International Journal of Advanced Engineering and Technology
www.allengineeringjournal.com
ISSN: 2456-7655
Received: 12-03-2026, Accepted: 12-04-2026, Published: 27-04-2026
Volume 10, Issue 2, 2026, Page No. 69-72

The Koyal's Egg Fallacy: AI deployment outpacing understanding, governance, and interpretability

Abhijeet Parshuram Salvi, Aaditya Arvind Singh, Sneha Kumari
Department of Computer Science, Pillai College of Arts, Commerce and Science, Panvel, Navi Mumbai, Maharastra, India

Abstract
The rapid embedding of AI systems into healthcare, law, finance, and governance has outpaced our capacity to understand, monitor, or predict their behaviour—a structural mismatch this paper terms the "Koyal's Egg Fallacy," after the cuckoo that lays undetected eggs in another bird's nest. This paper makes four documented contributions: (i) empirical evidence that AI failures—hallucinations, silent degradation, and high-stakes errors—are the statistical norm rather than the exception, with 91% of deployed models degrading in production and 75% of organizations lacking monitoring [5]; (ii) a systematic analysis of the mechanistic interpretability gap showing that current techniques are fundamentally non-scalable to billion-parameter production systems [3]; (iii) a complex-systems analysis demonstrating that multi-model interactions generate cascading failures undetectable at the component level, illustrated by the still-unexplained 2010 Flash Crash [4, 6, 7]; and (iv) a governance-deployment gap framework quantifying the asymmetry between exponential infrastructure growth (+115% AI workload capacity) and stagnant governance maturity (monitoring coverage: 25%, static). The paper argues that the structural mismatch between deployment pace and understanding represents a form of collective epistemic overconfidence—we design safeguards for mechanism A while system may operate via mechanism B. It calls for epistemic humility, mandatory monitoring before deployment, and investment in scalable interpretability before further expansion of AI into critical decision-making domains.

Keywords: AI governance, mechanistic interpretability, model degradation, hallucinations, systemic risk, flash crash, deployment safety, epistemic humility

Introduction

A koyal—the Indian cuckoo—lays its egg in another bird's nest. The host raises the imposter chick, unaware of what is happening inside. When the cuckoo hatches, it eliminates competition. The parent birds, blind to the mechanism in their own nest, continue feeding the chick that destroys their lineage. Today's AI adoption mirrors this scenario with unsettling precision. Organizations embed AI into healthcare, finance, law, and governance—yet these systems remain fundamentally opaque. Safeguards are built assuming AI operates via mechanism A. If it actually operates via mechanism B, the divergence may not become apparent until damage is irreversible.

The empirical evidence for this mismatch is not speculative. Thirty thousand medical professionals use AI transcription systems despite documented 1% hallucination rates, with 40% of those hallucinations causing harm including false treatment records, violent rhetoric, and racist commentary in patient files [1]. Lawyers have been sanctioned by courts for submitting AI-fabricated legal citations [1]. Enterprise AI adoption reached 45% by 2026—yet 91% of deployed models experience post-deployment degradation and 75% of organizations lack the monitoring infrastructure to detect it [5, 8].

This paper examines three critical structural dimensions: [1] the empirical prevalence and impact of unexpected AI behaviour in production environments; [2] the mechanisms by which component-level failures scale to systemic risk through feedback loops and cascade pathways; and [3] the governance gap between deployment velocity and organizational understanding capacity. The analysis draws on documented failure cases, mechanistic interpretability research, complex-systems literature, and deployment infrastructure data to argue that the mismatch between deployment pace and understanding constitutes a structural vulnerability of growing severity.

Literature Review

1. The Interpretability Gap

Neural networks represent information in ways that violate human intuition at every scale of analysis. Three properties make deployed models fundamentally resistant to interpretation [3]. Polysemanticity describes neurons that activate for multiple unrelated concepts simultaneously—a single neuron may respond to royal titles, large integers, and obscure proteins. Superposition compresses many features into fewer dimensions using high-dimensional geometry, making individual features inaccessible to inspection. Alien features are computational patterns discovered by networks that have no human-interpretable analogue and resist all explanation attempts.

