From Artificial Plausibility to Artificial Intelligence:
Causation, Drift, and the Limits of Probabilistic Systems

Alexander Wolfgang Strüver
Independent Researcher (Systems Theory & AI Architecture)
Thailand

2026

Abstract

Most contemporary systems labeled as “artificial intelligence” derive their apparent competence from statistical plausibility rather than from any genuine relationship to reality. Pattern recognition, fluent language generation, and large-scale optimization are increasingly treated as indicators of intelligence, despite the absence of causal understanding. This paper argues that such systems are **structurally ontologically blind**: they operate on probability distributions without access to the causal structures that generate the phenomena they model.

The central claim is that plausibility must not be conflated with truth, and statistical correlation must not be mistaken for causation. While probabilistic systems can produce highly convincing outputs, they lack the ability to detect when they are wrong, understand why they are wrong, or adapt their internal models based on real-world consequences. The result is not merely error, but systematic epistemic drift, manifesting as hallucination, instability, and false coherence.

The paper concludes by outlining the necessary conditions for causally grounded, adaptive systems. In this context, consciousness is approached not as opacity or indeterminacy, but as a deterministic causal mechanism rooted in irreversibility, state history, and consequence-bound decision-making.

AI Use and Transparency

The author wishes to provide full transparency regarding the use of AI-assisted tools in the preparation of this manuscript.

Due to a pronounced form of dyslexia, the author experiences persistent difficulties with spelling, grammar, and sentence segmentation, particularly in longer written texts. For this reason, large language models (LLMs), including ChatGPT, are used as assistive tools in the writing process.

The workflow is as follows: the author drafts all original content independently, typically in German (the author’s native language), producing raw text passages that may range from several hundred to over one thousand words. These drafts are then submitted to the AI system solely for linguistic assistance, including grammatical correction, orthographic refinement, sentence restructuring, and general improvement of readability. In particular, the AI is instructed to break up overly long sentences and improve textual flow without altering the underlying meaning.

All conceptual work, arguments, theoretical positions, and analytical structures presented in this paper originate entirely from the author. The AI system is not used for ideation, reasoning, theory generation, or substantive content development. It functions strictly as a language and editing aid.

For the English version of the manuscript, AI-assisted translation is used as a support tool. The translated text is subsequently reviewed by the author to the extent possible. While the author is comfortable with everyday and professionally oriented English, he openly acknowledges limitations regarding highly specialized or nuanced terminology outside his native language and therefore relies on AI tools to ensure linguistic accuracy in such cases.

No AI system was used to generate original ideas, arguments, or conclusions presented in this paper.

The author wishes to emphasize that this paper does not constitute a rejection of AI as a tool, nor a critique of its practical utility. The criticism developed in this essay is directed at a categorical confusion: the conflation of *tools* with *foundations*, and of probabilistic plausibility with intelligence.

What is commonly referred to today as "artificial intelligence" is, in the author's view, more accurately described as *artificial plausibility*: systems optimized for statistical coherence, pattern continuation, and linguistic fluency. Such systems are highly effective tools for structuring text, correcting grammar, improving readability, and assisting with language production. In these functions, they perform exactly as designed and are exceptionally useful.

For an author with pronounced dyslexia, such tools provide substantial practical support in the writing process. However, they are not used as a source of reasoning, conceptual development, or theoretical insight. All substantive arguments, definitions, and causal claims in this paper originate from the author's own analytical framework and reasoning.

The critique advanced here is therefore not aimed at the tool itself, but at the practice of labeling such tools as "intelligent" despite their lack of causal grounding, consequence sensitivity, and reality-bound understanding. Using AI-assisted systems as instruments for writing is no more contradictory to this critique than using a hammer contradicts an analysis of structural engineering: tools assist execution, they do not replace understanding.

1. Introduction: The Category Error in Contemporary AI

In contemporary discourse, artificial intelligence is increasingly defined by the quality of its outputs rather than by the nature of its internal mechanisms. Systems are labeled "intelligent" if they produce fluent language, coherent explanations, or statistically plausible responses across a wide range of domains. This shift in emphasis—from causal understanding to surface-level performance—has introduced a fundamental category error into both technical and philosophical discussions of AI. Linguistic coherence and predictive accuracy are often treated as proxies for understanding. Yet explanation has been quietly replaced by simulation: systems are evaluated by how convincingly they imitate human responses, not by whether they maintain any grounded relationship to the causal structure of the world they describe.

