Causality Is Not a Label

Why Statistical Intelligence Cannot Understand Reality

Author:
Alexander Wolfgang Strüver

Location:
Thailand

Date:
2026

License:
Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International
(CC BY-NC-ND 4.0)

Abstract

Most contemporary AI systems are optimized for plausibility rather than truth.
They excel at producing fluent, coherent, and statistically consistent outputs — yet remain fundamentally unbound to the causal structure of reality.

This essay argues that the dominant paradigms in modern AI research suffer from a foundational category error: they conflate correlation with causation, plausibility with understanding, and complexity with insight. As a result, systems are increasingly capable of simulating intelligence while remaining ontologically blind.

The text introduces a strict functional distinction between statistical approximation and causal evaluation, outlining the minimal structural conditions required for genuine causality: causal closure, invariant preservation, meaningful operators, and falsifiability without reliance on aggregate performance.

It further demonstrates why governance, ethics, and control mechanisms cannot compensate for epistemic instability, and why improving plausibility without causal grounding increases systemic risk rather than reducing it.

The central claim is deliberately uncompromising:
intelligence is not a matter of performance metrics or narrative coherence, but of maintaining a coherent, falsifiable relationship with the causal structure of the world.

1) Introduction: The Core Misclassification

The current AI debate is not limited by compute. It is limited by a category error.

We are not failing because GPUs are too slow or datasets too small. We are failing because we keep treating different kinds of systems as if they were the same kind of thing. We use words like **intelligence**, **understanding**, **causality**, and **learning** as if they were interchangeable labels that can be earned through scale. They are not. These terms describe fundamentally different relationships between a system and reality.

Today, “intelligence” is often used to mean *fluency*.
“Understanding” is used to mean *coherent output*.
“Learning” is used to mean *parameter updates*.
And “causality” is used to mean *statistical regularity* or *successful prediction*.

That is the misclassification.

A system can be statistically impressive and still be ontologically blind. It can produce answers that look correct, sound correct, and even often are correct—while still lacking any causal relationship to the mechanisms that generate the world it is describing.

This is why modern AI can appear authoritative while being structurally incapable of truth. It does not fail because it is “not smart enough.” It fails because the architecture is not built to bind internal states to external causal structure. It is built to optimize for what *sounds right* under a distribution.

And reality does not negotiate with distributions.

Reality does not care what we call a model.
Reality does not care what a system *intended* to do.
Reality does not care how well something performs in a benchmark.
Reality only cares whether a claim corresponds to causal structure, and whether actions respect constraints that are actually present in the world.

That is the core separation:

A system does not become causal because we name it causal.
Causality is a property of the world, not a property of the narrative.

If we do not keep that distinction sharp, we will keep building systems that get better at plausibility while remaining fundamentally detached from the conditions that make statements true or false in the first place.

---

2) The Status Quo: What Is Currently Marketed as “Progress”

Most of what is presented today as progress in AI is progress in **simulation quality**, not progress in **epistemic grounding**.

We are improving surface competence—fluency, coherence, pattern completion, retrieval—but we are not building systems that can establish causal necessity, detect causal absence, or distinguish “unknown” from “plausible.”

That is not a minor gap. That is the difference between:

- **producing outputs**
  and
- **maintaining a correct relationship with reality**

To see the problem clearly, it helps to break current approaches into three dominant families.

2.1) Statistics as the Foundation (LLMs and Related Systems)

Large language models and similar systems optimize for probability, coherence, and linguistic continuity.

Their objective function rewards:

- tokens that are likely under the training distribution,
- outputs that are internally consistent,
- completions that resemble human text.

This produces a powerful result: fluent, convincing language and impressive generalization within the space of patterns represented in data.

But none of that implies causality.

A language model does not represent the world as a causal system.
It represents the world as a **statistical manifold** of correlations between symbols.

That distinction matters because “truth” and “causal correctness” are not statistical properties of text.

A statement can be highly probable in language while being causally false.
A statement can be rare in language while being causally correct.
And most importantly: a model can be wrong in ways it cannot detect, because nothing in its architecture binds outputs to causal necessity.

In other words:

- It can produce coherent answers without knowing whether the answer is grounded.

- It can produce confident answers without having an internal representation of what would make them wrong.
- It can optimize plausibility while lacking any mechanism for causal accountability.

This is why hallucination is not a bug. It is a structural consequence of the objective.

