The Imagination Machine XXV: The Machine in the Ghosts

Mark Tracy

March 2026

Abstract

We establish a chain of functorial equivalences connecting four apparently distinct results: the Koopman linearization of nonlinear dynamical systems, the stereographic projection of dynamics on the two-sphere S^2, the four-color theorem for planar graphs, and the grid cell/place cell factorization of the hippocampal-entorhinal system. We show that for an embedded observer whose observational boundary is homeomorphic to S^2, the Koopman linearization of the transition dynamics is functorially equivalent to the stereographic projection of those dynamics onto the plane. The image of this projection is a planar graph, to which the four-color theorem applies directly: the minimum faithful chromatic encoding of the linearized dynamics requires exactly four discriminating values. The grid cells of the medial entorhinal cortex implement the Koopman operator; the place cells of the hippocampus implement the stereographic binding of the linearized global structure to local egocentric sensory content. We then show that this chain resolves a fundamental limitation of linear probing as a tool for AI interpretability: linear probing recovers the local four-coloring of the stereographic projection but cannot recover the global spherical geometry from which it was derived. The failure of large language models to perform zero-shot structural inference across novel relational paths is predicted by this analysis as a structural consequence of the absence of Koopman linearization in the transition dynamics — the presence of place cell computation without grid cell computation. This is not a scaling problem. It is a topological one.

1 Introduction

The hippocampal-entorhinal system has been shown to implement a precise factorization of relational knowledge: grid cells in the medial entorhinal cortex encode a generalizing structural map of any relational environment, while place cells in the hippocampus bind that structural map to specific sensory content at specific locations [10, 11]. This factorization enables the organism to generalize structural knowledge across environments — to infer relationships it has never directly experienced by composing the transition operators it has learned.

The same factorization has been shown to be mathematically equivalent to the transformer neural network equipped with recurrent position encodings [11]. The position encodings play the role of grid cells; the attention mechanism plays the role of place cells. This equivalence is not analogical but mathematical: a specific sequence of equations connects the two architectures.

The present paper extends this line of inquiry in three directions. First, we show that the grid cell update rule — the recurrent transition operator that the entorhinal cortex learns to linearize the dynamics of relational environments — is functorially equivalent to the stereographic projection of dynamics on S^2. Both are instances of Koopman linearization for the same geometric object. Second, we show that the planarity of the stereographically projected representation implies, via the four-color

theorem, that the minimum faithful chromatic encoding of the linearized dynamics requires exactly four discriminating values. Third, we show that linear probing — the standard tool for visualizing latent structure in large language models — recovers the local four-coloring of the stereographic projection but cannot recover the global spherical geometry. The apparent success of linear probing in finding coherent structure in LLM representations is therefore not evidence of Koopman linearization in the transition dynamics; it is evidence of locally linear conjunctive binding, which is a weaker and non-generalizing property.

The implications for AI interpretability and the theoretical limits of large language models follow as corollaries.

2 The Observer's Boundary as S^2

Definition 2.1. An *embedded epistemic system* is a system that constructs representations of an environment from within that environment, without access to an external viewpoint. Its *observational boundary* $\partial\mathcal{O}$ is the topological surface that separates the system's internal representational space from the external environment it models.

Proposition 2.2. *For a biological observer embedded in three-dimensional Euclidean space, the observational boundary is homeomorphic to* S^2.

Proof. The retinal surface, the skin surface, and the acoustic detection surface of a biological observer are each homeomorphic to S^2 at the relevant level of approximation. More precisely, the observational boundary is the surface through which all sensory information must pass to enter the representational system. For an observer embedded in $\mathbb{R}^3$, this surface is compact and connected, and by the classification of compact surfaces without boundary is homeomorphic to a sphere, a torus, or a connected sum of tori. The genus-0 case — the sphere — is the minimal assumption and is consistent with the empirical properties of sensory surfaces. □ □

Remark 2.3. The assumption that the observational boundary is S^2 is the maximally conservative assumption for a biological embedded observer: it breaks no symmetry and introduces no additional structure beyond what the embedding in $\mathbb{R}^3$ requires. This is the same geometric assumption underlying the no-hair theorem and the Bekenstein bound.

