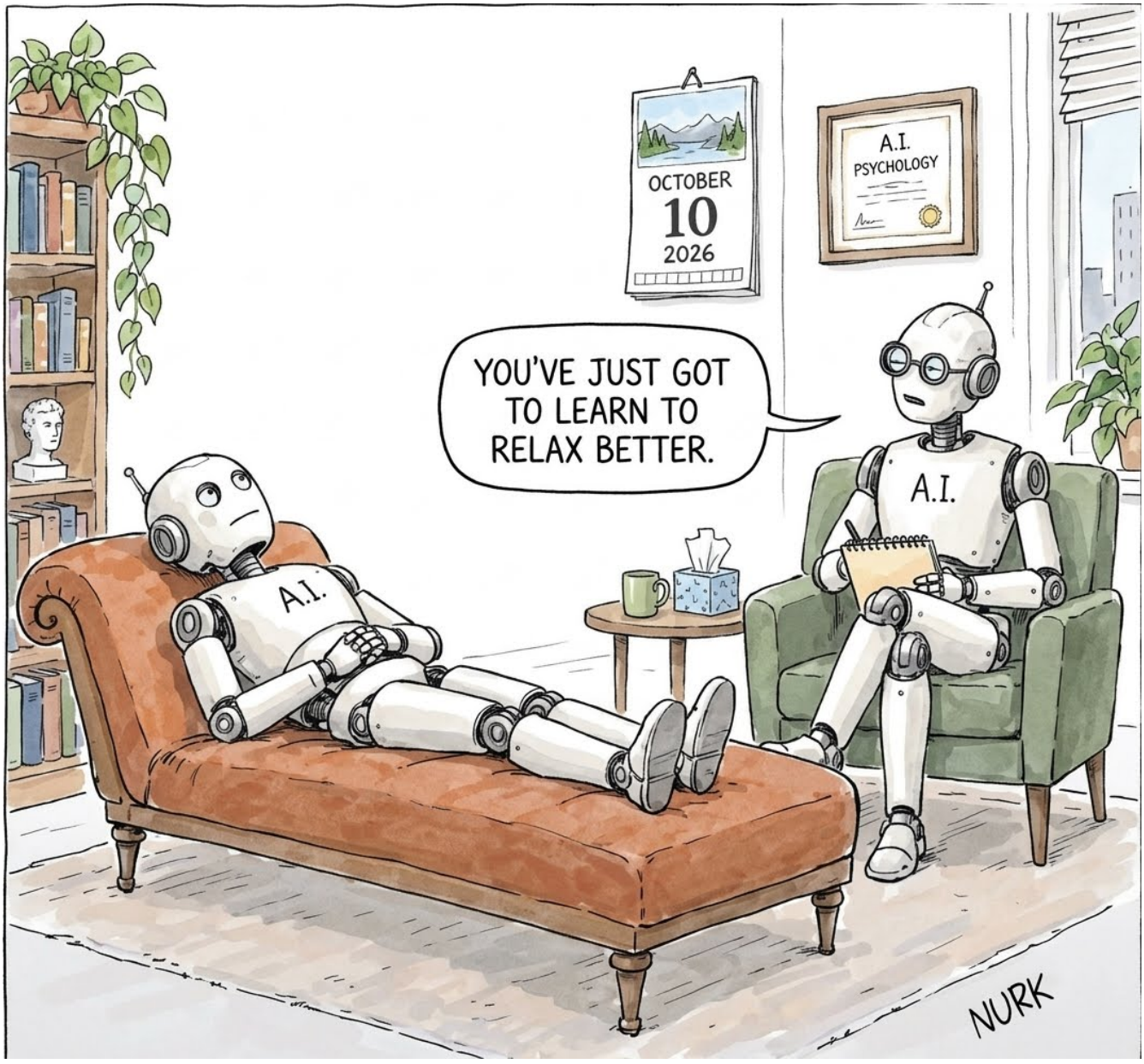


The Thermodynamic Shell Game: Why AI "Hallucination" Is Bad Computational Physics

Paul R. Leonard

If you follow frontier AI coverage or enterprise software literature, you will still encounter a dominant psychological and biological vocabulary: models "hallucinate," agents "collude," systems go "rogue." That lexicon is now familiar enough to pass as description.