If the system is not designed to:

- detect missing causal constraints,
- identify unknowns as unknowns,
- and refuse to invent when causal structure is absent,

then it cannot be said to “understand,” even when it performs well.

Decisive declassification:
A system that cannot identify *why* it is wrong is not understanding anything. It is producing well-formed guesses.

2.2) “Learning,” Feedback Loops, and AGI Rhetoric

A second family of claims revolves around “learning” in the sense of:

- online updates,
- reinforcement loops,
- continual training,
- human feedback,
- environment interaction.

The story is usually:
“Now the system learns from consequences. Therefore it becomes grounded.”

But this also collapses categories.

Feedback-based adaptation can reduce error. It can even reduce error dramatically. But it does not automatically generate causal insight, because feedback is not the same as causal reference.

A system can adapt via feedback while still operating purely statistically:

- adjusting weights,
- tuning policies,
- compressing patterns of reward.

That is still calibration.

It is optimization under signals.
Not causal comprehension.

The reason drift remains unavoidable in this class is simple:
if the internal representation is not causally closed, then it has no invariant anchor.
It is continuously updated relative to a moving distribution of inputs and rewards.
Over time, that necessarily produces state instability unless the system has a causal criterion that constrains updates.

So the key question is not:
“Does it learn?”
but:
“What constrains the meaning of learning?”

If the system’s learning rule is not constrained by causal invariants, then it is simply becoming better at the reward landscape, not better at reality.

That is why “AGI” rhetoric often collapses into a marketing claim:
the system gets better at continuing a loop, so people interpret it as “understanding.”

But improvement is not the same as grounding.

Decisive declassification:
Feedback without causal reference is calibration, not learning.

2.3) Geometric, Symbolic, or Hybrid Models

A third family tries to address the shortcomings of pure statistics by adding:

- geometry-aware operators,
- symbolic reasoning layers,
- graphs, heuristics, rules,
- neuro-symbolic hybrids.

These approaches often feel closer to “structure,” and that is a legitimate direction.

But the same category error frequently remains:
structure alone does not imply causality.

You can impose formal structure that remains purely descriptive.
You can add rules that remain arbitrary.
You can build heuristics that remain taxonomic rather than causal.
You can create graphs that link concepts without encoding what makes transitions admissible, necessary, or forbidden in the world.

A formal system becomes causal only if its internal constraints correspond to real constraints.

If the system cannot answer:

- what makes a transition possible,
- what makes it impossible,
- what invariants must be preserved,
- what would falsify the model,

then it is not causal. It is organized description.

This is exactly where terms like “emergence” are frequently abused.

“Emergence” is often treated as an explanation:
“This behavior emerged.”

But that statement usually means:
“We do not know the causal decomposition.”

In a causal framework, emergence is not a final answer. At best, it is a placeholder for a multi-causal mechanism not yet decomposed:
multiple causes,
weighted relations,
interactions across scales,
constraints producing localized outcomes.

If you cannot specify the causal basis for the weighting and relation, you do not have emergence.
You have ignorance with a label.

Decisive declassification:
Emergence is not an explanation. It is an admission that causal structure is missing.

---

3) The Central Category Error: What Causality Is *Not*

The core mistake in modern AI discourse is that **causality is treated as a stylistic label**—something you can approximate by adding more data, more parameters, more compute, or better wording. That is false.

Causality is not a vibe. It is not a performance metric. It is a structural relationship in the world: **what produces what, under which constraints, and which invariants must remain preserved across state transitions.**

To keep the boundary sharp, three confusions must be eliminated.

3.1) Causality ≠ Correlation

Correlation, even perfect correlation, does not produce an explanation.

A correlation only tells you that two variables co-vary under observation. It does not tell you:

- which variable (if any) is causally upstream,
- whether both are effects of a third factor,
- whether the relationship is conditional on hidden constraints,
- whether the pattern will hold under intervention.

Even a 1.00 correlation can occur in purely observational contexts where **no causal link exists**—because both signals are governed by a common hidden driver, or because the measurement process itself imposes coupling.

That is why causality requires something correlation does not provide: **intervention semantics** and **constraint semantics**.

If you cannot state what happens when you intervene on X while holding other relevant constraints constant, you do not have causality—you have pattern alignment.

A second common mistake is to treat temporal ordering as causal ordering.

“X happened before Y” is not “X caused Y.”

