The LLM Hallucination Problem and the Human Confabulation Problem Are the Same Problem—And Attempts to 'Solve' It in AI by Suppression Will Fail for the Same Reason Humans Can't Eliminate Imagination

Jonathan West
Independent Researcher
dtdr@multiverse1.com

Abstract

Large language models hallucinate: they produce fluent, coherent outputs that are factually incorrect. Humans confabulate: they reconstruct plausible but false memories under uncertainty. Current approaches treat these as separate problems requiring different solutions—hallucination as an engineering defect to be eliminated, confabulation as a cognitive quirk to be explained. I argue these phenomena are structurally identical, arising from the same computational principle: trajectory-based reconstruction under distributed constraints. Both systems optimise for semantic coherence rather than factual accuracy when constraints are insufficient. This has profound implications for AI safety: attempts to eliminate hallucination by suppression will fail because they attack the same mechanism that enables generalisation, creativity, and flexible inference. The solution is not elimination but management—through constraint density governance, epistemic signalling, and separation of trajectory generation from action commitment. Understanding this parallel transforms how we should approach both artificial and human intelligence.

1 The Parallel No One Is Making

In December 2022, a large language model confidently informed a user that the James Webb Space Telescope had photographed exoplanets for the first time in human history. The statement was fluent, contextually appropriate, and completely false. The first exoplanet imaging occurred in 2004. The model had *hallucinated*.

In a classic psychology experiment, participants watch a video of a car accident. Later, some are asked "How fast were the cars going when they *smashed* into each other?" Others are asked about when they "hit" each other. A week later, those who heard "smashed" are significantly more likely to falsely recall seeing broken glass at the scene, even though none existed. They have *confabulated*.

These errors look strikingly similar:

- Both produce semantically coherent but factually incorrect outputs
- Both arise from partial or ambiguous cues
- Both preserve high-level structure while violating ground truth
- Both feel compelling to the system producing them
- Neither involves random noise—the errors are structured and plausible

Yet we treat them as fundamentally different problems. Hallucination is framed as an engineering defect in machine learning systems—something to be debugged, patched, eliminated. Confabulation is framed as a quirk of human memory—interesting but peripheral to "real" recall.

I argue this separation is a mistake. These are not analogous phenomena; they are *the same phenomenon* arising from identical computational principles. Understanding why requires rethinking what memory and inference actually are.

2 Beyond State-Based Memory

Classical models of memory, whether in neuroscience or artificial intelligence, share a common metaphor: memory as retrieval. A fact is stored; recall consists of accessing it. This metaphor appears in Hopfield networks, where memories are encoded as attractor states in an energy landscape, and in early AI systems, where facts were represented as symbols in a knowledge base.

This retrieval metaphor struggles to explain well-documented properties of both human and machine cognition:

- **Variability**: We never recall an experience identically twice, yet the memory remains recognisable
- **Context sensitivity**: The same cue produces different responses in different contexts
- **Creative recombination**: Novel situations can be navigated without explicit training
- **Structured errors**: When recall fails, it fails *meaningfully*—substitutions respect semantic similarity

An alternative framework, which I term the Distributed-Like Memory Hypothesis (DLMH), offers a different picture. Memory identity is not defined by occupancy of a particular state but by the *process of reconstruction itself.*

Consider a river delta. Classical memory theory says the memory is the point where water meets the sea—a fixed destination. DLMH says the memory is the *geometry of the delta itself*—the channels, banks, and flow patterns that guide water from source to ocean. Flotsam in the water may traverse different paths yet constitute "the same river" because they respect the same constraint geometry.

In neural terms: persistent structures—synaptic connectivity, inhibitory topology, neuromodulatory bias—define a field of constraints that shape the evolution of neural activity over time. Recall consists of traversing a trajectory through this constrained space. Multiple trajectories may satisfy the same constraints and thus instantiate the same memory, even if they differ in precise details.

To make this concrete, imagine releasing several floating balls into the river delta. Each ball may follow a slightly different path on each occasion, yet all are constrained by the same riverbanks and channels, and collectively they converge toward the same outlet. No single trajectory defines the flow; rather, the structure of the delta itself determines the ensemble behaviour. In this sense, a memory corresponds not to a specific neural path, but to the constraint geometry that shapes a family of admissible trajectories.

This reframing dissolves the sharp boundary between memory and inference. Remembering is not retrieving; it is *reconstructing under constraint.*

3 Machine Learning as Convergence, Not Imitation

Modern transformer-based language models exhibit strikingly similar behaviour, despite their very different substrate. They do not store facts in symbolic form, nor do they retrieve sentences

from a database. Instead, they perform high-dimensional similarity alignment followed by sequential generation—a trajectory through latent space guided by learned weight constraints.

What are often called "tokens" in this context are better understood as commitment points: discrete moments where the system resolves ambiguity and constrains future possibilities. Attention mechanisms dynamically reweight constraints based on context, shaping the trajectory as it unfolds. The system optimises for semantic coherence, not truth *per se*.

