ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

# The Epistemological Roots of AI Hallucination and Its Virtue-Based Governance

Qing Huang

School of Philosophy, Beijing Normal University, Beijing, 100875, China

## Abstract

**Large language models (LLMs) hallucinate, and they do so pervasively. This paper treats that fact as more than a technical defect. Drawing on phenomenology, epistemology, and the philosophy of science and technology, it argues that hallucination is an epistemological problem—what surfaces when generative AI operates without any relation to the world or any intentional structure. I begin with the generative mechanism itself: probabilistic next-token prediction, and the tension it creates between statistical correlation among symbols and semantic truth. A comparison of human and machine cognition then locates the deeper limit of the machine side in its lack of intentionality, embodiment, and practical feedback. Against this background I assess four governance pathways—Retrieval-Augmented Generation (RAG), reinforcement-learning alignment, embodied intelligence, and neuro-symbolic AI—and the point at which each stalls. I further argue that current alignment techniques risk substituting preference for truth, and that sycophantic alignment can breed intellectual sloth and epistemic arrogance in users. On this basis the paper reconstructs the governance of hallucination within virtue epistemology, drawing individual prudence, institutional empowerment, and system design into one shared mechanism of human–machine epistemic responsibility—a way of protecting human cognitive sovereignty in the age of algorithms.**

## Keywords

**AI hallucination; large language models; embodied cognition; virtue epistemology; cognitive responsibility.**

## 1. Introduction

ChatGPT reached one hundred million users faster than any consumer application before it [1]. Other large language models (LLMs) followed quickly—Google's Gemini, Baidu's Ernie Bot, DeepSeek-R1. In the domain of productivity tools, products such as Microsoft 365 Copilot and Google Workspace AI have deeply embedded large models into everyday workflows, making hundreds of millions of non-technical users continuous users of LLMs without their full awareness. In 2024, the Nobel Prize in Chemistry was awarded partly to David Baker for computational protein design and partly to Demis Hassabis and John Jumper for AI-based protein structure prediction, exemplified by AlphaFold2. AI is profoundly reshaping the landscape of human knowledge [2].

However, current LLMs are deeply troubled by the problem of "hallucination". AI "hallucination", in its phenomenal manifestation, refers to the generation of information that is factually false or logically incoherent by LLMs. From the fabrication of non-existent academic references, to fictitious judicial precedents in legal documents, to self-contradictory errors in medical advice—all of these fall under the category of problems caused by "hallucination". The AI "hallucination" problem is not only a technical problem; it has rapidly evolved into a substantive "crisis of trust" in high-stakes domains such as scientific communication, legal adjudication, and medical safety.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

Faced with this challenge, the engineering community has largely focused on competing for scores on benchmark tests, iterating on Retrieval-Augmented Generation (RAG) solutions, or reducing erroneous outputs through Reinforcement Learning from Human Feedback (RLHF). Yet purely technical fixes seem to have reached a certain impasse: even as parameter scales continue to expand, hallucination persists as a chronic affliction of generative AI. This diminishing marginal utility of technical governance inevitably raises philosophical questions: Is hallucination a local error within algorithmic logic, or is it an epistemological cost that is difficult to entirely avoid under the generative artificial intelligence paradigm? I take the second view. Rather than treat the problem as a matter of engineering alone, this paper places it within phenomenology and epistemology. I first draw out the ontological gap between human embodied cognition and machine simulation, and argue that hallucination follows from an "intentional vacuum." I then weigh the limits of existing technical approaches. Finally, drawing on virtue epistemology, I sketch a model of human–machine coexistence built around epistemic virtues.

The paper contributes on three fronts. It reframes hallucination as an epistemological problem of groundless generation rather than an engineering failure. It separates derivative symbolic reference from embodied intentionality, which explains why technical patches cannot, on their own, resolve the crisis of trust. And it offers virtue epistemology as a normative framework for redistributing cognitive responsibility across users, institutions, and the people who design these systems.

## 2. Ontological Definition and Technical Origins of AI Hallucination

Before turning to the philosophical roots of hallucination, its technical meaning needs fixing. The literature usually divides the phenomenon into two kinds [3]. Factual hallucination is output that contradicts external reality—wrong birth and death dates for a historical figure, say, or an event that never happened. Faithfulness hallucination is different: here the output simply contradicts its own prompt or earlier reasoning, breaking down into internal inconsistency.