Current mechanistic interpretability techniques are both labor-intensive and inherently non-scalable: they can analyse small circuits in simplified models containing thousands of parameters, but billion-parameter production systems present a combinatorial challenge several orders of magnitude beyond current methods [3]. Critically, the relationship between model capability and model interpretability is inverse: as models grow more powerful and more widely deployed, they become harder to explain. We are building increasing reliance on systems whose internal reasoning processes are becoming progressively more opaque, not less.

2. Complex Systems and Cascading Failure

AI systems embedded in interconnected institutions do not fail in isolation—they interact. Complex-systems theory predicts that interconnected systems can generate emergent

failures undetectable from component-level analysis alone [4]. Three cascade pathways have been identified in deployed AI infrastructure: feedback loops, where a biased algorithm generates biased decisions that produce biased training data for the next-generation model, reinforcing rather than dampening error; cascade propagation, where a security breach in one model provides entry points to all connected models; and emergent interaction effects, where multiple systems operating within their individual design parameters collectively produce outcomes no single system was designed to generate [4, 6, 7].

The 2010 Flash Crash remains the canonical empirical illustration: independent algorithmic trading systems interacted in unforeseen ways, erasing over USD 1 trillion in market value within minutes. No single program was designed to crash the market; no human error triggered the event. Sixteen years later, the precise triggering mechanism remains unresolved [7]. This is not a historical curiosity—it is evidence that our models of how AI systems interact at scale are fundamentally incomplete.

Methodology

This study employs a multi-method documentary analysis approach across four evidence streams. First, AI failure case analysis: real-world failure cases were examined across high-stakes domains (healthcare, legal, financial services, consumer applications), drawing on user-reported error data from a sample of three million mobile application reviews, post-deployment performance degradation studies, and regulatory findings [1, 2, 5]. Second, interpretability research review: primary literature on mechanistic interpretability was examined to document the specific technical obstacles—polysemanticity, superposition, alien features—preventing scalable understanding of production models, with explicit attention to the gap between research-system capability and production-system scale [3]. Third, complex-systems analysis: The Flash Crash case was examined as an empirical anchor for cascade pathway theory, and multi-bank credit risk cascade scenarios were mapped to identify how bias propagates through interconnected AI systems [4, 6, 7]. Fourth, governance infrastructure assessment: deployment growth rates were compared against governance maturity indicators including model registry adoption, monitoring coverage, and retraining schedule prevalence, drawing on global infrastructure data from 2025–2026 [8, 10, 11, 12, 13].

Results

1. Hallucination Prevalence and Impact

Table 1: AI Hallucination Metrics Across Production Deployments

Metric	Value
Hallucination prevalence in mobile app AI-error reviews	1.75% of AI-error reports
Hallucinations as share of all reported AI failures	38% (largest single category)
Proportion of hallucinations causing real-world harm	40%
Average user rating: hallucination reviews	1.8 stars
Average user rating: other AI-error reviews	3.5 stars (-1.7 differential)
Daily vs. casual users checking AI output	Daily users 14x more likely to double-check

2. High-Stakes Failure Cases

Table 2: Documented AI Failures in Critical Domains

Domain	Case	Failure Mode	Consequence
Healthcare	Whisper transcription (30,000 users)	1% hallucination; 40% harmful	False treatments, violent rhetoric, racist commentary in patient records
Legal	ChatGPT citation generation	Fabricated case citations	Lawyer sanctioned by court
Travel	Air Canada chatbot	False refund policies generated	Tribunal ruled airline liable for chatbot statements
E-commerce	Chevy chatbot jailbreak	Exploited to sell vehicle at $1	Viral reputational damage
Finance	Multi-bank credit cascade	Bias propagated across 3 AI models	Millions of lending decisions affected; undetected by regulators