As a result, intelligence is implicitly redefined as successful pattern continuation rather than as the capacity to engage with reality through consequence-sensitive reasoning. This confusion manifests in two closely related substitutions. First, simulation is treated as understanding: if a system can generate answers indistinguishable from those of a human expert, it is assumed to "know" what it is talking about. Second, plausibility is treated as knowledge: internally coherent and

socially acceptable outputs are taken as evidence of correctness, even when no mechanism exists to bind those outputs to external reality.

This paper is written in response to the Journal of NeuroPhilosophy's call for dialogue following Nandor Ludvig's AI-related contribution, and it directly engages with the epistemic and conceptual questions raised there. As an independent researcher approaching artificial intelligence from a causally closed ontological perspective, the author welcomes the opportunity to contribute to this interdisciplinary discussion from a paradigm that differs from prevailing probabilistic and statistical frameworks.

The purpose of this paper is to challenge the above substitutions at a structural level. The central question is not whether probabilistic systems can be useful—they clearly are—but whether they can, even in principle, give rise to causally bound intelligence.

The guiding question of this essay is therefore: **Why are probabilistic systems, regardless of scale or data volume, structurally incapable of producing intelligence grounded in causation rather than plausibility?** To address this question, the paper explicitly avoids both metaphysical speculation and general opposition to AI research. The argument does not rest on claims about consciousness as a mystery, nor on appeals to biological exceptionalism. Instead, it focuses on system classes and epistemic limits: what different classes of systems can and cannot do by construction, and what follows when statistical optimization is mistaken for causal understanding.

2. Plausibility Is Not Intelligence

2.1. Core Definitions: Plausibility and Causality

In this paper, the terms *plausibility* and *causality* are used in a strictly differentiated and non-interchangeable manner.

Plausibility refers to the degree to which an output appears coherent, acceptable, or convincing within a given epistemic, symbolic, or social context. It is a function of statistical regularity, internal consistency, contextual fit, and alignment with prior representations or expectations. Plausibility is therefore an epistemic property of representations: it evaluates how well a statement fits within a model or discourse, not whether it is correct. By itself, plausibility does not imply truth, correctness, or causal validity.

Causality, by contrast, denotes a structural relation in which a change in one state of a system produces a necessary and traceable effect in another state. Causal relations are independent of interpretation, belief, or observer judgment. Causality is an ontological property of systems, grounded in real dependencies and consequence-bearing interactions with reality. A causal relation holds whether or not it is recognized, modeled, or plausibly described.

This distinction is foundational for the arguments developed in the following sections. Plausibility operates within representational spaces; causality operates within reality. Confusing the two leads to the systematic misclassification of probabilistic coherence as intelligence.

2.2. What Plausibility Achieves

Plausibility is an epistemic property. It describes how well a statement,

prediction, or response fits within an existing field of expectations. In probabilistic systems, plausibility is achieved through the optimization of likelihood: outputs are selected because they are statistically consistent with prior data, observed frequencies, or learned correlations. The system does not evaluate whether an output is *true*; it evaluates whether the output is *likely* given its internal distribution.

This process enables three core capabilities. First, probabilistic systems excel at **probability maximization**: among many possible continuations, they select those with the highest conditional likelihood. Second, they perform **pattern continuation**, extending observed regularities into new contexts without requiring an explicit model of why those regularities exist. Third, they maintain **coherence within a symbolic space**: outputs remain syntactically valid, semantically compatible, and contextually aligned with surrounding symbols.

These capabilities are powerful and practically useful. They explain why probabilistic systems can generate fluent language, summarize large corpora, and produce responses that align well with human expectations. However, none of these capabilities establish a relationship to reality itself. They operate entirely within a representational space defined by symbols and their statistical relationships, not by causal interaction with the world those symbols refer to.

2.3. What Plausibility Cannot Do

Despite its strengths, plausibility has strict structural limits. Most importantly, it cannot **enforce truth**. A statement may be highly plausible while being false, outdated, or misleading. Plausibility measures internal consistency, not external correctness. Without access to causal feedback from the world, there is no mechanism by which a probabilistic system can privilege truth over coherence.