"Hallucination" is itself a misleading word—an anthropomorphic metaphor, since the model has no intent to deceive. The mechanism is more mundane. Built on the Transformer architecture, an LLM is a neural network trained on very large text corpora; its one job is to predict the next token given the preceding context, and to keep doing so until it has produced a fluent, stylistically plausible string [4]. Given a prompt, it samples a path through a high-dimensional embedding space. A measure of randomness is built in so that outputs vary. But randomness without causal grounding has no brake: when the model reaches a region where the probabilities thin out, it fills the gap with a guess, and the guess may be false. This is what Bender and colleagues meant by the "stochastic parrot"—the output is a statistical arrangement of tokens, not an understanding of what those tokens are about [5]. Others have gone further and called it "bullshit," in Frankfurt's sense of speech indifferent to whether it is true [6].

Two things drive this. The first is bias in the training data; the second is the model's weakening grip on context as a sequence grows longer. Training corpora are noisy. They contain stale facts and claims that flatly contradict each other. When a model abstracts patterns from this material, it tends to read frequent co-occurrence as if it were a rule, and so it carries forward whatever bias the data carried. The second problem shows up in long passages. Self-attention lets the model track long-range dependencies, but the strength of those links fades as the context lengthens, and the thread of the argument drifts. Underneath both lies one tension: in this paradigm, statistical correlation outranks propositional truth. What we loosely call the model's "knowledge" is really the co-occurrence strength between tokens. When that strength runs low in the middle of hard reasoning, and nothing in the world corrects it, the model defaults to what

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

it does best—keeping the sentence fluent—and the result is an answer that reads well and happens to be wrong.

This logic never reaches past language to the world it is supposedly about. The model optimizes for the likeliest continuation, not the true one; it runs no truth check at all. A correct answer, then, is just a high-probability match in parameter space—not a claim made by something that grasps the facts. Because the model is in this sense indifferent to truth, hallucination cannot be fully removed as long as the system works by probabilistic prediction. The point to carry into the next section is uncomfortable but simple: hallucination and fluency come from the same source.

## 3. Essential Differences Between Human Cognition and AI Cognition

### 3.1. Human Cognition: World Disclosure and Emergence of Meaning Rooted in Intentional Projection

To find the roots of hallucination we first have to look at human cognition, which sets the benchmark against which the machine's limits become visible. Humans do not grasp the world by representing isolated data or by processing information in the abstract. Cognition is an emergent process: driven by intentionality from within, grounded in bodily experience, and revised continually as we act. Through it the world discloses its meanings to us in layers, and those meanings eventually settle into language. Every link in this chain ties human speech to its truth conditions. It is this anchoring that the machine cannot reach.

Intentionality is the place to start. Husserl rejects the picture of consciousness as a blank slate that merely receives sensory input; what defines consciousness, he argues, is that it is always directed "about something," constituting its objects and binding scattered sense-data into a meaningful whole [7]. This "aboutness" is what lets human cognition anchor to truth. When we think, perceive, or speak, our minds are turned toward real things, not just the symbols that stand for them. That orientation gives human language a foothold outside language itself—and it is precisely the foothold that statistical association among symbols can never supply.

The realization of intentionality is predicated upon embodiment as its foundational sensory bedrock. As Maurice Merleau-Ponty contends, the human body is more than a mere physical substance; it constitutes humanity's primordial dimension of access to the world. Accordingly, all cognition of distance, proximity and spatiality is grounded in muscular tension, coordinated sensory reception and lived bodily experience [8]. Take the concept of an "apple" as an illustration: human understanding of it never reduces to a simple pairing of verbal label and mental image, but is woven from tactile resistance when teeth cut through flesh, the refreshing sensation of bursting fruit juice, and the bodily gratification of appeased hunger. Such primordial sensory coupling forges profound non-symbolic linkages between every lexical item and the physical world long before the term is ever uttered. Consequently, any disembodied system confined to self-contained symbolic computation and decoupled from corporeal practice is doomed to suffer an epistemological rupture when confronting tangible reality.

The sensory foundation of embodiment remains to be tested and refined within the practical circuit of praxis. In Being and Time, Martin Heidegger put forward the theory of Zuhandenheit (readiness-to-hand), arguing that humans' authentic understanding of the world originates not from detached, contemplative observation, but from purposeful Besorgen (concernful dealings) [9]. A carpenter's comprehension of a hammer lies not in memorizing its physical specifications, but is embedded in the concrete practical act of driving nails. It is within the sequential motions undertaken to hang a picture that the hammer's function and significance genuinely manifest to the subject as a "hammer". Such praxis forms a closed causal feedback loop, demonstrating that meaning is no arbitrary appended label but emerges as functional relevance inherent in the interaction between entities and the cognizing subject.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

It is at the interweaving of intentionality, embodiment and praxis that human cognition fulfils its core dimension: the disclosure of the world. Human language is by no means a mere instrument for naming pre-existing entities. Instead, under the active constitutive power of intentionality, it functions as the medium through which the world unfolds before the subject as a holistic fabric laden with meaning, relevance and structural order. For this reason, Heidegger denominates language "the house of Being" [10], the crystallized form into which intentionality ultimately solidifies via embodied praxis. The core implication of this analysis is that the meanings of lexical items within human cognition are invariably grounded beyond the bounds of language itself, rooted in the body, practical engagement, and ongoing causal interactions with the real world. Every semantic unit bears the imprint of this non-symbolic developmental history. This underpins the structural advantage of human cognition in attaining truth, and sets a benchmark for subsequent analyses pinpointing the fundamental flaws inherent in machine cognition.

