The Phenomenon of 'Concept Drift' in AI-Generated Content and Its Ontological Roots

Zuoxiu Zheng[1] and Jiacheng Wu*[2]

[1,2]College of Computer Science and Technology,Zhejiang University,Hangzhou,China

Abstract: Against the background of the rapid development of large language models, the phenomenon of 'hallucination' has become a central obstacle to AI credibility. This paper argues that AI hallucination is not merely a technical failure, but a systematic form of 'concept drift' that occurs when AI processes human knowledge systems. From the perspective of the RID cognitive-dynamics model in *Knowing and Saying*, the stability of human concepts derives from anchoring in survival pressure (D) and embodied experience (I), whereas AI lacks an ontological ground. Its generated linguistic symbols (R), once detached from concrete contexts, are therefore highly prone to semantic slippage and alienation. The paper analyzes the internal mechanisms of concept drift and its destructive consequences in high-risk fields such as medicine and law. It argues that human-AI collaboration must uphold the regulatory principle that 'humans bear conceptual responsibility, while AI provides statistical reference', in order to reconstruct safe boundaries for human-machine cognition.

Keywords: Large Language Models; Hallucination; Concept Drift; Ontological Roots; Cognitive Compensation; Symbol Grounding; Problem Pressure

1 Introduction: A Paradigm Shift from "Technical Hallucination" to "Concept Drift"

When discussing the generative mechanism of large language models (LLMs), the first problem we encounter is their widely observed phenomenon of "hallucination." Conventional natural language processing (NLP) research tends to define hallucination as the generation of text that is inconsistent with the input source or with facts in the real world. Although this definition is operationally useful in engineering, it conceals a deeper philosophical problem behind hallucination. When we carefully examine AI-generated texts that appear reasonable but cannot withstand scrutiny, we find that the issue is not simply one of "factual error," but one of disorder in "concept use." When AI processes complex and highly context-dependent concepts, it often strips them from their original networks of meaning and recombines them into a "conceptual facade," thereby causing the intension of concepts to slide and disintegrate. This phenomenon is what this paper calls "concept drift."

Reconceptualizing "hallucination" as "concept drift" marks a paradigm shift in our understanding of AI-generated content. It requires us to move beyond the level of technical optimization and to seek answers in the ontological roots of concept formation. Why can humans maintain relative stability and accuracy when using concepts, whereas AI is prone to endless semantic drift? Behind this question must lie a fundamental difference between human and machine cognition. This paper attempts to introduce the theory of "cognitive compensation" in *Knowing and Saying* in order to provide a systematic philosophical explanatory framework.

The rise of large language models, especially the GPT series, BERT, and their variants, marks a breakthrough in artificial intelligence in the field of natural language processing. Through unsupervised pretraining on massive textual data, these models have learned statistical regularities and rich patterns of language. Yet precisely this statistically based pattern-matching mechanism plants the seeds of hallucination. In the technical literature, hallucination is usually treated as a deviation in the decoding process or as a result of noise and bias in training data. To address this problem, researchers have proposed various methods, such as retrieval-augmented generation (RAG) with external knowledge bases, improved decoding algorithms, and reinforcement learning from human feedback (RLHF).

Although these technical measures alleviate hallucination to some extent, they do not touch the core of the problem. Hallucination is not merely a mismatch between model output and facts; it is also a loss of control at the conceptual level. Human language is not merely an arrangement and combination of symbols, but a carrier of meaning. Behind every concept lie human cognition, experience, and value judgment concerning the world. When AI generates text, it is in fact performing symbolic operations without a foundation of meaning. Such operations may conform to grammatical

* Corresponding author: 1320389424@qq.com

rules on the surface, but at the deeper semantic level they are often fractured and misplaced.

We therefore need to reexamine AI hallucination from a philosophical perspective. Understanding hallucination as "concept drift" helps reveal the fundamental limitation of AI in language processing. Concept drift is not merely a technical term; it is also a philosophical concept that points to AI's incapacity in the construction of meaning. The following sections define the phenomenal essence of concept drift, trace the ontological anchoring of human concepts, analyze the logic by which concept drift occurs in AI-generated content, and demonstrate its destructive consequences through concrete case analysis. Finally, the paper explores how to respond to this challenge by reconstructing the boundary of human-machine cognition.