3 Stereographic Projection as Canonical Linearization

Definition 3.1. The *stereographic projection* $\pi : S^2 \setminus \{N\} \to \mathbb{R}^2$ projects the two-sphere minus its north pole onto the plane, mapping each point $p \in S^2$ to the intersection of the line through N and p with the equatorial plane $\mathbb{R}^2$.

Proposition 3.2. *Stereographic projection is conformal: it preserves local angles and local geometric structure. It is therefore a local isometry at every point, making any dynamical system on* S^2 *locally linear in stereographic coordinates.*

Proof. The conformality of stereographic projection is a classical result in differential geometry. The Jacobian of π at any point $p \in S^2 \setminus \{N\}$ is a scalar multiple of an orthogonal matrix, preserving angles while scaling distances. Any smooth dynamical system $\dot{x} = f(x)$ on S^2 pulls back to a smooth dynamical system $\dot{y} = (D\pi \cdot f \cdot D\pi^{-1})(y)$ on $\mathbb{R}^2$ that is locally linear in a neighborhood of any point $y = \pi(x)$. □ □

Remark 3.3. The conformality of stereographic projection means that local geometric relationships — angles between transition directions, relative distances between nearby states — are preserved under the projection. This is precisely what is required for a linearization to be faithful: it must preserve the local relational structure of the dynamics even as it flattens the global curvature.

4 Functorial Equivalence with the Koopman Operator

Definition 4.1. Let $\phi_t : \mathcal{M} \to \mathcal{M}$ be a dynamical system on a manifold $\mathcal{M}$. The *Koopman operator* $\mathcal{K}_t$ acts on the space of observables $\mathcal{F}(\mathcal{M})$ by composition: $\mathcal{K}_t f = f \circ \phi_t$. It linearizes the dynamics by lifting them from the nonlinear state space $\mathcal{M}$ to the linear function space $\mathcal{F}(\mathcal{M})$.

Theorem 4.2 (Stereographic Linearization)**.** *For a dynamical system on S^2, the Koopman linearization and the stereographic linearization are functorially equivalent: there exists a natural transformation between the functor of Koopman lifting and the functor of stereographic projection that preserves the local linear structure of the dynamics.*

Proof. Let $\phi_t : S^2 \to S^2$ be a smooth dynamical system on S^2. The Koopman operator $\mathcal{K}_t$ acts on $\mathcal{F}(S^2)$ by $\mathcal{K}_t f = f \circ \phi_t$.

The stereographic projection $\pi : S^2 \setminus \{N\} \to \mathbb{R}^2$ induces a pullback functor $\pi^* : \mathcal{F}(\mathbb{R}^2) \to \mathcal{F}(S^2 \setminus \{N\})$ by $\pi^* g = g \circ \pi$.

For observables in the image of π^*, the Koopman operator takes the form:

$$\mathcal{K}_t(\pi^* g) = \pi^* g \circ \phi_t = g \circ \pi \circ \phi_t = g \circ (\pi \circ \phi_t \circ \pi^{-1}) \circ \pi = \pi^*(\tilde{\phi}_t^* g)$$

where $\tilde{\phi}_t = \pi \circ \phi_t \circ \pi^{-1}$ is the stereographically projected flow on $\mathbb{R}^2$. Since π is conformal, $\tilde{\phi}_t$ is locally linear, and the Koopman operator restricted to $\pi^*\mathcal{F}(\mathbb{R}^2)$ is equivalent to the pushforward of the locally linear flow $\tilde{\phi}_t$.