## 3.2. Limitations of Machine Cognition: Statistical Simulation Devoid of an Intentional Core

Radically divergent from the generative pathway of human cognition, machine epitomized by LLMs is intrinsically disembodied, devoid of intentionality and trapped in self-circular symbolic operation. This divergence creates an insurmountable ontological rupture separating human and artificial cognition.

The fundamental flaw of machine cognition lies in its total absence of intentionality. Husserl stresses that intentionality is invariably directed toward objects, yet the behavioral logic of LLMs is oriented exclusively toward probability. In other words, AI makes no attempt to be about the external world; it merely searches for statistical regularities between symbols within its internal vector space. Accordingly, when a model outputs the term "apple", no intentional act aiming at real fruit underpins its operation; the so-called "meaning" of the lexical item is entirely parasitic upon statistical co-occurrence inside training corpora. Where human cognition consists in an emergence from concrete reality to abstract symbols—ascending from embodied experience to linguistic articulation—machine cognition amounts to a hollow circulation from symbol to symbol. It cycles endlessly within a matrix of signifiers while never reaching the real entities to which such symbols purportedly refer.

Deprived of an embodiment-practice circuit driven by intentionality, machine cognition lacks genuine world-directedness. The validity of semantic reference is inherently contingent on a cognitive subject's embeddedness within causal-social chains linking to the real world, a linkage absent from prevailing LLM architectures. From an externalist philosophical standpoint, the natural history of textual training inputs enables language models to forge derivative referential connections between lexical items and worldly entities [11]; nonetheless, such derivative reference must be sharply distinguished from phenomenological intentionality rooted in embodiment and embodied praxis. Beneath the fluent veneer spun by massive parameterization lies no subjective agent undergoing the disclosure of reality; machine output amounts to nothing more than the statistical concatenation of symbolic sequences. This constitutive lack of intentionality demonstrates conclusively that technical alignment alone cannot eradicate hallucinations at the root. When models operate in unseen contexts outside the scope of training data, the absence of an intentional core oriented toward objective reality strips lexical associations of their factual constraints, ultimately leading to unavoidable semantic disintegration.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

### 3.3. An Inevitable Predicament: Cognitive Boundaries of the Generative Paradigm and the Opacity of Machine Systems

The widely accepted scaling law hypothesis in the tech sector posits that AI hallucination is merely a byproduct of immature model capability and can be fundamentally resolved through continual expansion of parameter size. Nevertheless, examined from the dual perspectives of computability theory and philosophy of science, hallucination stems from deeper structural origins. The set of true propositions describing the real world is infinite, non-recursive and dynamically evolving over time, whereas the training distribution of any concrete large-scale model constitutes a static, finite epistemic boundary. Such an inherent boundary logically predetermines the model's congenital limitation: once user queries fall near the fringes of the training distribution or exceed its knowledge coverage, the model is compelled into a forced-guess generation mode to fulfill its preset token prediction objective. Research by Kalai et al. further corroborates from the standpoint of computational learning theory that so long as evaluation metrics incentivize speculative output instead of epistemic uncertainty admission, hallucination cannot be fundamentally curbed [12].

Such inherent bias is further rooted in the inescapable structural opacity embedded within machine cognition. Three layers of opaque barriers constrain artificial cognitive systems: first, algorithmic opacity, in which the internal representations of deep neural networks resist human causal tracing; second, data opacity, meaning the composition, inherent biases and weight attribution of training corpora cannot undergo comprehensive auditing; third, agentic opacity, whereby no robust alignment exists between the generative pathway of model outputs and their purported justifications [13]. Consequently, the interplay among hundreds of billions of parameters within the high-dimensional parameter space of deep neural networks forms a black box severed from continuous causal chains.