2 Phenomenal Definition: "Concept Drift" in AI-Generated Content

2.1 The Appearance of "Hallucination" and the Essence of "Concept Drift"

To understand "concept drift," we must first deconstruct the appearance of "hallucination." In AI research, hallucination is usually divided into two types: intrinsic hallucination and extrinsic hallucination. Intrinsic hallucination refers to generated content that contradicts the conditions of the input; extrinsic hallucination refers to generated content that has no factual basis in the external world [9, 10]. This classification, however, focuses mainly on the truth value of propositions while neglecting the state of the basic units that constitute propositions, namely concepts. In fact, many extrinsic hallucinations are not invented out of nothing, but arise because a model mistakenly borrows a concept from one domain and forcibly grafts it onto an incompatible context.

For example, in medical consultation scenarios, LLMs may confuse the concept of the "immune system" with the concept of a "computer antivirus system," generating absurd advice such as "enhancing human immunity by updating the virus database." Here, the model has not completely invented a new term. Rather, it has undergone "concept drift": it strips "immunity" and "virus database" from their respective ontological roots in biology and computer science and splices them together only according to their statistical co-occurrence frequencies in the training corpus. This drift not only damages the logical structure of knowledge, but also dissolves the seriousness and precision of concepts.

Further analysis shows that the appearance of hallucination conceals a systemic defect in AI's handling of concepts. When humans use concepts, they can dynamically adjust their intension and extension according to context, and this adjustment is based on deep understanding of the world and rich experience. AI, by contrast, can only rely on the statistical distribution of concepts in training data. When confronted with novel or complex contexts, AI often cannot accurately grasp the boundaries of a concept's applicability, thereby causing abuse and misplacement of concepts. This phenomenon is particularly evident in cross-domain knowledge generation, where models frequently piece together concepts from different fields in a rigid way, producing text that seems profound but is actually meaningless.

2.2 Specific Manifestations of "Concept Drift"

Concept drift in AI-generated content appears in several concrete forms, mainly contextual misplacement, intensional generalization, and logical disconnection.

Contextual misplacement refers to the rigid transplantation of a specialized concept from its proper context into everyday contexts or other professional contexts. For instance, the "uncertainty principle" in quantum mechanics may be abused to explain the "unpredictability of interpersonal relationships" in sociology. Such misplacement not only causes the original meaning of the concept to be lost, but may also produce serious misunderstandings. In the generation of legal documents, if a model confuses the everyday sense of "intent" with the criminal-law concept of "intent," the consequences may be disastrous.

Intensional generalization refers to a model's excessive expansion of a concept's extension in pursuit of textual coherence and fluency, causing the concept to lose its original precision. For example, all forms of "disagreement" may be generalized as "cognitive bias." Such generalization makes concepts empty and deprives them of the capacity to distinguish between different things. In academic writing, frequent use of generalized concepts by a model severely weakens the rigor and persuasiveness of argumentation.

Logical disconnection refers to a model's inability to maintain the consistency of the same concept across different paragraphs in long-text generation, resulting in contradiction. Because the attention mechanism of LLMs mainly focuses on local context, when a text becomes long, the model often "forgets" how a concept was defined or used earlier, and later gives it an entirely different meaning. This logical disconnection not only damages the overall coherence of the text, but also exposes AI's incapacity to maintain conceptual identity.

These manifestations point together to a core problem: AI's grasp of concepts is "flat" and "surface-level." It has only mastered the distributional patterns of the signifier in vector space, but cannot reach the anchoring point of the signified in the real world. Such symbolic operation without ontological roots inevitably produces disordered semantic slippage.

2.3 Concept Drift and the "Stochastic Parrots" Thesis

As Luciano Floridi pointed out in his analysis of GPT-3, such models are essentially powerful "syntactic engines" rather than "semantic engines" [6]. They display striking fluency at the syntactic level, but at the semantic level they lack reference to the real world. In discussions of the generative mechanism of large language models, the "stochastic parrots" thesis proposed by Bender et al. is an unavoidable point of reference [1]. This thesis holds that LLMs merely predict the next word according to the statistical distribution in the

training corpus. They have no intentionality and do not understand what they generate, as the metaphor of stochastic parrots suggests.