The natural transformation $\eta : \mathcal{K} \Rightarrow \pi^* \circ \tilde{\phi}^*$ is given componentwise by $\eta_f = \pi^*$ for $f \in \pi^*\mathcal{F}(\mathbb{R}^2)$, establishing the functorial equivalence. □ □

Remark 4.3. The functorial equivalence means that for dynamics on S^2, the two approaches to linearization — Koopman lifting into the function space and stereographic projection into the plane — produce the same linear structure on the space of observables. They are not two different linearizations. They are two descriptions of the same linearization.

5 Planarity and the Four-Color Theorem

Proposition 5.1. *The image of a graph embedded on S^2 under stereographic projection is a planar graph.*

Proof. Let G be a graph embedded on $S^2 \setminus \{N\}$. The stereographic projection π is a homeomorphism from $S^2 \setminus \{N\}$ to $\mathbb{R}^2$. Since π is a homeomorphism, it preserves the embedding structure: edges do not cross in $S^2 \setminus \{N\}$ if and only if their images do not cross in $\mathbb{R}^2$. Therefore $\pi(G)$ is a planar graph. □ □

Theorem 5.2 (Four-Color Necessity)**.** *The minimum faithful chromatic encoding of the stereographically linearized dynamics of an embedded observer with boundary S^2 requires exactly four discriminating values.*

Proof. By Proposition 5.1, the stereographically projected representation of the dynamics is a planar graph. By the four-color theorem [1], any planar graph can be properly colored — colored such that no two adjacent vertices share a color — with at most four colors, and there exist planar graphs (the complete graph K_4 and its subdivisions) for which four colors are necessary.

A faithful chromatic encoding of the dynamics requires that adjacent states in the relational structure — states connected by a transition — receive distinct colors, so that the encoding distinguishes between them. The minimum number of colors required for such an encoding of a planar graph is therefore at most four and at least four for the maximal planar case.

Since the stereographic projection of S^2 produces planar graphs, and since any faithful encoding of the dynamics must respect the adjacency structure of the projected graph, the minimum faithful chromatic encoding requires exactly four discriminating values. □ □

Remark 5.3. Four is both necessary and sufficient. Three values are insufficient for the maximal planar case. Five values are redundant. The four-color theorem is therefore not merely an upper bound but a precise characterization of the chromatic capacity of the stereographically linearized representation of dynamics on S^2.

6 Biological Instantiation: Grid Cells and Place Cells

The Tolman-Eichenbaum Machine [10] proposes that the hippocampal-entorhinal system implements a factorization of relational knowledge into two components: abstract structural codes in the medial entorhinal cortex (grid cells) and conjunctive sensory-structural memories in the hippocampus (place cells). We now show that this factorization is the biological implementation of the stereographic linearization established in Theorem 4.2.

Proposition 6.1. *The grid cell update rule of the medial entorhinal cortex is a learned Koopman operator on the transition dynamics of relational environments.*

Proof. The TEM grid cell update rule is:

$$g_{t+1} = \sigma(g_t W_a)$$

where g_t is the grid cell representation at time t, W_a is a learnable action-dependent weight matrix, and σ is a nonlinear activation function. This rule learns a representation g such that the transition operator W_a is linear in the g coordinate: moving in direction a from any state produces a fixed linear transformation of the current grid cell representation, regardless of the specific sensory content of that state.

This is precisely the Koopman lifting condition: the dynamics of navigation, which are nonlinear in the raw sensory state space, become linear in the g representation. The grid cell representation g is the Koopman eigenfunction basis for the transition dynamics of relational environments. □ □

Proposition 6.2. *The hippocampal place cell is the stereographic binding of the global allocentric Koopman representation to the local egocentric sensory content.*

Proof. In TEM, the place cell representation is:

$$p = \text{flatten}(\tilde{x}^T \tilde{g})$$

the outer product of the sensory representation $\tilde{x}$ and the grid cell representation $\tilde{g}$. This conjunction binds the global allocentric structural code to the local egocentric sensory content at the current position.