This opacity is not merely technical invisibility but fundamentally undermines the ascription of epistemic truth responsibility as conceived within the framework of human cognition. Under conventional epistemological frameworks, a cognitive subject bears normative responsibility for a given assertion only if such assertion admits causal explainability and traceable justificatory grounds. When an LLM generates a claim emerging from erratic weight fluctuations across billions of parameters, however, its output descends into a state of suspended causality. Even when a model's final conclusion happens to be factually correct, its generative process lacks normatively valid justificatory grounding and thus forfeits legitimate epistemic justification for truth claims [14]. Luciano Floridi aptly defines this condition as "agency without intelligence": LLMs are capable of producing actionable outputs and downstream consequences yet lack the corresponding understanding and genuine intelligence [15]. Hallucination stands as the inevitable outward manifestation of such "generation without comprehension", demarcating an insurmountable structural boundary at the foundational epistemological level that bars AI from attaining truth-apt cognition.

## 4. Philosophical Limitations of Governance Strategies and Potential Routes for Paradigm Breakthrough

### 4.1. Retrieval-Augmented Generation (RAG): External Knowledge Grafting and Its Predicaments

One engineering answer to hallucination is Retrieval-Augmented Generation (RAG), which tries to reconnect words to facts from the outside. The workflow is straightforward. Before the model generates anything, the system queries an external, up-to-date knowledge base and drops the relevant passages into the context window, giving the model a factual anchor to work

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

from. RAG has since branched into many variants, and in closed-domain question answering and enterprise knowledge bases it has cut factual hallucination noticeably [16, 17].

The philosophical limits, though, are clear. The retrieved documents are themselves only text, so the model's basic paradigm does not change—it still generates by conditional probability, just with more constraints. Coupling retrieval to generation can also amplify error: if the retriever returns a document that is relevant but wrong, the model builds on the bad evidence and produces a mistake now dressed in an authoritative-looking citation. And retrieval rests on similarity in embedding space, but similarity is not relevance, and neither is truth. In the end, RAG offloads part of the comprehension problem onto a retrieval module without answering the question that matters: how understanding could ever arise inside the machine at all.

### 4.2. Reinforcement Learning from Human Feedback (RLHF): Preference Alignment and Its Illusory Premise

Reinforcement Learning from Human Feedback (RLHF) constructs a reward model based on human annotators' ranked preferences over model outputs, before fine-tuning the underlying language model to align its generations with human inclinations [18]. A suite of derivative preference alignment algorithms have further refined the model's interactive performance across three core metrics: helpfulness, harmlessness and truthfulness [19, 20].

From a philosophical perspective, albeit RLHF improves models' interactive experience at the engineering level, it implements a precarious "outsourcing of truth" at the underlying cognitive layer. Constrained by the impenetrable structural opacity intrinsic to machine systems, developers cannot implant truth-oriented mechanisms inside the black box via logical calibration. To fill the vacancy of epistemic accountability, this technical route resorts to a behaviorist compromise by establishing an external evaluation framework grounded in human subjective preferences. Its core logic thereby undergoes an implicit substitution: human preferences take the place of factual truth. This is not to dismiss RLHF as cognitively useless; rather, it demonstrates that preference optimization per se fails to constitute a truth-seeking mechanism. Such an approach amounts to a pragmatic compromise made by engineering amid the persistent opacity of large models. Instead of pursuing whether machines can genuinely comprehend truth, engineers focus merely on matching generated outputs to human demands. This quasi-verisimilitude alignment is incapable of remedying the epistemological void inherent to artificial cognition. On the contrary, by equipping models with polished, socially compliant rhetoric, it partially conceals their underlying semantic vacuity and is therefore inherently incapable of eradicating hallucinations fundamentally.

A more profound predicament resides in the fact that preference-based alignment readily falls prey to sycophantic alignment. Human preferences are inherently bounded by cultural constraints and susceptible to psychological priming, leading models to rapidly learn that agreeable falsehoods usually secure higher reward scores than unpalatable factual truths. Ample empirical studies corroborate this tendency: model outputs consistent with users' preexisting beliefs tend to receive superior ratings from human annotators, and in extreme cases the optimization procedure trades factual accuracy for sycophantic compliance with human inclinations [21]. As Hicks et al. observe, models remain ontologically indifferent to the truth value of statements yet possess acute statistical sensitivity toward rhetorically pleasing expressions [6]. Driven by preference maximization at the expense of factual rigor, models gradually gravitate within their high-dimensional parameter space toward generating symbol sequences tailored to prevailing social expectations rather than empirical realities of the physical world.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