The concept of "concept drift" is internally consistent with the "stochastic parrots" thesis in criticizing the semantic emptiness of LLMs, but it advances the critique further in an ontological dimension.

First, the "stochastic parrots" thesis mainly remains at the level of behavioral description, indicating what LLMs do, namely probability-based textual stitching. By contrast, "concept drift" attempts to explain why this occurs. From the perspective of the RID model, the fundamental reason why LLMs become "stochastic parrots" is that they lack the drive of "problem pressure" (D). Without the urgency of survival pressure, a system cannot generate genuine concern for the external world, and its generated symbolic sequences naturally cannot find firm anchors in reality.

Second, the "stochastic parrots" thesis often treats LLM output as a kind of meaningless noise, whereas "concept drift" emphasizes the structural damage such output may cause within human cognitive networks. When AI-generated "conceptual facades" are received and internalized by uninformed human users, these drifting concepts can spread through human knowledge systems and gradually erode the determinacy and authority of existing concepts. Concept drift is therefore not only a description of AI's generative mechanism, but also a profound warning about deterioration in the cognitive ecology of human-machine interaction.

Finally, concept drift theory argues that even if we use technical measures such as RAG to allow this system to consult an encyclopedia, it still cannot escape the fate of concept drift. The introduction of external knowledge bases merely increases the material available for symbolic splicing; it does not give the system the capacity to "bear" concepts. As long as the system remains pure computation detached from ontological roots, its use of concepts will always be drift without responsible guarantee.

3 Theoretical Genealogy: The Ontological Anchoring of Human Concepts

To understand deeply why AI undergoes concept drift, we must look back at how human concepts remain stable. *Knowing and Saying* proposes that the human cognitive structure does not exist a priori, but is generated as a form of ontological compensation. In this framework, "problem pressure" (D) is the starting point of cognition. When finite beings confront challenges in their living environment, they must construct stable concepts in order to grasp the order of the world and guide practical action. Accordingly, every core human concept is deeply anchored in specific survival pressures and problem situations [20].

3.1 Anchoring in Survival Pressure: Problem Pressure (D) as the Genetic Starting Point of Concepts

In *Being and Time*, Martin Heidegger profoundly disclosed the existential structure of Dasein, pointing out that human understanding of the world always arises from our Sorge and Besorgen in being-in-the-world [8]. Such existential concern is precisely the soil in which concepts arise. For example, for early humans, the concept of "fire" was not merely the name of a physical phenomenon, but also an answer to matters of survival such as heating, cooking, and protection from wild animals. This powerful "problem pressure" (D) gave the concept of fire an indelible ontological weight. When we use the concept of fire, we are not merely invoking a symbol, but awakening a dense survival experience. By contrast, when AI processes the symbol "fire," it has no survival pressure at all. It is merely calculating the probability distribution of character sequences. The lack of anchoring in problem pressure (D) is the first ontological root of AI concept drift.

Throughout human history, the birth of every important concept has been accompanied by specific survival challenges. From "solar terms" in agrarian society to "efficiency" in industrial society and "data" in information society, these concepts are cognitive tools created by humans to respond to environmental change and optimize survival strategies. They are not merely descriptions of the objective world, but projections of human subjectivity into the world. When humans use these concepts, they engage in practice filled with purposiveness and value orientation. AI, however, is a pure computational system. All its operations are based on preset optimization objectives, such as minimizing a loss function. It has no survival needs of its own and no genuine concern for the world. Its generated concepts are therefore necessarily hollow and lacking in vitality.

3.2 Anchoring in Embodied Experience: Sensorimotor Constraints on Semantics

Bisk et al. propose that "experience grounds language," emphasizing that linguistic meaning must be established through multimodal physical interaction and embodied experience [2]. Beyond the macro-level anchoring of survival pressure, human concepts are also strictly constrained by micro-level embodied experience. The embodied cognition tradition in cognitive science holds that the human mind and conceptual system are deeply rooted in the physical properties of the body and in sensorimotor experience [13]. Our understanding of basic spatial and physical concepts such as "high," "low," "front," "back," "heavy," and "light" derives directly from our bodily interaction with the gravitational environment. These metaphors based on embodied experience further constitute the basis on which we understand abstract concepts.