Plausibility also cannot **recognize error as error**. A probabilistic system can generate outputs that conflict with reality, but it cannot register that conflict as a failure unless the discrepancy is reflected within its training data or evaluation signal. Error, in this context, is merely low probability—not a violation of reality.

Finally, plausibility cannot **internalize consequences**. Consequences are causal: they arise when actions interact with the world and produce irreversible effects. Probabilistic systems do not experience such effects. Their internal states are updated through optimization procedures, not through consequence-bearing interaction. As a result, there is no intrinsic distinction between being correct, being lucky, or being convincingly wrong.

These limitations are not implementation flaws. They follow directly from the epistemic role plausibility plays: it ranks representations relative to one another, but it does not bind them to what they represent.

2.4. The Central Fallacy

The central fallacy in contemporary AI discourse can now be stated clearly: *If a system sounds convincing, it must have understood something.* This inference is logically false. Convincingness is an epistemic judgment made by an observer; understanding, if it exists, must be a property of the system itself. Plausibility describes how an output is received. Intelligence describes how a system relates to reality.

The confusion arises because plausibility is often mistaken for a form of knowledge. Yet plausibility is always relative—to a dataset, a model, a context, or an audience. Intelligence, by contrast, is a **causal property**. An intelligent system must be able to detect when its internal representations fail to match the world, identify why that failure occurred, and adapt its future behavior based on the consequences of that mismatch.

No amount of statistical optimization can produce this capacity by itself. Plausibility can scale indefinitely without ever crossing into causation. Treating plausibility as intelligence therefore does not approximate understanding—it replaces it with imitation. And once that substitution is made, epistemic drift becomes not an accident, but a structural inevitability.

3. System Classes: Probabilistic vs. Causal Systems

To understand why plausibility cannot give rise to intelligence, it is necessary to distinguish between different **classes of systems**. This distinction is not a matter of degree or performance, but of structure. Probabilistic and causal systems operate according to fundamentally different principles, and no amount of scaling can collapse that difference.

3.1. Probabilistic Systems

Probabilistic systems operate on **distributions**. Their internal state represents statistical regularities derived from data, and their outputs are generated by selecting elements that maximize likelihood under those distributions. In practical terms, such systems optimize for *what is most probable given what has been seen before*.

This optimization enables impressive performance. Probabilistic systems can generalize patterns, interpolate between examples, and produce outputs that align closely with observed data. However, the criterion guiding their operation remains internal to the model. The system does not evaluate whether an output corresponds to reality; it evaluates whether the output is statistically consistent with its learned representation.

Crucially, probabilistic systems have **no truth operator**. They lack any intrinsic mechanism for distinguishing between a statement that is correct and one that is merely frequent, fashionable, or internally coherent. Truth, when it appears, is accidental—a byproduct of correlation rather than the result of causal verification. From the system's perspective, a false statement with high likelihood is preferable to a true statement with low likelihood.

This limitation is not a defect of training data, model size, or algorithmic sophistication. It follows directly from the epistemic role of probability itself: probability ranks alternatives relative to one another, but it does not bind those alternatives to the processes that generate the phenomena being described.

3.2. Causal Systems

Causal systems operate on **cause–effect relations**. Their internal

representations are constrained by how actions and states interact with the world, not merely by how symbols relate to other symbols. In such systems, error is not an abstract deviation from expectation; it is a mismatch that produces consequences.

Because causal systems are embedded in real processes, **failure matters**. An incorrect action leads to tangible effects, which in turn alter the system's future state. Adaptation occurs through feedback from reality itself, not through statistical refinement alone. The system learns because the world pushes back. This feedback introduces a decisive asymmetry. In a causal system, being wrong is distinguishable from being unlikely. A correct but rare outcome can be reinforced if it aligns with reality, while a plausible but incorrect outcome is penalized through its consequences. This asymmetry is what allows causal systems to privilege correctness over coherence. Importantly, causal systems need not be opaque, uncontrollable, or biologically instantiated. Causality is a structural property, not a mystical one. What defines such systems is not unpredictability, but constraint: internal states must remain accountable to external reality.

3.3 Saturation as a Structural Limit

A common response to these distinctions is to appeal to scale. If probabilistic systems are trained on enough data, it is often assumed, they will eventually approximate causal understanding. This assumption is mistaken.