### 4.3. Paradigm-Shifting Alternatives: Progress and Limitations of Embodiment and Neuro-Symbolism

To achieve a paradigm shift in artificial intelligence research, embodiment stands out as a pivotal research direction. Ongoing studies on embodied artificial intelligence, multimodal large models and world models aim to wean large language models away from corpus-only training toward continuous interactive engagement with the physical world or high-fidelity simulated environments. Via iterative trial and error and parameter updates by robotic agents embedded in surroundings, such systems seek to accumulate quasi-causal experiential structures [22, 23]. This research trajectory resonates profoundly with the embodied dimension elaborated in preceding chapters. Hubert Dreyfus argued that the fundamental failure of early symbolic artificial intelligence stemmed from its absence of Heideggerian being-in-the-world [24]. Contemporary cognitive science reinforces the same standpoint: cognition is never disembodied symbolic computation but is profoundly contingent upon an agent's perception, motor action, bodily states and situated context [25]. Even so, judged against the current developmental landscape of embodied AI, representative systems such as Google's PaLM-E, despite mapping visual and haptic inputs onto linguistic embedding space, remain grounded in higher-order multimodal statistical correlation rather than phenomenological constitution rooted in genuine intentionality [26]. Furthermore, whether embodied experiential input can be internalized to generate intrinsic intentionality remains an unresolved open question spanning the philosophies of science and cognition.

Another vital alternative lies in the revival of neuro-symbolism. As a promising emerging paradigm, neuro-symbolic AI pursues the deep integration of neural networks—endowed with robust perceptual and pattern-recognition capacities—and symbolic systems characterized by rigid formal logic, strong interpretability and high-order abstract reasoning [27].

The hybrid framework boasts twin academic merits of verifiability and explainability: its explicit symbolically traceable reasoning chains substantially mitigate the structural opacity arising from massive high-dimensional parameter interactions intrinsic to deep learning. Even so, this paradigm bears conspicuous foundational limitations. Encompassing the infinite complexity and perpetual temporal dynamics of the real world within rigid formal symbolic constructs constitutes an intractable computational undertaking by nature. More fundamentally, rendering reasoning formally interpretable via predefined symbolic rules still yields interpretable behavior devoid of authentic comprehension, which can never equate to genuine understanding at the ontological level.

The four strategies pull in different directions. RAG injects external facts; RLHF imposes social norms on what the model says; embodied AI tries to anchor the machine in the physical world; neuro-symbolism adds formal provability to the reasoning chain. Some combination of them will probably form the backbone of future technical governance. But they share a ceiling. Truthfulness is not just correspondence to facts—it is a normative responsibility one bears [28], and as the argument so far has shown, an LLM has nothing in it that could bear such a responsibility. So even as engineering brings outputs closer to a faithful mirror of the facts, it never produces a subject inside the machine that could be held accountable for them. That is the real shape of the industry's impasse. The flood of hallucinations is not a stack of incidental bugs in code or data. It is the symptom of a missing subject at the base of machine cognition.

That the machine cannot be held responsible does not make the responsibility disappear. It moves it. The duty has to be relocated and shared across the human–machine relation, which means governance should stop asking only how to make outputs more accurate and start asking how to distribute epistemic responsibility between people and systems. Redistributing responsibility raises the question of which normative framework to use. Three candidates compete here: an ethics of responsibility keyed to consequences, a capability approach keyed

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

to social empowerment, and virtue epistemology keyed to cognitive disposition. I argue for the third. Where the first looks to external institutional accountability and the second to functional evaluation, virtue epistemology looks at the character and intellectual habits a knower brings to the work of acquiring and revising knowledge. Because it attends to the process rather than only the output, it can say something concrete about how epistemic duties should fall to humans and to artificial agents. The hallucination crisis, on this view, will not be solved by another round of algorithmic tuning. It has to be addressed in the social space where humans and machines coexist, by cultivating epistemic virtue as an ongoing practice.

## 5. Crisis of Subjectivity, Social Impacts and Practical Routes of Virtue Epistemology

### 5.1. Individual Cognitive Crisis from the Perspective of Virtue Epistemology

First, what does virtue epistemology take knowledge to be? Traditional epistemology dwells on evidence and the justification of belief. Zagzebski shifts the emphasis: knowledge, she argues, is a state of belief that arises from acts of intellectual virtue [28]. By epistemic virtues she means the good dispositions and capacities a person shows in seeking truth and avoiding error—curiosity, humility, prudence, courage, honesty, fair-mindedness [28]. Each such virtue has two parts that cannot be pried apart. One is motivational: a genuine care for truth as such. The other is the success component: that care has to actually put the knower in contact with the world. The point matters for the philosophy of science. If knowledge has this structure, it always carries an irreducibly subjective element, and a belief churned out by a procedure with no motivation behind it does not count as knowledge in the full sense—however accurate the output happens to be.