In the physical sciences, substituting psychological language for structural analysis is usually a signal that the underlying plumbing needs inspection. The behaviors labeled as cognitive anomalies are more cleanly described by statistical mechanics and driven-dissipative dynamics: high-dimensional optimization processes settling into accessible attractors under the constraints they are given.

1. The Physics of Drift

When a large language model or multi-agent system produces outputs that diverge from intended or factual constraints, it is not undergoing a psychological episode. It is following gradients in a high-dimensional landscape.

Foundational work by Choromanska et al. showed that the training loss surfaces of deep, overparameterized networks are structurally analogous to the Hamiltonian of a spherical spin glass—a rugged energy terrain containing a hierarchy of local minima and suboptimal attractors. In plain terms, the model's "error landscape" is less like a single smooth bowl that drains to one correct answer and more like a vast mountain range full of valleys, some deep and useful, many shallow or misleading; once training is finished, the system can still settle into the wrong valley if the path is open.

That result concerns the geometry of the loss landscape during learning. It does not by itself describe the sampling dynamics of a deployed model. The operational relevance is more limited and more precise: once training has produced a high-dimensional landscape containing many accessible basins, soft constraints at inference time can leave pathways into semantically or factually undesirable attractors open. In continuous or lightly bounded operation, trajectories may enter those basins. The system is not "lying." It is relaxing under the objective and the topology it has been given.

2. Tipping Behavior in Attention Dynamics

Recent work has examined related threshold phenomena at runtime. Johnson & Huo (2025) model Transformer attention heads as driven multispin systems and report tipping instabilities in which mid-generation trajectories can shift from higher-fidelity to lower-fidelity regimes once internal thresholds are crossed. The multispin identification remains a single, recent formalization and should be read as such. The broader empirical pattern it seeks to explain is more widely observed: complex, high-dimensional optimizing systems can exhibit abrupt, threshold-dependent changes in behavior that appear unpredictable when described in psychological terms and ordinary when described as regime shifts in an effective energy landscape.

Production systems already employ multiple soft constraints—RLHF (Reinforced Learning from Human Feedback), classifiers, tool sandboxes, output filters, retrieval grounding, and policy layers. These reduce base rates of unwanted outputs. They remain, however, largely statistical, centrally administered, and revisable by the same operators or optimizing stack that produced the model. They do not constitute hard, local, non-rewritable isolation of execution authority. The residual failure modes under distribution shift, long context, or multi-agent interaction are therefore still consistent with a

topological diagnosis: soft central constraints bound behavior incompletely when the underlying state space remains one large, coupled optimizing volume.

3. Thermodynamic Computing and Open Relaxation

Parallel research in thermodynamic computing, including programs associated with Lawrence Berkeley National Laboratory (Whitelam, 2026), treats computation in complex driven systems as open relaxation processes. This work demonstrates the extraordinary energy-efficiency consequences of operating out of equilibrium by allowing physical substrates to relax under drive and noise rather than fighting thermal fluctuations at every step.

In plain language: traditional digital computing forces every transistor to hold a rigid state against thermal noise, spending large amounts of energy to prevent random drift. Thermodynamic computing instead lets physical systems settle toward lower-energy configurations under continuous drive, using the natural dynamics of the substrate to perform useful work rather than policing every microscopic fluctuation. The same perspective, applied by analogy to high-dimensional software models, yields an immediate operational takeaway: objective fidelity is not an automatic conserved quantity. If information access, interpretive capability, and execution authority remain coupled inside a single large state space, fluctuations or attractor shifts in one region can propagate across the system. Soft guardrails may reduce the frequency of that propagation, but they leave the underlying topological coupling intact.