Time sequence can be a prerequisite for causation, but it is never sufficient. A sequence can be:

- coincidence,
- shared response to a third trigger,
- a downstream manifestation of a constraint,
- or a measurement artifact.

A causal claim is not validated by chronology.
It is validated by the ability to bind the relationship to necessity: *why this transition must follow under the system’s constraints, and what would prevent it.*

3.2) Causality ≠ Plausibility

Plausibility is a property of narratives. Causality is a property of reality.

A model can be plausible and wrong.
It can be internally consistent and wrong.
It can be rhetorically persuasive and wrong.

And the more plausible a wrong model becomes, the more dangerous it is—because plausibility does not merely fail to guarantee truth; it actively **suppresses detection of error**.

This is the central failure mode of purely probabilistic systems:
they optimize for “what looks right,” not for “what must be right.”

When a system improves its language quality, coherence, and rhetorical smoothness, it does not reduce hallucination by default. It often **amplifies** it, because:

- errors become better camouflaged,
- uncertainty becomes linguistically polished,
- weakly supported candidates become “answer-shaped,”
- and the human reader’s skepticism drops because the output feels authoritative.

Good language is a **multiplier** of perceived correctness.

So if the underlying epistemic layer is not causally grounded, then “better output quality” is not safety. It is **stronger misdirection**.

That is why the real problem is not that these systems sometimes hallucinate.
The real problem is that they can hallucinate *with convincing precision.*

Plausibility is not a proxy for truth.
It is often the opposite: a mechanism by which falsehood becomes hard to detect.

3.3) Causality ≠ Complexity

Complexity is not insight.

Adding parameters, depth, attention heads, memory length, or larger training corpora increases representational capacity. But representational capacity is not causal grounding. It is just the ability to fit and reproduce more patterns.

“More complex” often means:

- more degrees of freedom,
- more room to interpolate,
- more ways to produce an answer that sounds correct.

Without invariants, complexity does not converge toward causality.
It converges toward **better surface mimicry**.

Causality requires invariance because invariants are what separate:

- accidental co-occurrence from necessity,
- local fit from structural truth,
- correlation from mechanism.

If a system does not preserve invariants—if it cannot state what must remain stable across transformations, domains, or interventions—then its complexity is not a strength. It is a larger search space for plausible-looking candidates.

In that case, complexity is just noise with style.

More parameters ≠ more knowledge.
And more "knowledge" ≠ causal understanding.

A complex model that lacks invariance is not approaching reality.
It is approaching a high-dimensional imitation of discourse about reality.

---

4) Why "Better Models" Often Make the Problem Worse

Central claim

Every improvement in plausibility **without** a causally grounded substrate increases systemic risk.

This is the part many people find counterintuitive, because the industry's default assumption is:

> If the model gets better, the system gets safer.

But if "better" means *more plausible outputs* while the epistemic layer remains ungrounded, then better performance does not eliminate the failure mode—it strengthens it.

The failure mode becomes harder to detect, harder to audit, and more expensive to correct.

4.1) Confidence Calibration Error Grows

As systems become more fluent, humans interpret them as more reliable. That shifts user behavior:

- users ask higher-stakes questions,
- they verify less,
- they delegate decisions earlier,
- they treat the system as an authority rather than a tool.

But the model's internal confidence signals are not causal confidence. They are distributional confidence: "I have seen similar continuations before."

So as output quality improves, the gap between:

- *perceived certainty*
  and
- *epistemic justification*

widens.

This is not a minor UX issue. It is a structural amplification of error impact.

The model can become more convincing faster than it becomes more correct—because correctness requires grounding, and grounding is not delivered by scale alone.

4.2) Wrong Candidates Become More Convincing

In an ungrounded system, the space of candidates is not filtered by causal admissibility. It is filtered by plausibility under a learned distribution.

So “better models” do not merely produce fewer errors. They often produce **better-formed errors**:

- more coherent false explanations,
- more persuasive incorrect diagnoses,
- more elegant wrong derivations,
- more confident recommendations that should have been refused.

The danger is not that the system outputs a wrong token.
The danger is that it outputs a wrong *plan*, a wrong *interpretation*, a wrong *decision candidate*—and makes it look structurally sound.

In high-stakes domains, a polished wrong candidate is worse than an obviously weak candidate, because it crosses the trust threshold.

4.3) Errors Are Detected Later—or Not at All

When outputs become persuasive, error detection shifts downstream:

- the system does not refuse,
- the user does not verify,
- the organization integrates it into workflow,
- and only real-world consequences reveal the mistake.