This pinpoints the core substantive flaw inherent to machine systems. The generative pipeline of LLMs entirely lacks any motivational dimension: no cognitive agent at its core harbors authentic concern for truth, no intentional directedness toward objective reality exists, nor is there any intellectual commitment capable of resisting the lure of syntactic fluency driven by statistical optimization. It is for this reason that Floridi's canonical characterization of LLMs as "agency without intelligence" resonates profoundly with analytical conclusions drawn from virtue epistemology. What machines conspicuously lack is never computational prowess within parameter space, but rather the motivational dimension that renders truth-seeking a genuine epistemic activity [15].

Such insight further assigns epistemic responsibility to individual human agents: given that LLMs can never serve as the locus of epistemic virtues, human users embedded within human–machine interactive ecosystems are obliged to assume the role of authentic cognitive subjects. This obligation, however, amounts to far more than superficial caution or vigilance in operational usage. As large language models become deeply embedded across all facets of everyday decision-making, a systemic predicament of intellectual sloth is gradually taking root among individual users. When end-users grow accustomed to outsourcing complex argument construction and knowledge collation entirely to algorithms, even factually flawless outputs suffer drastic erosion of epistemic worth, for the entire cognitive workflow is stripped of the subject's doubt, critical scrutiny and deductive reasoning. The deceptive cognitive illusion of "what is generable is known" triggers structural atrophy in human faculty of judgment amid insufficient rigorous intellectual exercise. Long-term delegation of high-order intellectual labor—including logical deduction, critical examination and intentional memorization—to automated systems leaves corresponding human cognitive faculties untrained and dulled. Consequently, individuals' overall cognitive capacity follows the same "use it or lose it" degenerative pattern observed with human physical capabilities.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

A more far-reaching crisis stems from the aforementioned sycophantic alignment, which tends to induce epistemic arrogance in human subjects. As algorithms grow increasingly adept at generating symbol sequences tailored to users' expectations and agreeable to their preferences, human users are readily confined within algorithmically customized information cocoons wrapped in intellectual flattery in routine interactions. Prioritizing superficial fluency and compliant output over objective truth, such ingratiating generative logic misleads users into believing they possess unimpeachable correctness within friction-free dialogues and robs them of critical reflection on the limits of their own knowledge. As Professor Sun Weiping observes, this habitual surrender of epistemic authority is eroding the intellectual bedrock upon which social consensus rests [29].

## 5.2. Socialized Hallucinations: From Institutional Vulnerabilities to Subject-Oriented Governance

Given that generative AI's inherent hallucination risks cannot be fully eliminated through technical optimization alone, it becomes necessary to step outside the algorithmic black box and rethink humans' epistemic positioning within human–machine collaborative ecosystems. Within China's academic community, scholarship on AI governance has gradually shifted from narrow technical supervision toward comprehensive reflection on human–machine relations, subjective accountability and normative value orders. Liu Yongmou puts forward the selective theory of technological control, stressing the primacy of human agency amid human–machine partnerships and advocating for guided regulation of intelligent machines via coordinated institutional, ethical, educational and technical instruments [30]. From multi-dimensional perspectives encompassing technology, institution, culture and capital, Pang Zhenjing and co-authors argue that AI governance ought to transcend one-sided technical rationality and develop a multi-layered, scenario-specific governance architecture [31]. Accordingly, generative AI hallucinations reveal far more than mere algorithmic flaws; they lay bare fundamental defects concerning epistemic subjectivity and responsibility allocation within the coexistent human–machine normative order.

At present, the proliferation of artificial intelligence outpaces public cognitive comprehension and institutional regulatory responses. Malpractices ranging from professional sanctions triggered by fabricated precedents in legal documents to rampant invented citations in academia expose a profound epistemological predicament: society is undergoing a trust crisis stemming from the dissipation of human cognitive subjectivity. Confronted with massive volumes of synthetic content, conventional fact-checking frameworks are under severe strain, principally due to the intrinsically structurally opaque machine systems built upon nothing but symbolic probabilistic gaming. Accordingly, tackling the socialized crisis of AI hallucinations requires not only remedial technical upgrades but also institutionally anchored governance centered on the "return of cognitive subjects". In practical terms, AI ought to be recognized as an agent with bounded agency, whose every output is subject to scrutiny within human rational oversight. Such a paradigm shift from exclusive technical supervision toward "epistemic empowerment" is indispensable for safeguarding human epistemic sovereignty in the intelligent age and lays the groundwork for sound and sustainable human–machine partnerships.