3. Post-Deployment Model Degradation

Table 3: Post-Deployment Performance Degradation Statistics

Metric	Value
ML models experiencing post-deployment degradation	91%
Organizations with adequate monitoring infrastructure	25%
Organizations reporting major revenue losses from silent failures	50%
Typical accuracy loss within 6 months of deployment	Up to 35%
Bank credit model: controlled testing accuracy	95%
Same model at 9 months production (unchanged code)	87% (-8 pp, undetected)

4. Mechanistic Interpretability Scaling Gap

Table 4: Interpretability Obstacles and Scaling Limitations

Phenomenon	What It Means	Scaling Implication
Polysemanticity	Neurons activate for multiple unrelated concepts	Cannot determine which concept a neuron represents at inference time
Superposition	Features compressed into fewer dimensions via geometry	Individual features become inaccessible to human

		interpretation
Alien features	Non-human-interpretable computational patterns	Resist all explanation attempts at any scale
Technique scalability	Works on small circuits in simple models only	Does not scale to billion-parameter production systems

5. Governance-Deployment Gap

Table 5: Deployment Growth vs. Governance Maturity (2025-2026)

Dimension	2025	2026 (proj.)	Growth	Governance Status
Global AI workload capacity	38 GW	82 GW	+115%	No tracking
Enterprise AI adoption	27%	45%	+67%	No standard
AI infrastructure spending	$965B	$1.37T	+42%	No oversight
Organizations with adequate monitoring	25%	~25%	0%	Stagnant
Organizations with model registries	Insufficient	Insufficient	Minimal	Stagnant

The asymmetry in Table 5 is the paper's central empirical finding: infrastructure capacity grows exponentially while governance maturity remains static. AI workload capacity doubles in a single year; the proportion of organizations with adequate monitoring has not moved in the same period [5, 10, 11, 12, 13].

Discussion

1. The Structural Mismatch

The evidence assembled in this paper describes a coherent structural pattern, not a collection of isolated incidents. Organizations design AI governance assuming that testing provides safety, that hallucinations are rare, that failures are detectable, and that interpretability tools provide meaningful understanding of deployed systems. All four assumptions are contradicted by the empirical record. Testing does not predict production: bank credit models maintain 95% accuracy in controlled environments before falling to 87% undetected in nine months [5]. Hallucinations are not rare: they constitute 38% of all reported AI failures and cause real-world harm in 40% of cases [1]. Silent failures are the norm: 50% of organizations discover degradation only after losses accumulate, months after onset [5]. And interpretability tools fundamentally do not scale: the techniques that work on small research models cannot be applied to the billion-parameter systems making critical decisions at global scale [3].

The Flash Crash provides the starkest illustration of what this mismatch implies at systemic scale. Independent algorithms, each operating within design parameters, interacted to erase USD 1 trillion in sixteen minutes—and the precise mechanism remains unresolved sixteen years later [7]. The multi-bank credit cascade demonstrates the same dynamic in slower motion: three separately validated models reinforce each other's bias at scale, with no single model appearing problematic under individual audit [4]. These are not failure modes that better component testing would have prevented. They are emergent properties of interconnected systems that cannot be understood through component-level analysis alone.

2. Toward a Governance-First Framework

The paper does not argue against AI development. It argues that the sequence of deployment decisions should be reordered. For organizations: treat AI as poorly understood rather than fully understood—build comprehensive monitoring before deployment, maintain model registries as a precondition for system expansion, and conduct systemic interaction risk assessments rather than only component performance tests. For high-consequence domains specifically —healthcare, criminal justice, financial infrastructure—deployment should be conditional on demonstrated monitoring capability, not merely testing accuracy. For policymakers, the precedent set by aviation and pharmaceutical regulation is instructive: proof of safety at scale must precede mass deployment, not follow it. The koyal metaphor captures the epistemic problem precisely: the host birds are not negligent, they are actively nurturing. The problem is that they cannot see inside the nest. The responsible response is not to nurture faster; it is to build tools that let us see into the nest before expanding reliance on what is growing there.