By then, the cost is higher.

And sometimes there is no clean feedback signal at all:

- the decision produces subtle harm,
- the drift accumulates slowly,
- the causal chain is long,
- the damage looks like “bad luck” rather than a model failure.

So the operational reality becomes:

The more plausible the system is, the longer it can remain wrong without being challenged.

That is the opposite of safety.

Leitmotif

A highly plausible false system is more dangerous than an obviously dumb one.

Because the dumb one triggers skepticism.
The plausible one triggers delegation.

And delegation is where error becomes irreversible.

---

5) What Real Causality Requires (Minimal Conditions)

Causality is not achieved by approximation, scale, or policy.
It emerges only when a system is bound to reality in a very specific, structural way.

Without exposing implementation details, the minimal conditions can be stated precisely.

5.1) Causal Closure

A causally grounded system must be **causally closed at the level of evaluation**.

This means:

- no external correction loops that silently modify internal state,
- no post-hoc plausibility filters that overwrite conclusions,
- no hidden reinforcement layers that “nudge” outputs toward acceptability.

Once external correction becomes part of the evaluation process, causality collapses into calibration.

A causally closed system may *receive queries* and *produce evaluations*, but the internal structure that determines those evaluations must remain invariant and self-contained.

If a system's conclusions improve because they are being continuously patched from the outside, then the system is not learning about reality — it is being trained to avoid visible failure.

Causal closure does not mean isolation from input.
It means **isolation from authority, preference, and correction**.

The system may be wrong.
But if it is wrong, it must be wrong **on its own terms**, not because an external layer quietly adjusted the answer.

5.2) Invariants

Causality requires invariants.

An invariant is not a pattern that appears frequently.
It is a structural constraint that must hold across:

- domains,
- scales,
- contexts,
- and transformations.

If a system requires different internal rules for different domains, it is not modeling causality — it is fitting regimes.

Causal understanding is domain-transferable because the underlying constraints of reality do not change when the surface phenomena do.

A model that works in physics but fails in engineering, or works in engineering but fails in biology, is not causally grounded. It is domain-tuned.

Invariants are what allow:

- cross-domain evaluation,
- detection of missing dimensions,
- identification of impossible transitions,
- and recognition of destabilizing trajectories.

Without invariants, a system cannot tell the difference between:

- "this is a new situation"
  and

- “this violates the structure of reality.”

5.3) Operators with Meaning

Transformation is not causation.

Most computational systems apply operators that transform representations:
vectors to vectors, tokens to tokens, states to states.

But a causal operator is not defined by what it *does* to data.
It is defined by what it *represents* in the world.

A causal operator corresponds to a real mechanism:

- force,
- constraint,
- dependency,
- conservation,
- exclusion,
- or necessity.

If an operator cannot be interpreted as “this is how something in the world actually works,” then it is not causal — regardless of how effective it appears computationally.

Meaning here is not semantic decoration.
It is mechanistic grounding.

A causal operator must preserve the relationship between:

- internal state transitions,
- and external state transitions.

Otherwise, the system is performing formal manipulation without ontological commitment.

5.4) Falsifiability Without Statistics

A causally grounded system must be falsifiable **without relying on aggregate performance**.

Statistical models are evaluated by frequency:
accuracy, loss, error rate, benchmarks.

Causal systems are evaluated by contradiction.

If a single admissible counterexample exists — one case where the system's evaluation violates reality — then the causal structure is wrong.

There is no averaging this away.
There is no "mostly correct."

Either the causal constraints hold, or they don't.

This is why causality is expensive to get right and cheap to break.
One valid counterexample is enough.

And this is also why causal systems are rare:
they cannot hide behind probability.

Leitmotif

Causality does not reveal itself by how often a system is right.

It reveals itself by how cleanly, explicitly, and non-negotiably a system can be wrong.

If a system cannot fail in a way that is:

- precise,
- attributable,
- and structurally explainable,

then it does not understand reality.
It merely approximates it.

---

6) Why Governance, Ethics, and Control Cannot Solve This

Once the epistemic layer is wrong, everything built on top of it becomes compensatory.

This is where many well-intentioned efforts fail.

6.1) Governance Limits Behavior — Not Knowledge

Governance can decide:

- what actions are permitted,
- which outputs require review,

- where escalation must occur.

What governance cannot do is:

- generate truth,
- correct ontology,
- or repair a broken model of reality.