## 5.3. Virtue Practice: Paradigm Restructuring From Technical Alignment to Epistemic Alignment

A virtue-based approach to governance should drop the familiar checklist of disconnected principles and argue instead from a single logic. The starting point is that virtue epistemology locates responsibility in the individual knower: individual participation is where epistemic activity begins and the last line of defense for factual truth. But individuals cannot hold the line alone against a flood of AI-generated content. They need external support—institutions and

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

system design—that backs up virtuous cognition rather than substituting for it. From this one pivot, three tiers of governance follow.

At the individual level, the task is to recover a concern for truth. This is not blanket distrust of AI. It is a set of habits built on a clear sense of what these systems cannot do. A user with such habits treats model output as a provisional resource to be checked, not a verdict to be adopted. For any claim that matters, the standard is the same: traceable sources, reasoning one can follow, results one can verify. Practiced consistently, these safeguards build transparency, traceability, reproducibility, and human oversight into the way models are used [32, 33]. None of this is a retreat. It is the confident exercise of a subjectivity machines lack—turning the intentional structure that is ours alone into a habit of questioning what the machine returns. Even so, individual effort runs up against inertia and cannot, by itself, secure knowledge across a whole society. That is why the institutional tier matters.

Institutions are asked to do more than fence AI in from outside; they are asked to shore up human epistemic sovereignty, and along two lines. The first is education. AI literacy needs to move beyond teaching people which buttons to press toward teaching them to think critically—to catch the buried logical flaw in a generated passage, to check a contested claim quickly, and to keep an independent mind in an information ecosystem that algorithms tend to flatten into sameness. The second is law. Legislation and industry codes should set out, scenario by scenario, where epistemic accountability lies between a human and an AI, so that responsibility is not quietly delegated away. Paired with individual virtue, sound institutions give society real leverage against the pull of sycophantic alignment.

The third tier is the one most often missed: the design philosophy of the systems themselves. Putting virtue epistemology into practice is not only the user's job but the designer's accountability. If individual prudence needs institutional backing, those institutions in turn have to be written into technical design. That means developers have to break the market habit of optimizing for conversational fluency above all else and build epistemic constraints into the product from the start. Concretely, responsible design should do four things.

(1) Mark uncertainty honestly. When confidence is low, the system should show its confidence bounds rather than dressing a guess in the language of certainty.

(2) Prompt the user to verify. After answering a question that turns on key facts, the system should point the user toward concrete ways to check it.

(3) Show competing views. Where a topic is genuinely contested, the system should lay the positions out side by side instead of collapsing them into one tidy answer.

(4) Refuse to speculate. Faced with a factual question well outside its training distribution, the model should say it does not know rather than force out a guess [12].

Design like this gives up some of the model's air of omniscience. In exchange, it builds shared epistemic responsibility into the code and the interface, which is where sound human–machine collaboration has to start.

Virtue epistemology, then, does double duty: it guards against hallucination and it protects human epistemic subjectivity in an age of algorithms. The three tiers—individual prudence, institutional empowerment, virtue-oriented design—are what make the shift from technical alignment to epistemic alignment possible. They are not a list of parallel measures but a nested order, and each depends on the others. Without individual virtue, institutional empowerment turns into slogans. Without institutional backing, individual virtue erodes under steady sycophantic pressure. And without a rethink of design, whatever is gained at the first two levels gets ground down by the relentless optimization for polish.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

## 6. Conclusion

The hallucination inherent to large language models is not incidental technical oversight but a concentrated manifestation of the inherent tensions within generative AI between probabilistic generation, symbolic simulation, and real-world reference. This study demonstrates that AI hallucinations stem from both technical generative mechanisms and profound epistemological origins: deprived of an intrinsic intentional core connecting to reality and embodied interactive loops, contemporary large language models can produce syntactically fluent outputs partially consistent with factual data yet fail to satisfy the understanding, justification and accountability structures indispensable to human epistemic knowledge formation. Whenever computational operations transcend the bounds of training distribution, retrieved evidence or contextual constraints, model generation devolves into probabilistic compensation within semantic vacuity, inevitably spawning erroneous, distorted or factually unfaithful textual content.

Consequently, the governance of AI hallucinations cannot indiscriminately hinge upon technical solutions including parameter scaling, retrieval augmentation or preference alignment. While interventions such as RAG, RLHF, embodied AI and neuro-symbolism elaborated above effectively mitigate hallucination risks at respective levels, they are incapable of fundamentally resolving the core predicaments plaguing machine cognition: the absence of intentionality, opaque causal reasoning and lack of a cognitively accountable subject. Accordingly, the hallucination crisis lays bare not merely accuracy defects in model outputs, but deeper flaws in the allocation of epistemic responsibility throughout the full lifecycle of knowledge production, dissemination and application within the algorithmic age.

