cambridge, UK, 1987.
27. Chaitin, G.J. *Meta Math! The Quest for Omega*; Pantheon: New York, NY, USA, 2005.
28. Li, W.; Packard, N.H. The structure of the elementary cellular automata rule space. *Complex Syst.* **1990**, *4*, 281–297.
29. Gutowitz, H.A. Local structure theory for cellular automata. *Physica D* **1987**, *28*, 18–48. [CrossRef]
30. Zenil, H. Compression-based investigation of the dynamical properties of cellular automata and other systems. *Complex Syst.* **2010**, *19*, 1–28. [CrossRef]
31. Martínez, G.J.; Seck-Tuoh-Mora, J.C.; Zenil, H. Wolfram's classification and computation in cellular automata classes III and IV. *J. Cell. Autom.* **2013**, *8*, 233–259.
32. Martínez, G.J. A Note on Elementary Cellular Automata Classification. *J. Cell. Autom.* **2013**, *8*, 233–259.
33. Jin, X.; Chen, H. Symbolic dynamics and semantic pattern extraction in cellular automata. *J. Complex Syst.* **2025**, *34*, 217–241. [CrossRef]
34. Gautam, S. Asynchronous cellular automata for dynamic data clustering. *Pattern Recognit. Lett.* **2025**, *174*, 58–66. [CrossRef]
35. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
36. Pradhan, S.S.; Ward, W.; Hacioglu, K.; Martin, J.H.; Jurafsky, D. Shallow semantic parsing using support vector machines. In Proceedings of the HLT-NAACL 2004: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Boston, MA, USA, 2–7 May 2004; pp. 233–240.
37. Nayaka, P.R.; Ranjan, R. An efficient framework for algorithmic metadata extraction over scholarly documents using deep neural networks. *SN Comput. Sci.* **2023**, *4*, 341. [CrossRef]
38. Singer, A.C.; Clark, A.M.; Cash, S.S. Real-time neural decoding using hybrid Kalman filter–deep learning architectures for speech and motor BCI. *J. Neural Eng.* **2025**, *22*, 016012.
39. Betel, H.; de Oliveira, P.P.B.; Flocchini, P. Solving the parity problem in one-dimensional cellular automata. *Nat. Comput.* **2013**, *12*, 323–337. [CrossRef]
40. Shojaee, P.; Liang, C.; Woods, D.; Kannan, S. *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*; Apple Machine Learning Research Report; Apple Machine Learning Research: San Diego, CA, USA, 2025.
41. Solomon, E.A.; Schapiro, A.C. Plasticity of semantic representations in the human brain. *Trends Cogn. Sci.* **2024**, *28*, 89–102. [CrossRef]
42. Appelbaum, L.G.; Bethune, J.; Tzovara, A.; Cohen, J.D. Representational drift in human cortex during concept learning. *eLife* **2022**, *11*, e76504. [CrossRef]
43. Simon, H.A. *Models of Man: Social and Rational*; Wiley: New York, NY, USA, 1957.
44. Kahneman, D.; Tversky, A. Judgment under uncertainty: Heuristics and biases. *Science* **1974**, *185*, 1124–1131. [CrossRef]
45. Gigerenzer, G.; Todd, P.M. *Simple Heuristics That Make Us Smart*; Oxford University Press: Oxford, UK, 1999.
46. Dennett, D.C. *The Intentional Stance*; MIT Press: Cambridge, MA, USA, 1987.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.