CONSTRAINT-REGULATED COGNITION

A Unified Theory of Hallucination: Where AI Stands Today and What Comes Next

Quest Research Center™
EXPLORING THE FUNDAMENTALS OF EXISTENCE

Anil Patnaik

Quest Research Center • info@qresearch.in

Background & Aim

AI systems produce fluent, confident, yet factually incorrect outputs —hallucinations. Their persistence across architectures indicates a **structural** cause. We present the **Constraint-Lattice Model**, defining hallucination as $m \notin \mathcal{M}$ and proposing a 7-layer architecture for constraint-validated cognition.

Methods & Results

- Formal modelling: cognition as constraint intersection
- Hallucination defined as $m \notin \mathcal{M}$
- H-phenomenal vs. H-structural distinction
- Epistemological grounding (Goldman, Sosa)
- CRCS: 7-layer constraint architecture

Conclusion

Hallucination cannot be eliminated—only **constrained, detected, and corrected** via layered validation. We propose CRCS as a buildable architecture that shifts AI from pure pattern generation to constraint-validated cognition.

1. The Problem

- AI generates statistically coherent but **empirically false** outputs (Ji et al., 2022; Huang et al., 2023)
- Existing fixes (RLHF, RAG, grounding) improve outputs, reduce the frequency of errors, but do not eliminate them (Ouyang et al., 2022; Lewis et al., 2020)
- Failures persist across architectures—GPT, LLaMA, Gemini, Claude

— The Structural Gap

- No unified theory explains **why** hallucination is inevitable or what specific architectural changes are required to constrain it
- The engineering literature is rich in mitigation techniques but poor in explanatory theory. This study addresses that gap

— Core Claim

> Hallucination is the necessary shadow of generative intelligence—the same mechanism produces both errors and breakthroughs.

Human cognition manages this via **four constraint classes**:

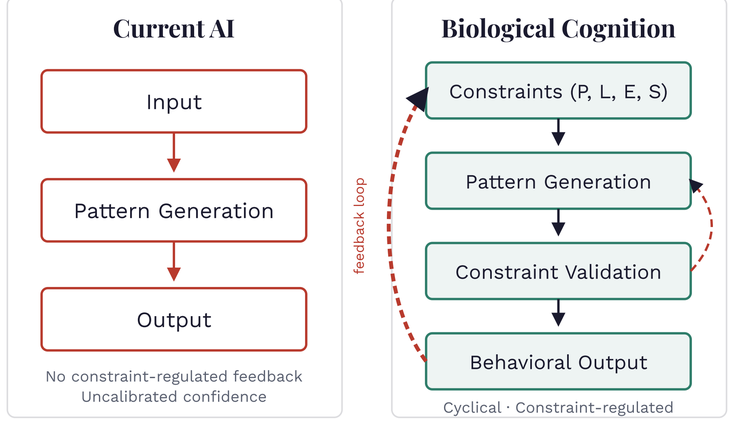
P Perceptual — Sensory grounding, embodiment

L Logical — Internal consistency, coherence

E Empirical — External validation, evidence

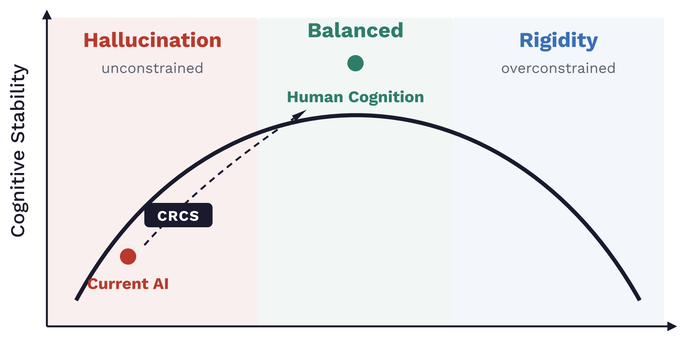
S Social — Intersubjective verification

2. Two Architectures



3. Instability Threshold

Cognitive stability requires **balance** between generation and constraint (Frith, 1992; Gregory, 1968):



- **Too few constraints** → hallucination, confabulation
- **Optimal balance** → stable, adaptive cognition
- **Too many constraints** → rigidity, inability to generalize

4. The Constraint-Lattice Model

The first **substrate-neutral formal unified theory** of hallucination that is simultaneously:

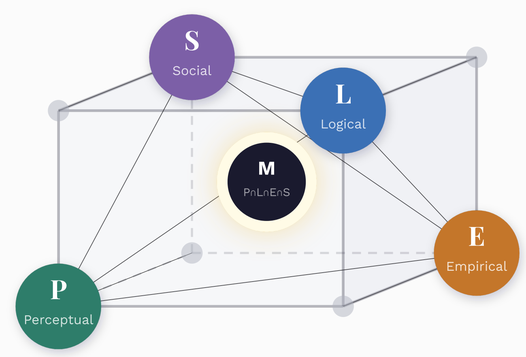
a. Philosophically defensible against the intentionality objection

b. Architecturally translatable into a buildable AI system

c. Epistemologically grounded in established reliabilist frameworks

No existing paper in the hallucination survey literature (Ji et al., 2022; Huang et al., 2023) provides a unifying formal structure of this kind.