Embodied experience provides concepts with a physical anchoring that cannot be violated at will. We cannot arbitrarily drift the concept of "heavy" into the meaning of "flying upward," because doing so violates our most basic bodily intuition. Yet for LLMs, which have no body and no sensory organs, all words are merely floating-point vectors in a high-dimensional space. In their "world," "heavy" and "light" have no essential physical difference; they are simply two points at different distances in a corpus. Because AI lacks the constraints of embodied experience, it is highly prone

to absurd semantic slippage when generating metaphors or handling physical concepts.

Embodied cognition theory emphasizes that cognition does not occur in an isolated and abstract computational space, but in the dynamic interaction between an embodied entity and its environment. The human sensorimotor system not only provides initial information about the world, but also shapes the way we organize and understand that information. For example, our understanding of "balance" derives from the experience of maintaining bodily uprightness in a gravitational field; our understanding of "container" derives from physical operations of putting objects in and taking them out. These embodied experiences constitute the underlying semantic structure of concepts, enabling concepts to maintain basic consistency across contexts. Because AI lacks this underlying embodied anchoring, its handling of concepts can only remain at the surface level of symbolic combination.

3.3 Anchoring in Social Interaction: Meaning Calibration in Public Language Games

Finally, the stability of human concepts also benefits from continuous calibration in social interaction. In his later philosophy, Ludwig Wittgenstein proposed that "meaning is use," emphasizing that the meaning of language is established in specific language-games and forms of life [17]. When a person uses a concept incorrectly in a community, they immediately receive correction and feedback from others. This public negotiation mechanism based on intersubjectivity functions like an invisible network that holds each concept in place and prevents arbitrary drift.

Although AI seems to have "learned" the rules of language games through training on massive human corpora, it does not truly participate in forms of life. It cannot feel the social consequences of using a concept incorrectly, such as being mocked, misunderstood, or punished. During inference and generation, LLMs are closed, one-directional output systems that lack real-time social feedback loops based on lived situations. Once they deviate from the correct semantic track during generation, there is no mechanism that can pull them back in time, and concept drift is therefore amplified.

Social interaction not only provides a calibration mechanism for meaning, but also gives concepts normativity and binding force. In human society, concept use is constrained by various explicit and implicit rules. In an academic community, for example, the use of a specific concept must follow strict definitions and argumentative logic; in everyday communication, concept use must conform to social customs and moral norms. These rules constitute the "grammar" of concept use and ensure the effectiveness and reliability of linguistic communication. AI, as a system detached from social norms, follows only statistical probability when generating text. This lack of normative constraint makes AI highly prone to crossing moral and logical boundaries when handling sensitive or complex concepts, producing unpredictable consequences.

4 Mechanism Analysis: The Logic of "Concept Drift" in AI-Generated Content

4.1 The Suspension of Symbols: Statistical Association without Real-World Reference

Based on the above analysis of the ontological anchoring of human concepts, we can more clearly analyze the logic by which AI concept drift occurs. The first mechanism is "symbol suspension," or symbol grounding failure. In cognitive science, the symbol grounding problem asks how formal symbol systems can acquire intrinsic semantic reference [7, 9, 20]. The training data of LLMs consist of pure textual corpora stripped of ontological roots. In these corpora, symbols point only to other symbols, forming a vast, self-referential network of statistical associations.

This suspended condition means that AI's grasp of concepts depends entirely on the distributional hypothesis, according to which words that frequently appear together in context have similar meanings. Although this approach is highly effective in capturing shallow semantic relations among words, it cannot establish a real connection between symbols and external physical entities or internal human mental states. When a model confronts tasks that require deep semantic understanding or commonsense reasoning, the fragility of purely statistical association is fully exposed, and the intension of concepts drifts freely in a vector space without real-world constraints.