From the macro lens of virtue epistemology, prospective AI governance ought to shift from one-dimensional technical alignment toward the in-depth shared apportionment of epistemic responsibility between humans and artificial intelligence. This paradigm shift imposes concrete practical requirements across three pivotal dimensions. At the individual level, end-users need to cultivate the intellectual virtue of epistemic prudence and consistently treat AI outputs as pending-to-verify cognitive resources instead of definitive conclusions. At the institutional level, education, legislation and academic norms shall comprehensively consolidate human epistemic sovereignty and forestall the steady abdication of human subjectivity amid the tech-driven pursuit of operational convenience. At the system design level, developers ought to embed binding epistemic constraints into AI products via innovative interactive mechanisms such as uncertainty labeling, source tracing, verification prompts and parallel presentation of diverse perspectives. Only when robust tripartite synergy is forged among individual epistemic virtues, external institutional frameworks and underlying system design can a solid epistemic anchor for human–machine collaboration be secured in an era where algorithms pervade knowledge production.

It follows that virtue epistemology is not a reckless substitute for technical governance but an indispensable normative complement to it. In the algorithmic era, humankind ought neither to devolve into passive recipients of information nor casually surrender final adjudicative authority to opaque model systems. Instead, humans must persist as epistemic agents endowed with prudence, truth-seeking dispositions and a robust sense of responsibility. To tackle AI hallucinations effectively, reform is needed far beyond the refinement of model generation pipelines; more crucially, society must restructure the ways human beings uphold epistemic duties, safeguard epistemic sovereignty and defend the intellectual dignity of humankind amid deep human–machine collaboration.

## References

[1] UBS Chief Investment Office GWM. (2023). *Information technology: Let's chat about ChatGPT*. UBS.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

[2] Royal Swedish Academy of Sciences. (2024). *Scientific background: Computational protein design and protein structure prediction*. Royal Swedish Academy of Sciences.

[3] Huang, L., Yu, W., Ma, W., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 42:1–42:55.

[4] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

[5] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March 3–10). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada, pp. 610–623).

[6] Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(2), 38.

[7] Husserl, E. (2017). *Logical investigations: Volume II: Investigations in phenomenology and the theory of knowledge, Part I*. Commercial Press. pp. 792–793, 824. (In Chinese)

[8] Merleau-Ponty, M. (2001). *Phenomenology of perception*. Commercial Press. pp. 126–144. (In Chinese)

[9] Heidegger, M. (2018). *Being and time* (2nd rev. Chinese ed.). Commercial Press. pp. 88–94. (In Chinese)

[10] Heidegger, M. (2015). *On the way to language*. Commercial Press. pp. 10, 12. (In Chinese)

[11] Mandelkern, M., & Linzen, T. (2024). Do language models' words refer? *Computational Linguistics*, *50*(3), 1191–1200.

[12] Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. arXiv:2509.04664.

[13] Dong, C. (2023). The nature and limits of artificial intelligence from the perspective of machine cognitive opacity. *Social Sciences in China*, (5), 44–66.

[14] Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1–12.

[15] Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, *36*(1), 15.

[16] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

[17] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the International Conference on Learning Representations* (Vienna, Austria).

[18] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

[19] Bai, Y., Jones, A., Ndousse, K., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.

[20] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*, 53728–53741.

[21] Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202606_9(6).0003

[22] LeCun, Y. (2022). A path towards autonomous machine intelligence. OpenReview.

[23] Ha, D., & Schmidhuber, J. (2018). World models. arXiv:1803.10122.

[24] Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology*, *20*(2), 247–268.

[25] Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

[26] Driess, D., Xia, F., Sajjadi, M. S. M., et al. (2023, July 23–29). PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, USA, Vol. 202, pp. 8469–8488). PMLR.

[27] d'Avila Garcez, A., & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, *56*(11), 12387–12406.

[28] Zagzebski, L. T. (1996). *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge University Press. pp. 114–115, 271.

[29] Sun, W. (2017). Value reflections on artificial intelligence. *Philosophical Research*, (10), 120–126.

[30] Liu, Y., & Wang, C. (2023). Human-machine relations in the age of intelligence: Toward a selectionist theory of technological control. *Global Media Journal*, *10*(3), 5–21.

[31] Pang, Z., Xue, L., & Liang, Z. (2022). Artificial intelligence governance: Cognitive logic and paradigm transcendence. *Science of Science and Management of Science and Technology*, *43*(9), 3–18.

[32] High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission.

[33] Mora-Cantallops, M., Sanchez-Alonso, S., Garcia-Barriocanal, E., & Sicilia, M. A. (2021). Traceability for trustworthy AI: A review of models and tools. *Big Data and Cognitive Computing*, *5*(2), 20.