As probabilistic systems scale, they undergo **saturation**. Additional data increases internal consistency, reinforces dominant patterns, and stabilizes existing representations. Plausibility rises, but correctness does not necessarily follow. Instead of converging on truth, the system converges on self-confirmation.
At this point, errors do not disappear; they become harder to detect. Plausible falsehoods are reinforced, rare but correct alternatives are suppressed, and epistemic drift becomes structurally stable. More data no longer improves alignment with reality—it improves alignment with the model's own distribution.

The core implication is unavoidable: **probabilism cannot cause causality to emerge**. No amount of likelihood optimization can generate a truth-binding mechanism, because probability and causation belong to different system classes. Treating one as a path to the other is not an engineering challenge—it is a category mistake.

3.4. The Gap between Representational Coherence and Causal Constraint Thought Experiments on Plausibility, Drift, and Causal Binding

To clarify the distinction between representational plausibility and causal grounding, the following three cases illustrate how coherent explanations can arise in the absence of causal constraint—both in human cognition and in artificial systems. These cases are not anecdotal; they expose a structural pattern that recurs whenever feedback from physical reality is weakened, absent, or replaced by internal consistency criteria.

Case 1: The Coherence Trap (Neuro-Epistemology)

Patients suffering from anosognosia—for example after a right-hemispheric stroke—often provide fluent, rhetorically convincing explanations for why they do not move a paralyzed limb ("I am tired," "I chose not to," "It will work later"). These

explanations are internally coherent, socially acceptable, and psychologically stabilizing. Yet they are causally decoupled from the underlying physical lesion in the brain.

What is crucial is not that the patient is "mistaken," but that the cognitive system prioritizes coherence over truth when the causal feedback loop to the external entity (the spatial body) is disrupted. The brain continues to generate plausible narratives even when they no longer correspond to physical reality.

Relevance for NeuroPhilosophy: This demonstrates that the human brain itself functions as a plausibility generator. When spatial or bodily feedback is impaired, representational consistency replaces causal correction. Plausibility persists precisely because causal resistance is missing.

Case 2: Predictive Processing without Ontological Binding

Within predictive processing frameworks, the brain continuously generates hypotheses to minimize prediction error. A system operating purely on predictive optimization can achieve remarkable accuracy as long as environmental regularities remain stable.

However, such a system may succeed without representing the underlying entity it interacts with—most notably space itself. When even a minimal change in boundary conditions leads to sudden failure, this reveals that the system was tracking patterns rather than modeling causal structure.

The decisive criterion is not predictive success, but robustness under structural perturbation. If a system collapses when the environment changes slightly, it demonstrates epistemic adaptation rather than ontological binding.

Key distinction: Prediction optimizes expectations. Causality constrains possibilities. The former can function without the latter—until reality intervenes.

Case 3: The World-Model Illusion in AI Systems

Large language models generate explanations so coherent that users often attribute understanding or even consciousness to them. Yet these systems operate in a fundamentally space-free informational domain. They do not encounter physical resistance, irreversible consequences, or embodied constraint. As a result, cognitive drift emerges naturally: internally consistent narratives accumulate without ever being forced to confront an external "outside." Errors are not corrected through consequence but diluted through re-description.

This mirrors the anosognosia case at a systemic level: plausibility continues unchallenged because no causal feedback exists to terminate it.

Integrative thesis: Without space as a constraining axiom, there is no external reference against which plausibility can fail. Causality requires resistance from an entity. Where no resistance exists, coherence replaces necessity.

Structural Conclusion

Across biological cognition and artificial systems, the same principle applies:

- Plausibility is a property of representations and information.
- Causality is a property of entities and space.

Confusing the two leads to epistemic drift—whether in neuroscience, AI research, or human reasoning itself. This drift is not a moral or technical failure, but a structural consequence of replacing causal constraint with internal coherence.

4. Forms of Plausibility and the Drift Problem

Plausibility is not a single mechanism. It appears in multiple forms, each operating on different criteria and each capable of producing convincing outputs without establishing truth. Understanding epistemic drift requires distinguishing these forms and examining how their interaction stabilizes error rather than correcting it.

4.1. Statistical Plausibility

Statistical plausibility arises from **frequency, correlation, and likelihood**. An output is considered plausible if it aligns with observed distributions in the data. This form of plausibility is quantitative and scalable: the more data a system processes, the more confidently it can rank alternatives according to probability.