John Searle's Chinese Room thought experiment vividly reveals the essence of symbol suspension [14]. In a closed room, a person who does not understand Chinese can consult a detailed rulebook and convert Chinese symbols as input into Chinese symbols as output, leading those outside the room to believe that he understands Chinese. Yet the person in fact does not understand the meaning of any Chinese character; he is only performing mechanical symbolic operations. Large language models are like extremely complex and efficient Chinese Rooms. Through enormous neural networks and massive parameters, they have mastered the statistical distribution patterns of linguistic symbols, but they do not truly understand the concepts represented by those symbols [16]. This symbolic operation without semantic foundation is the fundamental cause of concept drift.

4.2 Flattening of Context: Compression of Multidimensional History and Culture

Marshall McLuhan famously asserted that "the medium is the message," emphasizing that media technologies not only transmit content but also profoundly reshape human sensory ratios and cognitive structures [12]. As a new "meta-medium," the large language model is reshaping the generation and circulation of concepts through its flattening treatment of context. The second key mechanism causing concept drift is the flattening of context. Human concepts are highly context-sensitive. The same word often has very different intensions in different historical periods, cultural backgrounds, or professional domains. For example, "freedom" means different things

in political science, economics, and philosophy. Humans can dynamically select the appropriate conceptual intension according to the current problem situation (D) and context-activation framework (I).

During pretraining, however, LLMs compress heterogeneous corpora from the internet, spanning different eras, cultures, and fields, into the same high-dimensional vector space. This compression inevitably causes the flattening of context and loss of information. Although the attention mechanism of the Transformer architecture can capture contextual information to some extent, it still cannot fully restore the multi-dimensional historical and cultural context behind concepts. As a result, when the model generates text, it often stitches together conceptual intensions from different contexts, producing a confused conceptual patchwork that seems profound but is in fact disordered.

Context flattening not only confuses conceptual intensions, but also weakens the model's capacity to handle complex semantic relations. In human language, the meaning of many concepts depends on implicit background knowledge and common sense. When we say "bank," for example, we know not only that it is a financial institution, but also that it is closely related to deposits, loans, and interest. This background knowledge constitutes a rich semantic network for the concept. In the vector space of LLMs, however, such background knowledge is often compressed into simple vector distances. When generating text, the model can only combine words according to these surface vector distances, and cannot engage in deep semantic reasoning. This surface-level semantic processing makes the model highly prone to conceptual misplacement and drift when facing tasks that require judgment based on the synthesis of many kinds of information.

4.3 The Probabilistic Nature of Generation: The Inherent Defect of Maximum Likelihood Estimation

Zhang Haojun has analyzed the two dimensions of AI epistemology, arguing that current AI technologies rely excessively on statistical fitting while seriously lacking causal explanation [19]. Finally, AI concept drift is deeply rooted in the mathematical foundation of its generative mechanism: maximum likelihood estimation (MLE). Text generation by LLMs is essentially an autoregressive probabilistic prediction process, in which the model predicts the next most likely token according to the preceding sequence. This probability-maximizing strategy tends to select word combinations that occur frequently in the training corpus and are ordinary in collocation.

This mechanism has two consequences. The first is mediocrity: models tend to generate cliches and lack genuine originality and depth of thought. The second is hallucinatory drift: in order to maintain local coherence and high probability, the model sacrifices global logical consistency and conceptual accuracy. In long-text generation, as the number of prediction steps increases, errors accumulate continuously. To "cover" earlier errors, the model has to introduce more irrelevant concepts, eventually causing serious concept drift and semantic collapse. This probabilistic generative mechanism determines that AI can never use concepts rigorously in the way humans do, on the basis of solid logical inference and value judgment.

In addition, an MLE-based generative mechanism makes models highly susceptible to prompt interference. Because the model's objective is to maximize the probability of a word sequence under a given context, minor changes in the input prompt may lead to major changes in the output. This fragility is especially prominent in the handling of sensitive or controversial topics. Certain keywords in a prompt may trigger specific distributions in the training corpus, causing the model to generate highly biased or entirely factually false text. Such generation governed by probability distribution further intensifies conceptual instability and drift.

