World Models, Structural Constraints, and Deterministic Histories: A Unified Critique of Autoregressive Intelligence

Flyxion

December 2025

1 Introduction

Recent debates concerning the limitations of large language models have converged on a recurring diagnosis. Despite their fluency and breadth of apparent competence, such systems fail in situations requiring stable reasoning across time, intervention, or counterfactual dependence. Yann LeCun has argued that this failure is not incidental but architectural: autoregressive language models cannot achieve human-level intelligence because they lack *world models*, understood as internal systems capable of representing causal structure, predicting the consequences of actions, and supporting planning under constraints.

In parallel, Elliot Murphy's ROSE architecture advances a mechanistic account of linguistic competence grounded not in probabilistic sequence continuation but in invariant structural operations. Syntax, on this view, is enforced by admissibility conditions rather than statistical preference.

This essay argues that these critiques are not merely compatible but formally equivalent when expressed at the correct level of abstraction. Both LeCun's world models and Murphy's structural constraints impose the same functional requirement: the existence of an *authoritative internal history* that stabilizes cognition against statistical drift. Without such a history, systems may generate plausible continuations but cannot reliably distinguish between hypothetical extension and irreversible commitment.

We make this equivalence explicit by introducing a deterministic event-log formalization, which grounds intelligence not in token prediction but in invariant-preserving replay. On this account, autoregressive models fail not because they operate over language, but because they operate entirely within a *view-only loop* disconnected from any authoritative causal order.

2 LeCun's World Model Requirement

LeCun's critique of autoregressive models is architectural rather than empirical. He does not deny that such systems can approximate many human behaviors; instead, he argues that they lack the internal machinery required for reliable, sample-efficient, and grounded intelligence. In particular, they lack mechanisms for representing state transitions under intervention.

At the core of this critique is the concept of a *world model*. Importantly, a world model is not a pixel-level simulator of the environment. Rather, it is a predictive system operating in an abstract latent space that suppresses unpredictable detail, preserves task-relevant invariants, and predicts state transitions conditioned on actions. Planning, in this framework, is optimization over trajectories generated by the world model rather than imitation of past behavior.

Language-only systems, on this account, are disconnected from reality because they never commit to a stable internal notion of state. They predict the next symbol in a sequence rather than the next state of a world.

Definition 1 (World Model)**.** *A world model is an internal predictive system*

$$W : (\sigma_t, a_t) \mapsto \widehat{\sigma}_{t+1}$$

that operates in an abstract representation space, preserves task-relevant invariants, and supports counterfactual evaluation of action sequences.

This definition presupposes a separation between surface-level externalizations, such as text or images, and an internal authoritative state whose evolution is constrained.

3 Autoregression as a View-Only Loop

Autoregressive language models operate by iteratively extending a surface representation according to conditional distributions of the form

$$x_{t+1} \sim P(x_{t+1} \mid x_{\leq t}).$$

While this procedure can produce locally coherent sequences, it lacks any mechanism for enforcing long-range causal consistency. Each step is conditioned on prior outputs rather than on an independently maintained internal state.

This architectural choice collapses the distinction between speculative representation and authoritative update. Generated outputs simultaneously function as proposals and as the sole record of system history.

Proposition 1 (Autoregressive Drift)**.** *In a purely autoregressive system with no invariant-governed internal state, projection error accumulates monotonically with generation depth. As a result, long-horizon coherence cannot be guaranteed even when local predictions are accurate.*

What is commonly described as hallucination is therefore not an anomaly or a narrow failure mode, but a structural consequence of view-only iteration without authoritative replay.

4 Deterministic Event Logs and Authoritative History

To make the world-model requirement precise, we introduce the notion of an *authoritative history*: a deterministic event log that defines the system's internal state independently of any particular external description.

Definition 2 (Authoritative Event Log)**.** *An authoritative event log is a finite ordered sequence*

$$\mathcal{L} = (e_1, e_2, \ldots, e_T)$$

together with a deterministic replay operator

$$\mathrm{Replay}(\mathcal{L}) = \sigma_T,$$

where σ_T is the internal state obtained by applying events in order under fixed transition rules.

External inputs are treated as *views* that may or may not be committed to the log. Commitment is constrained by invariants that define admissible state transitions.

Theorem 1 (Replay-Stabilized Consistency)**.** *If all state updates occur via invariant-preserving replay from an authoritative log, then incoherent or contradictory views cannot corrupt the system's internal state. Errors remain confined to the level of uncommitted views.*

This property directly addresses the failure modes observed in autoregressive systems.

5 ROSE and Structural Causality in Language

Murphy's ROSE architecture provides a domain-specific instantiation of invariant-governed state evolution in the linguistic domain. Rather than modeling syntax as probabilistic sequence regularity, ROSE decomposes linguistic competence into Representation, Operation, Structure, and Encoding.

Merge operations function as constrained state transitions, hierarchical structure defines admissibility conditions, and encoding ensures reliable propagation across neural subsystems. In effect, ROSE specifies a language-domain world model whose authority derives from structural legality rather than statistical likelihood.

Remark 1. *From this perspective, Murphy's so-called Platonic constraints and LeCun's world models are not distinct metaphysical commitments, but alternative realizations of the same architectural requirement: invariant-governed construction of authoritative internal state.*

6 Planning, Objectives, and Safety

LeCun emphasizes that intelligence requires planning: the ability to evaluate sequences of actions with respect to objectives under constraints. This presupposes a predictive model of state transitions, an objective function, and hard constraints that cannot be violated.

In a log-based system, safety constraints are naturally implemented as commit rules. Actions whose predicted consequences violate invariants are not downweighted or filtered post hoc; they are rendered inadmissible. Safety is therefore enforced by impossibility rather than by penalty.

7 Conclusion

The limitations of autoregressive language models do not arise from their use of symbols, language, or large-scale learning. They arise from the absence of an authoritative internal history governed by invariant-preserving replay. Without such a structure, prediction remains a form of unconstrained imitation.

LeCun's world models and Murphy's structural constraints converge on the same conclusion. Intelligence requires systems that can commit to what has happened, reason about what could happen, and refuse what cannot happen. Deterministic event logs provide a minimal formal substrate for this capability.

Artificial general intelligence, on this view, is not a matter of better statistics. It is a matter of building systems that can tell the difference between a view and a fact.

References

[1] Y. LeCun. A path towards autonomous machine intelligence. OpenReview, 2022.

[2] E. Murphy. ROSE: A neurocomputational architecture for syntax. arXiv:2303.08877, 2023.

[3] Meta AI. Joint-Embedding Predictive Architectures for World Modeling. 2024.

[4] G. Buzsáki. *The Brain from Inside Out.* Oxford University Press, 2019.