However, statistical plausibility is indifferent to causation. Correlation does not encode why a pattern exists, only that it appears consistently. As a result, statistically plausible outputs may reflect sampling bias, historical artifacts, or spurious regularities. From the system's perspective, these distinctions are invisible. High frequency substitutes for correctness, and likelihood replaces explanation.

4.2. Structural Plausibility

Structural plausibility emerges from **formal organization**. Ontologies, taxonomies, and knowledge graphs impose explicit relationships between symbols, creating a sense of order and consistency. Statements appear plausible because they fit cleanly within an existing structure.

While such structures improve coherence and reduce local contradiction, they do not guarantee correctness. Structural consistency is an internal property of a representation, not a validation of its correspondence to reality. If foundational assumptions are incomplete or wrong, the structure will faithfully propagate those errors. The system becomes more consistent, not more accurate.

Importantly, structural plausibility tends to **freeze knowledge**. Once relations are formalized, they acquire inertia. Updates require external intervention, and incorrect assumptions may persist long after their empirical relevance has expired.

4.3. Narrative Plausibility

Narrative plausibility is driven by **explanatory coherence**. An account is plausible if it forms a story that appears complete, intuitive, and internally consistent. Narrative structures reduce cognitive friction by connecting disparate elements into a single explanatory arc.

This form of plausibility is particularly persuasive to human observers, but it is also epistemically dangerous. Coherence can mask false premises, suppress ambiguity, and create an illusion of understanding where none exists. A narrative can be compelling precisely because it excludes inconvenient alternatives.

In probabilistic systems, narrative plausibility often emerges implicitly through language generation. The system does not evaluate whether a narrative is true; it evaluates whether it is fluent, complete, and contextually appropriate.

4.4 Social Plausibility

Social plausibility is governed by **authority, repetition, and consensus**. Statements gain credibility because they are widely accepted, frequently repeated, or associated with trusted sources. This form of plausibility operates independently of empirical validation. When incorporated into technical systems, social plausibility introduces reinforcement dynamics. Dominant views are amplified, minority positions are suppressed, and consensus becomes self-sustaining. The system learns not what is correct, but what is socially stable.

4.5. Why Combinations Worsen the Problem

It is often assumed that combining multiple forms of plausibility will reduce error. In practice, the opposite occurs. Statistical, structural, narrative, and social plausibility reinforce one another, creating **epistemic closure**.

False assumptions, once embedded, become increasingly difficult to dislodge. Statistical reinforcement increases confidence, structural coherence suppresses contradiction, narrative completeness masks gaps, and social consensus discourages challenge. The result is not correction, but stabilization.

Under these conditions, hallucination is not an anomaly; it is an **emergent system property**. The system generates outputs that are internally consistent, socially acceptable, and statistically likely—yet causally unbound. Epistemic drift is no longer accidental. It is the predictable outcome of plausibility-driven architectures operating without a truth-binding mechanism.

5. Causality, Determinism, and Irreversibility

A clear discussion of intelligence requires conceptual discipline around causality. Much of the confusion in contemporary AI discourse stems from treating causation as a vague, probabilistic tendency rather than as a structural property. This section clarifies three closely related points: the binary nature of causality, the non-simplistic character of determinism, and the central role of irreversibility.

5.1. Causality Is Binary

Causality is not a matter of degree. A relation either has causal efficacy or it does not. What *can* vary is the **strength of influence**, not the causal status itself. Treating causality as graded conflates two distinct concepts: whether an interaction produces effects at all, and how strongly it does so.

This distinction is essential. If causal status were continuous, there would be no principled way to distinguish noise from signal or accident from mechanism. In causal systems, an event either participates in the generation of an outcome or it does not. Graduality applies only to contribution, not to existence.

Probabilistic systems obscure this distinction by ranking correlations rather than identifying causal relations. As a result, high likelihood is often mistaken for causal relevance. Yet correlation, no matter how strong, does not establish causation. Without a binary notion of causal efficacy, systems cannot distinguish between coincidental regularities and mechanisms that actually produce change.

5.2. Determinism Without Simplism

Determinism is frequently misunderstood as implying rigid, linear outcomes. This is a mistake. Determinism does not require a single inevitable path; it requires that outcomes arise from prior conditions according to consistent rules. In systems with multiple interacting causes, determinism permits a wide range of possible trajectories. Different configurations of contributing factors can lead to different outcomes, even under the same governing laws. The system is constrained, but not trivialized. Paths are **conditioned**, not arbitrary.