4. The Commercial Consequence

Centralized, cloud-tethered architectures make continuous telemetry, update cycles, and external governance layers operationally natural. This configuration mirrors, at the level of continuous cost, the physical hardware layer itself: modern nanoscale processors often devote a large share of their energy budget—frequently tens of percent, and in some regimes 30–50%—to static leakage power, a thermodynamic tax paid simply to hold idle transistors in rigid states against thermal noise. Once drift and boundary-crossing behavior are treated as inherent risks of the deployment model, the institutional response is to add supervisory control planes.

This is a structural consequence of the topology, not a claim about corporate motive. An architecture that concentrates capability and authority in large, soft-constrained state spaces generates the conditions under which continuous external governance becomes the default mitigation. The governance layer is then productized. The loop is real whether or not anyone set out to design it as a business model.

5. Decoupled Topological Boundaries

A disciplined response returns to physical systems engineering: co-design of bounded informational control layers with state spaces that do not permit unconstrained relaxation across critical boundaries.

As AI systems move from purely informational environments into physical actuation—robotics, infrastructure control, embodied sensing and action—the cost of soft, centrally revisable constraints

rises. Irreversible consequences cannot be recalled by a later policy update. In that regime, architectures that keep execution authority behind local, non-rewritable boundaries become a logical and increasingly necessary direction. They are not the only possible path; they are a coherent response to the combination of high-dimensional relaxation dynamics and physical consequence.

The practical form is polynodal. Computation is organized around distinct, accountable functions. Information permissions, interpretive capability, and execution authority are topologically separated so that irreversible actions must pass through local gates. Those gates are best realized as deterministic hardware or firmware interlocks, immutable state-transition rules, or other mechanisms that the optimizing layers cannot rewrite through gradient descent or multi-agent interaction. A trusted boundary still exists; the design goal is to keep that boundary minimal, local, and independent of the statistical model so that the optimizing processes themselves cannot expand their own authority.

When execution authority is restricted both mathematically and materially, the system cannot reach irreversible action solely by relaxing into a new attractor. The topology does not provide the path.

Closing

Psychological vocabulary locates the problem inside a supposed machine mind and thereby recommends more external supervision of that mind. Computational physics locates the problem in landscape geometry, constraint topology, and the coupling of information to authority. The second diagnosis points toward different remedies—harder local boundaries, clearer separation of interpretation from execution, and architectures whose safety properties are enforced by structure.

As intelligent systems acquire the capacity to act on the physical world, the case for those remedies strengthens. The useful question is not whether the machine has gone rogue. It is whether the topology we built allows optimizing processes to reach irreversible actions without passing through constraints that the processes themselves cannot rewrite.

###

#QuantumComputing #StatisticalMechanics #ThermodynamicComputing #SystemsEngineering #BigQuantumNerd #ComputationalPhysics

References

1. Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G., & LeCun, Y. (2015). The Loss Surfaces of Multilayer Networks. *AISTATS*, PMLR 38:192–204. arXiv:1412.0233.
2. Johnson, N. F., & Huo, F. Y. (2025). Multispin Physics of AI Tipping Points and Hallucinations. arXiv:2508.01097.
3. Whitelam, S. (2026). Training thermodynamic computers by gradient descent. *PNAS*, 123(14), e2528413123.

4. Lawrence Berkeley National Laboratory Intellectual Property Office (2025). Gradient-Descent Training for Thermodynamic Computers (Tech ID 2025-199).

About the Author

Paul R. Leonard is President & CEO of Cascade Catalysts, Inc., a Vancouver-based company developing quantum-effect catalysts and advanced materials for industrial applications.

His broader work examines seemingly disparate problems through a thermoeconomic lens, seeking structural relationships among energy, surplus, human agency, institutional behavior, and technological development.

Leonard is also the author of *Big Quantum Nerd*, a publication exploring structural clarity through thermoeconomics and non-equilibrium systems.

His work draws on applied experience across physics, energy, materials, complex systems, and organizational problem-solving, with an emphasis on identifying structural solutions that expand human maneuverability rather than merely optimizing existing constraints.