Conclusion

This paper has documented a structural mismatch between AI deployment pace and the organizational capacity to understand, monitor, and govern the systems being deployed. The core findings are: hallucinations and silent degradation are statistically normal across production AI systems, not exceptional; mechanistic interpretability fundamentally does not scale to the models now making critical decisions; complex multi-model interactions generate systemic failures that component-level testing cannot predict; and infrastructure capacity is growing exponentially while governance maturity has stagnated. Collectively, these findings establish what the paper terms the Koyal's Egg Fallacy—the structural overconfidence of deploying systems whose true mechanisms diverge from the assumptions underpinning our safeguards.

Three specific interventions are prioritized. First, mandatory pre-deployment monitoring infrastructure: no organization should expand AI deployment in high-consequence domains without demonstrating monitoring capability equivalent in scope to the systems being deployed. Second, systemic interaction assessment: regulatory frameworks should require assessment of how new AI deployments interact with existing systems, not merely how they perform in isolation—treating interconnected AI as an infrastructure risk category equivalent to interconnected financial institutions post-2008. Third, honest governance of interpretability limits: policymakers and organizations should require explicit disclosure of which aspects of a deployed model's reasoning are interpretable and which are not, treating opacity as a governance risk factor rather than an acceptable technical condition. The question is not whether AI development continues. The question is whether the current pace of deployment is justified by the current level of understanding. On the evidence presented here, it is not.

References

1. Weidinger L, Mellor J, Rauh M, Gabriel I, Krueger D, Andersson J, Gabriel I. *et al*. LLM hallucinations and failures: Lessons from real-world applications. EvidentlyAI Research Report, 2024.
2. Liu Y, Zhang W, Wang J, Qu B, Chen S. User-reported LLM hallucinations in AI mobile applications: A large-scale empirical study. Nature Scientific Reports,2025:15(1):2847.
3. Nanda N, McGrath T, Rauh M, Hubinger E, Schiefer N, Joly A. *et al*. Mechanistic interpretability for AI safety: A comprehensive review. arXiv Preprint, arXiv,2024:2401:08307.
4. Kondor D, Posfai M, Csabai I, Vattay G. Complex systems perspective in assessing risks in artificial intelligence deployment. Philosophical Transactions of the Royal Society B,2024:379(1901):20230268.
5. Barash G, Farchi E, Izsak A, Kaplan H, Matalon Y. Model drift in machine learning: Comprehensive analysis of post-deployment performance degradation. IEEE Transactions on Software Engineering,2024:50(8):1847-1863.
6. SEC. Findings regarding the market events of May 6, 2010. U.S. Securities and Exchange Commission Report to Congress, Washington, DC¸ 2010.
7. Rambachan A, Roth J, Stoye J. The cascading effect in AI/ML systems: From component failures to systemic collapse. Machine Learning Safety Research,2025:8(2):127-145.
8. Statista. AI infrastructure and datacenters 2026: Global trends and statistics. Statista Digital Report Database, 2026.
9. Rev.com. Study: Heavy AI users encounter hallucinations 3x more frequently than casual users. Technology Research Report, 2025.
10. Jones Lang LaSalle. 2026 Global Data Center Outlook: Market trends and forecasts. JLL Research & Forecasting Division, 2026.
11. Programs.com. Measuring the data center boom: Facts and statistics 2026. Technology Infrastructure Analysis Report, 2026.
12. Technavio. Enterprise AI market analysis, size, and forecast 2025-2029. Market Intelligence Report, 2025.
13. Gartner & Express Computer. Gartner forecasts global AI spending to reach $2.5 trillion in 2026. Executive Summary and Analysis, 2026.