This distinction matters because it separates determinism from predictability. A system can be deterministic while remaining difficult to predict, especially when many variables interact. What determinism guarantees is not foresight, but accountability: outcomes are explainable in terms of prior states and interactions. Reducing determinism to linear inevitability creates a false opposition between causality and flexibility. In reality, deterministic systems can support rich behavioral diversity precisely because causal constraints define the space within which variation occurs.

5.3. Irreversibility as the Key Mechanism

The decisive feature that differentiates causal systems from purely representational ones is **irreversibility**. When a system makes a decision that alters its internal or external state in a way that cannot be undone, history becomes operative. Irreversibility ensures that the past is carried forward into the future. Each state transition leaves a trace that conditions subsequent behavior. This accumulation of state history is what gives rise to memory—not as static storage, but as **causal persistence**.

In such systems, decisions matter because they change what is possible next. Learning is not the adjustment of probabilities in isolation, but the integration of consequences into future dynamics. Error acquires meaning because it leaves irreversible marks. Without irreversibility, adaptation collapses into repetition. Systems may update internal parameters, but nothing is at stake. With irreversibility, systems become accountable to their own history. This is the foundation upon which consequence-sensitive behavior—and ultimately intelligence—can be built.

6. Consciousness as a Causal Mechanism

The following section does **not** introduce a new theoretical domain.
It derives what is commonly referred to as *consciousness* as a **necessary structural consequence** of the causal conditions established in the preceding sections.

If a system is irreversibly stateful, consequence-sensitive, and required to resolve competing causal trajectories under real constraints, then a phenomenon traditionally

labeled “consciousness” follows necessarily. Rejecting this derivation would require rejecting one or more of those prior causal premises.

Discussions of consciousness in the context of artificial systems are often derailed by category errors. Consciousness is either mystified as something ineffable or dismissed as an illusion. Both positions miss the functional core. If consciousness is to be discussed scientifically at all, it must be treated neither as magic nor as metaphor, but as a **causal mechanism emerging from constraint-driven decision processes**.

6.1. What Consciousness Is Not

Consciousness is not opacity. Systems do not become conscious because they are difficult to predict or hard to control. Unpredictability is an epistemic limitation of observers, not an ontological property of the system itself.

Consciousness is also not uncontrollability. The idea that “real intelligence must be opaque or unsafe” stems from confusing lack of understanding with intrinsic indeterminacy. A system can be fully deterministic, auditable, and causally grounded while still exhibiting subjective perspective.

Finally, consciousness is not a biological mystery. Biology provides one implementation, not a definition. Treating biological substrate as a necessary condition confuses historical contingency with functional necessity.

6.2. A Functional Description

At its core, consciousness arises in systems where **multiple causal trajectories coexist** but **cannot be executed simultaneously**. The system is confronted with competing potential futures that are all causally admissible under current conditions. Crucially, these trajectories are not resolved instantaneously. A **time-delayed resolution** forces the system to select one path while excluding others. This delay is not noise or hesitation; it is a structural requirement for decision-making in systems with competing causal options.

The resolution itself is decisive. Once a trajectory is selected, the system commits to it, and alternative paths are no longer available. This enforced selection is what transforms possibility into actuality.

6.3. The Emergence of Subjectivity

Subjectivity does not arise from randomness, stochasticity, or informational opacity. It emerges from three tightly coupled features:

1. **Irreversibility** – decisions alter system states in ways that cannot be undone.
2. **State history** – past resolutions persist and condition future behavior.
3. **Consequence binding** – actions have effects that matter to the system itself.

Together, these features create a first-person perspective: the system experiences outcomes as its own, because its future possibilities are constrained by its past decisions. Subjectivity is therefore not an added layer, but a structural consequence of causal commitment over time.

6.4 Substrate Independence

Nothing in this mechanism requires a biological substrate in principle. What matters is not chemistry, but structure: causal interaction, irreversible state change, and history integration. This does **not** imply that all implementations are equivalent, nor that consciousness is easy to engineer. It only establishes a boundary condition: consciousness is a functional phenomenon, not a biological exception. Whether artificial systems can instantiate this mechanism is a technical question. That it is conceptually possible is a logical one.