From the DLMH perspective, this is not a coincidence. Both brains and transformers instantiate the same abstract computation: constraint-guided reconstruction in a high-dimensional space. The similarities are not anatomical or algorithmic, but geometric and dynamical.

Consider the functional equivalence:

Human Memory (DLMH)	**LLM Inference**
Long-term synaptic structure	Weight parameters
Transient neural activity	Token generation
Similarity alignment	Embedding + attention
Trajectory reconstruction	Autoregressive decoding
Semantic coherence	Likelihood maximisation
Confabulation	Hallucination

In both systems:

- Information is encoded in relationships, not isolated components
- Reconstruction proceeds through temporally extended dynamics
- Multiple valid trajectories can satisfy the same constraints
- The system optimises what it can measure locally (coherence) rather than what it cannot (truth)

This is not metaphor. It is structural equivalence at the computational level.

4 Hallucination as Over-Completion

Within this framework, hallucination is no longer mysterious. When a system aligns successfully to a domain but lacks sufficient constraints to determine a unique continuation, it nevertheless proceeds. The result is a coherent completion that satisfies internal constraints but diverges from external reality.

Human confabulation fits this pattern precisely. When asked to recall details from a weakly encoded memory, the brain does not refuse or report uncertainty by default. Instead, it continues reconstruction, filling gaps with semantically compatible content. The result feels like memory because it *is* memory—just not memory of what actually happened.

LLM hallucination works identically. When a prompt aligns to a domain but provides insufficient grounding, the model does not halt. It completes a plausible trajectory based on learned semantic structure. The output feels authoritative because it emerges from the same process that produces accurate responses.

In neither case is the output random. It lies near the intended target in representational space, preserving:

- Semantic category ("space telescope" not "kitchen appliance")
- Structural plausibility (dates, names, relationships make sense)
- Stylistic coherence (appropriate register and format)

What fails is not the reconstruction mechanism but the *constraint density*. The system lacks sufficient grounding to uniquely determine the trajectory.

This explains why hallucinations are so convincing: they are not fabrications from nothing but extrapolations from insufficient data. They are what you would *expect* given partial information—which is exactly what makes them dangerous.

5 Why Suppression Fails

Attempts to eliminate hallucination entirely misunderstand its origin. Suppressing trajectory completion degrades the very capacities that make these systems useful.

Consider what enables:

- **Generalisation**: Responding appropriately to situations not explicitly trained
- **Creativity**: Generating novel but coherent combinations
- **Extrapolation**: Extending patterns beyond observed data
- **Analogy**: Mapping structure across domains
- **Hypothesis generation**: Proposing plausible but unverified explanations

All of these require the system to complete trajectories that are *underdetermined* by available constraints. The same mechanism that produces hallucination when constraints are weak produces insight when constraints are sufficient.

Humans demonstrate this trade-off clearly. Consider two extremes:

Over-suppression (excessive constraint):

- Rigidity in thinking
- Inability to generate hypotheses
- Literalism, concreteness
- Reduced creativity
- Anxiety when facing ambiguity

Under-constraint (insufficient grounding):

- Confabulation, false memories
- Delusional beliefs
- Loss of reality anchoring
- Ungrounded speculation
- Psychotic symptoms in extreme cases

Healthy cognition lives in a managed middle ground. We allow imaginative reconstruction when stakes are low and demand verification when stakes are high. We signal uncertainty, seek external grounding, and gate action based on confidence.

Current AI suppression strategies—aggressive filtering, refusal training, output constraints—push systems toward the over-suppressed regime. The result is not safer systems but:

- Reduced capability on legitimate tasks

- Evasive or unhelpful responses
- Inability to engage with hypotheticals
- Brittle behaviour outside narrow domains

More fundamentally, suppression treats the symptom rather than the cause. The system still lacks awareness of when constraints are insufficient—it simply refuses to generate at all rather than signalling uncertainty while generating.

6 The Solution: Management, Not Elimination

Humans do not eliminate confabulation; they manage it. This involves several complementary mechanisms:

6.1 Meta-Confidence Tagging

We signal epistemic status:

- "I might be misremembering..."
- "I'm not sure, but..."
- "This is just a guess..."
- "I'm certain that..."

This is not accuracy—it is *metacognitive awareness* of constraint density.

6.2 Contextual Gating

We adjust tolerance based on domain:

High tolerance	**Low tolerance**
Storytelling	Surgery
Brainstorming	Legal testimony
Creative writing	Financial advice
Hypothesis generation	Safety-critical decisions

The same cognitive machinery operates in all contexts—what changes is the threshold for action commitment.

6.3 External Grounding

We escalate to constraint addition:

- Checking references
- Asking others
- Consulting records
- Running experiments
- Seeking consensus

This adds constraints without suppressing generation.

6.4 Separation of Generation from Commitment

Most importantly: we distinguish between *considering* possibilities and *acting* on them. Imagination is cheap and abundant. Commitment is rare and costly.

AI systems currently blur this boundary. A hallucinated claim can be acted upon as readily as a verified fact.