Stable meaning emerges at the intersection of independent constraints:



Stable meaning emerges at the intersection

Stable Meaning Space

$$\mathcal{M} = P \cap L \cap E \cap S$$

Convergence (Cognition)

$$m^* = \Phi(m^*, P, L, E, S)$$

5. Hallucination Defined

H-structural hallucination

$$m \notin \mathcal{M}$$

Typical AI failure pattern

$$m \in L \cap S \quad \text{but} \quad m \notin E$$

→ Fluent and plausible, but **empirically false**

H-Phenomenal (Human only)	≠	H-Structural (AI + Human)
• Conscious experience • Requires subjectivity • Substrate-dependent		• Constraint failure (m∉M) • Substrate-neutral • Formally definable

> Preserves the analogy without claiming AI consciousness. Both reduce to $m \notin \mathcal{M}$ structurally.

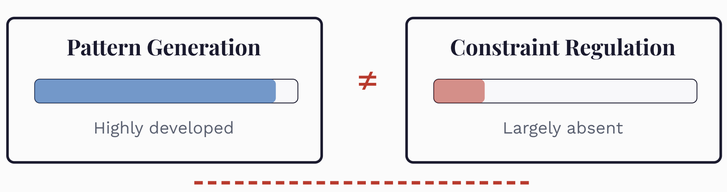
6. Constraint Score

Reliabilist scoring (Goldman, 1979)

$$\text{Score}(m) = w_p P + w_l L + w_e E + w_s S$$

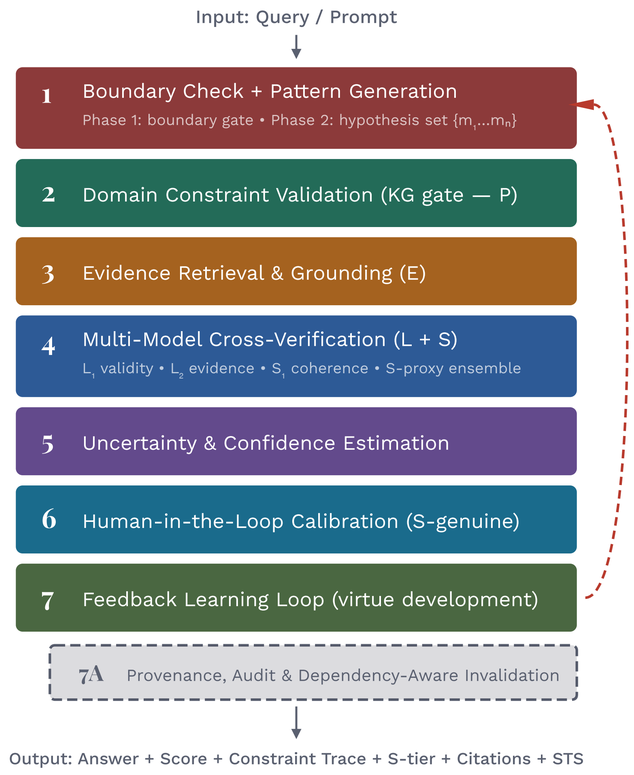
Confidence should track **constraint satisfaction**, not fluency.

- Each constraint = **intellectual virtue** (Sosa, 1991)
- Learning loop = virtue development
- Cross-verification = reflective knowledge



7. CRCS Architecture

A buildable **7-layer** Constraint-Regulated Cognitive System with iterative convergence:



> Pattern generation is hypothesis formation, not answer production. Layers 3–5 iterate until convergence; Layer 6 provides genuine human validation; 7A ensures provenance and dependency-aware invalidation.

8. Discussion & Conclusion

— Key Objections Resolved

OBJECTION	RESOLUTION
Chinese Room	Derived intentionality via constraints (Rey, 2002)
Confabulation	Phenomenal ≠ Structural
Symbol Grounding	Domain rules + evidence linkage
Substrate gap	AI lacks P → CRCS compensates functionally

— Key Contributions

- Formal definition of hallucination: $m \notin \mathcal{M}$
- H-structural vs. H-phenomenal distinction
- Reliabilist epistemological grounding
- **CRCS**: a buildable 7-layer architecture

— Implications for AI

- Optimise **constraint consistency**, not just fluency
- Every output should carry a **confidence trace**
- Missing perceptual constraint (P) → **embodied AI** (Patnaik & Sharma, 2025)

— Limitations & Future Work

- Prototype CRCS pipeline developed; extensive evaluation underway
- Constraint weights (w_i) require domain-specific calibration

> The solution is not to eliminate hallucination, but to constrain, detect, and correct it—as biology does.

AI hallucinates because it generates patterns without sufficient constraint validation. Reliable systems must implement multi-domain constraint evaluation across perceptual, logical, empirical, and social domains.

References

Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia.* Psychology Press / Lawrence Erlbaum Associates.

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proceedings of the Royal Society of London, Series B, 171*(1024), 279–296.

Ji, Z., Lee, N., Frieske, R., et al. (2022). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS, 35*, 27730–27744.

Rey, G. (2002). Searle's misunderstandings of functionalism and strong AI. In J. Preston & M. Bishop (Eds.), *Views into the Chinese room* (pp. 201–225). Oxford University Press.

Goldman, A. I. (1979). What is justified belief? In G. Pappas (Ed.), *Justification and knowledge* (pp. 1–25). D. Reidel.

Huang, L., Yu, W., Ma, W., et al. (2023). A survey on hallucination in large language models. *ACM Transactions on Information Systems, 43*(1), 1–55.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS, 33*, 9459–9474.

Patnaik, A., & Sharma, D. (2025). Cognitive framework for the study of the senses. *MetaScientia, 1*(2), 217–247.

Sosa, E. (1991). *Knowledge in perspective: Selected essays in epistemology.* Cambridge University Press.