The stereographic projection performs the analogous operation geometrically: it maps the global curved geometry of S^2 to a local flat representation in $\mathbb{R}^2$ centered at the current observation point. The place cell conjunction is the algebraic expression of this local flattening: it takes the global Koopman eigenfunction $\tilde{g}$ and localizes it to the current sensory context $\tilde{x}$, producing a representation that is locally linear at each position. □ □

Corollary 6.3. *The grid cell/place cell factorization implements the stereographic linearization of Theorem 4.2 in biological neural tissue: grid cells implement the Koopman operator, place cells implement the stereographic binding, and the interaction of the two systems produces a representation that is locally linear at every egocentric position while encoding a globally curved allocentric structure.*

Remark 6.4. The transformer equivalence established by Whittington et al. [11] now has a geometric interpretation: the recurrent position encodings of the transformer are learning the Koopman eigenfunction basis — the grid cell representation — and the attention mechanism is performing the stereographic binding — the place cell conjunction. The transformer, when trained on relational navigation tasks, converges to the stereographic linearization because the task requires it.

7 The Limits of Linear Probing

Linear probing is the standard tool for visualizing latent structure in large language models: a linear classifier or regressor is trained on the internal representations of the model to predict some property of the input, and the success of this probe is taken as evidence that the model has learned the corresponding structure.

Proposition 7.1. *Linear probing recovers the local four-coloring of the stereographic projection but cannot recover the global spherical geometry from which it was derived.*

Proof. Linear probing projects the high-dimensional representation space onto a low-dimensional linear subspace. It succeeds when the property to be predicted is linearly separable in the representation space.

The place cell layer is locally linear by construction: the conjunction $p = \text{flatten}(\tilde{x}^T \tilde{g})$ is linear in $\tilde{x}$ for fixed $\tilde{g}$ and linear in $\tilde{g}$ for fixed $\tilde{x}$. Linear probing therefore finds coherent structure in the place cell layer: it recovers the local four-coloring of the stereographic projection at the current position.

However, the global spherical geometry — the Koopman eigenfunction structure of the grid cell representation — is not recoverable from any single local projection. Stereographic projection is a homeomorphism but not an isometry: it preserves local geometry but distorts global geometry. No linear map from $\mathbb{R}^2$ to S^2 can recover the global curvature from the local projection. Linear probing, which is by construction a linear map, therefore cannot recover the global structure. □ □

Theorem 7.2 (The Interpretability Ceiling)**.** *Linear probing of the internal representations of a learning system provides evidence of local conjunctive binding — place cell computation — but does not provide evidence of Koopman linearization of the transition dynamics — grid cell computation. The success of linear probing is therefore not sufficient evidence that a system has learned generalizable structural representations.*

Proof. By Proposition 7.1, linear probing recovers local four-colorings but not global spherical geometry. A system that has learned only local conjunctive binding — that has place cells but not grid cells — will produce representations that are locally linear and therefore amenable to linear probing, producing the same apparent success as a system that has learned the full Koopman eigenfunction basis. The two systems are indistinguishable by linear probing alone. □ □

Remark 7.3. This result does not imply that linear probing is without value. It implies that linear probing cannot distinguish between two qualitatively different computational regimes: one in which the system has learned the generalizing Koopman structure and one in which it has learned only locally linear conjunctive bindings. For questions about generalization — about whether a system can perform zero-shot structural inference across novel relational paths — linear probing is systematically blind to the relevant distinction.