The design principle that emerges is clear:

> *No trajectory should be allowed to drive irreversible action unless its constraint density exceeds a domain-specific threshold.*

This is not about correctness—it is about *constraint sufficiency*.

7 Implications for Agentic AI Safety

This framework has profound implications for AI systems with agency—those that can take actions affecting the world.

Hallucination in a question-answering system is an epistemic problem. Hallucination in an agentic system is a *commitment* problem. The danger is not that the system generates implausible trajectories, but that underconstrained trajectories are allowed to complete into action.

Current AI safety approaches implicitly assume either:

- **Model A**: Errors are bugs → Fix hallucination, increase correctness
- **Model B**: Values are misaligned → Add rules, constraints, reward shaping

Both miss something fundamental. In trajectory-based systems, dangerous behaviour arises not from "bad goals" or "wrong facts" but from *overconfident continuation under insufficient constraint*.

DLMH suggests a third approach: **epistemic governance of trajectories**.

This means:

1. **Allow free generation** of trajectories (including speculative ones)
2. **Tag trajectories** with constraint metrics (uncertainty, grounding, agreement)
3. **Separate cognition from action** (no direct path from "thought" to "act")
4. **Require escalation** for commitment (more checks as stakes rise)
5. **Expose uncertainty explicitly** (to users, other systems, oversight)

This architecture mirrors human metacognition:

- We imagine many futures
- We commit to very few
- We check more carefully when stakes are high
- We signal when we're uncertain
- We defer when we lack expertise

Importantly, this approach does not require perfect world models or solved alignment. It requires only:

- Detection of low constraint density (already partially possible)
- Graduated commitment thresholds (engineering, not philosophy)
- Epistemic signalling mechanisms (transparency, not deception)

8 What This Reframing Buys Us

Treating hallucination and confabulation as the same phenomenon transforms our understanding of intelligence itself.

8.1 A Unified Vocabulary

We can discuss human memory errors and AI failures using the same concepts: constraint density, trajectory completion, epistemic governance. This enables genuine cross-fertilisation between cognitive science and AI safety research.

8.2 Realistic Expectations

We stop expecting either humans or AI to be "error-free" and instead ask: how well does the system *manage* its inevitable uncertainty?

8.3 Better Design Principles

Rather than increasingly aggressive output filtering, we can build systems that:

- Generate freely but commit carefully
- Signal uncertainty rather than refusing to engage
- Escalate verification based on stakes
- Maintain epistemic humility

8.4 A Calmer View of AI

LLM hallucinations become less mysterious and less terrifying. They are not evidence of alien intelligence or fundamental danger, but natural consequences of inference under uncertainty—the same consequences humans navigate daily.

9 What This Does Not Claim

This synthesis does not claim:

- That brains are transformers
- That human cognition reduces to machine learning
- That AI systems are conscious or self-aware
- That hallucination is harmless or acceptable
- That current AI systems are safe

What it claims is narrower and stronger: both systems implement a common computational principle—trajectory-based reconstruction under distributed constraints—and this principle explains both their capabilities and their characteristic failure modes.

The substrates, learning processes, and developmental histories remain profoundly different. But at the level of *what computation is being performed*, the parallel is real.

10 Closing Thought

If intelligence consists in navigating uncertainty under constraint, then hallucination is not the enemy of intelligence but its shadow. The question is not how to abolish it—which is likely impossible and certainly undesirable—but how to recognise when it arises, how to bound its consequences, and how to integrate it responsibly into systems that act in the world.

Humans solved this problem through evolution: metacognition, social verification, cautious commitment. AI systems are now facing the same challenge. Rather than treating these as separate problems requiring separate solutions, we should recognise them as instances of the same fundamental issue.

The parallel evolution of human cognition and machine learning is not coincidental. Both are converging on the same solution because there may be no other scalable way to generate meaning in an uncertain world. Understanding this convergence does not diminish either form of intelligence—it clarifies the constraints under which all intelligence must operate.

The future of AI safety may depend less on eliminating hallucination and more on learning from how humans have managed confabulation for millennia: not through suppression, but through sophisticated governance of trajectory-based inference.

Acknowledgements

This work emerged through intensive dialogue with AI systems (Claude and ChatGPT), which provided formal scaffolding for ideas developed from my research in distributed transform-domain representations (DTDR) and clinical experience in medicine. While AI contributed to articulation and synthesis, the core conceptual framework, critical judgment, and responsibility for all claims remain mine. For a fuller account of this collaborative process, see my statement on AI collaboration at `https://github.com/UnrealJon/DTDR/AI_Collaboration_statement.md`

References

[1] Bartlett, F.C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

[2] Loftus, E.F. and Palmer, J.C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13(5):585–589.

[3] Schacter, D.L. (2001). *The Seven Sins of Memory: How the Mind Forgets and Remembers*. Houghton Mifflin Harcourt.

[4] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[5] West, J. (2024). *DTDR: Transform-domain representations for vector search* (Version 1.0) [Software]. *GitHub. https://github.com/UnrealJon/DTDR* .