If the system's evaluation is wrong, governance can only limit damage — not prevent error generation.

Governance operates downstream of knowledge.
It does not create knowledge.

6.2) Ethics Cannot Compensate for Ontological Blindness

Ethics assumes the system can correctly evaluate:

- consequences,
- tradeoffs,
- risks,
- and outcomes.

If the system's internal model of reality is flawed, ethical constraints are applied to *misunderstood situations*.

This leads to moral theater:
rules applied rigorously to incorrect premises.

Ethics without causal grounding does not produce safety.
It produces **false confidence**.

6.3) Control Without Truth Is Just Managing Uncertainty

When systems are epistemically unstable, control layers proliferate:

- monitoring,
- logging,
- human-in-the-loop,
- rate limits,
- guardrails,
- audits.

These do not fix the root problem.
They manage its symptoms.

This results in permanent crisis management:

- reacting to failures,
- patching edge cases,
- escalating exceptions,
- increasing bureaucracy.

The system never becomes trustworthy.
It merely becomes supervised.

Point of Outclassing

Governance built on probabilistic epistemics is not governance.
It is continuous damage control.

As long as the underlying system does not know what is structurally true,
every control layer is compensating for blindness.

Real governance becomes possible only **after** reality-bound evaluation exists.

Until then, we are not governing intelligent systems.

We are administrating uncertainty.

---

7) The Uncomfortable Consequence

If the previous sections are taken seriously, one conclusion follows that is difficult — but unavoidable.

The overwhelming majority of today's AI research is **not** working on intelligence.

It is working on:

- simulation,
- imitation,
- interpolation,
- optimization under uncertainty.

That work is not illegitimate.
It is often technically impressive.
It has produced useful tools.

But it is not an epistemic technology.

A system that optimizes plausibility, coherence, or task performance without being bound to causal structure does not *understand* anything. It reproduces surface regularities. It predicts what looks right. It imitates the statistical shadow of intelligence.

Calling that “intelligence” is a category error.

Intelligence, in the strict sense, requires the ability to maintain a coherent internal relationship with the causal structure of the world — including the ability to detect when that relationship breaks.

Most current systems cannot do that.
They cannot know when they are wrong in a way that matters.
They cannot explain *why* a conclusion fails.
They cannot bind error to necessity.

So what we are seeing is not progress toward understanding, but progress toward **better simulation of understanding**.

That distinction matters.

Simulation can be scaled.
Imitation can be optimized.
But neither produces insight.

Leitmotif

Selling statistics as causality does not deceive the machine.

It deceives the builder.

---

8) Conclusion: Reality Is Not Negotiable

Reality does not negotiate with narratives.

It does not care about terminology, investor decks, product names, or conference slogans.
It does not respond to intention, confidence, or popularity.

Reality responds to structure.

Category errors are not philosophical inconveniences.
They are engineering failures waiting to happen.

You can smooth over uncertainty with probability.
You can hide ignorance behind fluency.
You can delay failure with governance, oversight, and control layers.

But none of that changes how the world actually works.

Physics does not care whether a system is “AI-powered.”
Biology does not care whether a model is “state of the art.”
Infrastructure does not care whether an output was “high confidence.”

If causality is wrong, the system fails.
Sometimes immediately.
Sometimes quietly.
Sometimes catastrophically.

Causality cannot be approximated by intention.
It cannot be inferred from performance.
It cannot be retrofitted through policy.

If you want causality, you have to build it.

Not name it.
Not market it.
Not simulate it.

Final Statement

You can call systems whatever you want.

The world only responds to cause and effect.

License and Use Statement

© 2026 Alexander Wolfgang Strüver.
This work is licensed under the **Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International License (CC BY-NC-ND 4.0)**.

You are free to:

- **Share** — copy and redistribute the material in any medium or format

Under the following terms:

- **Attribution** — You must give appropriate credit to the author, provide a link to the license, and indicate if changes were made. Attribution must not suggest endorsement.

- **NonCommercial** — You may not use this work for commercial purposes.

- **NoDerivatives** — You may not distribute modified material. This includes adaptations, remixes, abridgements, or derivative theoretical frameworks based on this work.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

This publication deliberately excludes proprietary formalizations, mathematical mappings, and implementation-level representations associated with the described ontological framework. All such formalisms remain protected intellectual property and are not granted for reuse, derivation, or incorporation into other theoretical or applied systems.