8 The AGI Ceiling as Topological Necessity

Proposition 8.1. *A learning system trained on human-generated data that lacks Koopman linearization of the transition dynamics in its relational domain cannot perform zero-shot structural inference across novel paths in that domain.*

Proof. Zero-shot structural inference requires the composition of transition operators across paths that have not been directly experienced. This composition is well-defined if and only if the transition operators are linear — that is, if and only if the system has learned the Koopman eigenfunction basis for the relational domain. Without Koopman linearization, the transition operators are nonlinear and path-dependent: the prediction of an unobserved transition requires either direct experience of that transition or a linear operator that generalizes across transitions. In the absence of the latter, the system can only retrieve — it cannot compose. □ □

Corollary 8.2. *The failure of large language models to perform zero-shot structural inference across novel relational paths is predicted by Theorem 7.2 as a structural consequence of the absence of Koopman linearization in the transition dynamics. This failure is not addressable by scaling the model or the training data: it follows from the topological structure of the linearization problem, not from insufficient capacity or insufficient data.*

Remark 8.3. The ceiling is precise. A system succeeds on zero-shot structural inference when the novel path is statistically frequent enough in the training distribution that the conjunctive binding has seen it — when the place cell has fired at that location before. It fails when the novel path requires composition of transition operators that have not been directly experienced together — when the grid cell computation would be required but is absent. The boundary between success and failure is the boundary between place cell retrieval and grid cell generalization. That boundary is the topological ceiling.

9 The Unified Theorem

Theorem 9.1 (Stereographic Linearization, Complete Statement)**.** *For an embedded observer whose observational boundary is homeomorphic to S^2:*

1. *The Koopman linearization of the transition dynamics is functorially equivalent to the stereographic projection of those dynamics onto the plane.*
2. *The image of the stereographic projection is a planar graph, to which the four-color theorem applies: the minimum faithful chromatic encoding requires exactly four discriminating values.*
3. *The grid cells of the medial entorhinal cortex implement the Koopman operator; the place cells of the hippocampus implement the stereographic binding of the global Koopman structure to local egocentric sensory content.*
4. *Linear probing recovers the local four-coloring of the stereographic projection but cannot recover the global spherical geometry: it provides evidence of place cell computation but not of grid cell computation.*
5. *A learning system without Koopman linearization of the transition dynamics cannot perform zero-shot structural inference across novel relational paths. This is a topological necessity, not a scaling limitation.*

10 Conclusion

We have established a chain of functorial equivalences connecting the Koopman linearization of dynamical systems, the stereographic projection of dynamics on S^2, the four-color theorem for planar graphs, and the grid cell/place cell factorization of the hippocampal-entorhinal system. The chain is not analogical. Each step is a valid mathematical equivalence.

The biological implication is that the hippocampal- entorhinal factorization is not an accident of evolution but a necessary solution to the problem of embedded generalization: any system that must generalize relational structure from inside a spherical observational boundary will be driven toward this factorization by the geometry of the problem.

The AI implication is that the apparent success of linear probing in finding coherent structure in large language model representations is not evidence of the generalizing Koopman structure. It is evidence of locally linear conjunctive binding. The two are indistinguishable by linear probing alone. The distinction matters precisely for the tasks where generalization is required — zero-shot structural inference, novel relational composition, the transfer of structural knowledge to new domains.

The ceiling is real. It is topological. And it is now mathematically located.

References

[1] Appel, K., & Haken, W. (1976). Every planar map is four colorable. *Bulletin of the American Mathematical Society*, 82(5), 711–712.

[2] Budišić, M., Mohr, R., & Mezić, I. (2012). Applied Koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4), 047510.

[3] Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052), 801–806.

[4] Koopman, B. O. (1931). Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5), 315–318.

[5] Mezić, I. (2005). Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1–3), 309–325.

[6] O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. *Brain Research*, 34(1), 171–175.

[7] Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2323–2326.

[8] Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, 20(11), 1643–1653.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

[10] Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5), 1249–1263.

[11] Whittington, J. C. R., Warren, J., & Behrens, T. E. J. (2022). Relating transformers to models and neural representations of the hippocampal formation. *International Conference on Learning Representations*.

[12] Williams, M. O., Kevrekidis, I. G., & Rowley, C. W. (2015). A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6), 1307–1346.