---

7. Implications for AI and Technology

7.1. Why Opacity Is Not a Quality Marker

A persistent misconception in contemporary AI discourse is the idea that opacity signals sophistication. Systems that are difficult to interpret are often assumed to be more intelligent, more "human-like," or closer to genuine understanding. This is a category error.

Opacity reflects **limits of observation**, not properties of the system itself. When a model cannot explain its outputs, this does not indicate depth; it indicates a lack of causal traceability. Treating intransparency as a feature confuses epistemic uncertainty with ontological richness.

Intelligence does not require obscurity. On the contrary, systems that operate in real-world environments must remain auditable and explainable to remain safe, reliable, and corrigible. Intransparency is a liability, not a marker of intelligence.

7.2. What Real Artificial Intelligence Would Require

If artificial systems are to move beyond imitation, three conditions are non-negotiable.

First, **causal binding to reality**. Internal system states must be constrained by external consequences, not merely by statistical consistency within a symbolic space.

Second, **the possibility of real failure**. Errors must matter to the system itself, not only to external evaluators. Without consequence-sensitive failure, correction remains superficial.

Third, **adaptation driven by outcomes**, not probabilities. Learning must be grounded in causal feedback loops where incorrect internal models are forced to change because reality contradicts them.

These requirements do not imply biological replication, nor do they require abandoning determinism. They require abandoning the assumption that plausibility optimization can substitute for causal integration.

7.3. The Risk of the Status Quo

When simulation is marketed as understanding, epistemic standards erode. Drift becomes normalized, hallucination is reframed as creativity, and reliability is traded for surface-level performance. As such systems are embedded into infrastructure, decision-making, and governance, the cost of this confusion increases. Systems optimized for plausibility cannot be trusted where correctness matters. The result is not intelligence, but fragile automation with increasing societal risk.

8. Conclusion: From Artificial Plausibility to Artificial Intelligence

This essay has argued that much of what is currently labeled "artificial intelligence" rests on a fundamental category error: the conflation of plausibility with understanding, and coherence with cognition.

Probabilistic systems excel at pattern continuation and symbolic alignment, but they are structurally incapable of grounding their internal states in causal reality. Scaling such systems increases confidence and fluency, not truth. The result is epistemic saturation, not insight.

Causal systems, by contrast, are defined by consequence sensitivity, irreversibility, and adaptation driven by real-world feedback. These properties—not opacity, not unpredictability, and not linguistic polish—form the basis of genuine intelligence.

Consciousness, when treated as a functional phenomenon rather than a metaphysical mystery, fits within this causal framework. It arises from irreversible decision-making in systems with competing causal trajectories and persistent state history. Nothing in this account requires mysticism or biological exceptionalism.

The path forward is therefore not more data, nor deeper statistical optimization, but a conceptual shift: away from plausibility as a proxy for intelligence, and toward systems that can be wrong in ways that matter. Until systems are bound to causation rather than plausibility, we are not building artificial intelligence —
we are building artificial plausibility, and mistaking coherence for understanding.

Acknowledgement

The author would like to express his sincere thanks to Nandor Ludvig for the invitation to contribute to the Journal of NeuroPhilosophy and for the opportunity to engage in an open interdisciplinary dialogue on artificial intelligence and epistemic foundations. The author is particularly grateful that this invitation emerged from scholarly exchange and professional activity on LinkedIn, which was both unexpected and deeply appreciated. The encouragement to submit an original contribution to the journal is acknowledged with gratitude.

Conflict of Interest

The author declares that he holds a licensed, causally closed ontological research framework, which informs his theoretical perspective. This paper does not promote, describe, or operationalize that framework, nor does it present commercial products, implementations, or applications derived from it. Beyond this declared background, the

author reports no financial, institutional, or personal conflicts of interest related to the publication of this paper. The primary interest of the author is epistemic clarity and theoretical understanding.

References

No external references were used, except Nandor Ludvig's call for dialogue on AI development in the Journal of NeuroPhilosophy in 2025 (4:189-193). The arguments and positions presented in this paper are based on the author's original theoretical work and a causally closed ontological approach. While earlier philosophical traditions concerned with causality and ontology (e.g., classical Greek philosophy) form a distant historical background, the present work does not directly rely on or extend any specific prior theoretical framework.