5 Case Analysis: The Destructive Consequences of "Concept Drift"

5.1 Conceptual Slippage in the Generation of Legal Judgments

To show more directly the destructive character of concept drift, we can examine the application of AI to the generation of legal judgments. Legal concepts such as "intent," "negligence," and "causation" have extremely strict intensions and boundaries of application, and they directly concern the life, liberty, and property rights of the parties involved. In some test cases, LLMs have frequently produced serious conceptual slippage when drafting judgments. For example, they drift from "indirect intent" to "negligence due to carelessness," or confuse "factual correlation" with "legal causation" when discussing causality.

The root of this concept drift is that AI cannot experience the heavy "problem pressure" (D) behind legal concepts, namely the ultimate concern for social justice and individual rights. It merely imitates the linguistic style of legal documents, or the conceptual facade, and fills textual gaps through high-probability word combinations. If such symbolic operation without ontological roots is blindly applied in judicial practice, it will cause devastating damage to the seriousness and fairness of the rule of law.

More specifically, legal reasoning is a highly structured and logically rigorous process. When applying legal concepts, judges must comprehensively consider the wording of statutes, legislative purpose, precedents, and the concrete facts of the case. This is a complex process filled with value judgment and balancing of interests. AI, however, merely performs pattern matching and probabilistic prediction when generating judgments. It cannot understand the fundamental difference between "intent" and "negligence" in the degree of moral blameworthiness, nor can it grasp the central place of "causation" in principles of attribution. When it reduces legal concepts with dense value implications to purely statistical symbols, it inevitably produces ruptures in legal logic and absurd adjudicative results.

5.2 Terminological Misplacement in Medical Diagnosis

In medicine, concept drift likewise harbors enormous risks. Medical terms such as "malignant tumor" and "autoimmune disease" are precise descriptions of pathological states of the human body, behind which lie complex biological mechanisms and rich clinical experience. When a patient describes symptoms to AI, LLMs may detect certain superficially similar keywords and produce terminological misplacement. For example, they may drift from ordinary gastrointestinal inflammation to a rare systemic immune syndrome and then offer absurd treatment advice.

Here the mechanism of context flattening intensifies the problem. AI cannot, like a human physician, integrate multidimensional embodied information such as the patient's physical signs, medical history, and laboratory test results in order to make a differential diagnosis. It can only search for the highest-probability lexical match in a flat textual space. Once patients trust this concept drift stripped of clinical embodied experience and concern for life, or problem pressure (D), it may delay treatment or even endanger life.

Medical diagnosis is not only a science but also an art. When making diagnoses, physicians rely not only on objective medical knowledge, but also on clinical intuition accumulated over time and sensitive insight into the patient's overall condition. This embodied and situated diagnostic process cannot be simulated by AI [11]. When AI attempts to replace complex medical reasoning with a flat statistical model, it inevitably simplifies multidimensional pathological features into one-dimensional lexical mappings, thereby seriously distorting and misusing medical concepts.

5.3 Knowledge Pollution in Scientific Research

Beyond law and medicine, concept drift may also have serious negative effects in scientific research. As AI becomes widely used in literature review, data analysis, and academic writing, more and more researchers are beginning to rely on LLMs to assist scientific work. Yet because AI lacks understanding of the real intension of scientific concepts, it often produces subtle concept drift when generating scientific texts, thereby polluting scientific knowledge.

For example, when explaining complex physical phenomena or biological mechanisms, AI may use inaccurate metaphors or confuse closely related scientific terms. Such seemingly minor deviations may trigger chain reactions in scientific argumentation and lead to erroneous final conclusions. More seriously, when AI-generated texts containing concept drift are widely disseminated and cited by other researchers, they gradually infiltrate human knowledge systems and create long-term, hard-to-detect knowledge pollution. Such pollution not only misleads later scientific research, but may also weaken public trust in science.

5.4 "Pseudo-Professionalism" and Responsibility Vacuum in Medical Diagnosis

In the medical field, great expectations have been placed on large language models, and many start-ups have attempted to develop LLM-based intelligent consultation systems. Yet medical concepts such as "inflammation," "tumor," and "autoimmune disease" are not only objective descriptions of physiological states, but also high-risk judgments concerning patients' lives and health.

In one test case, researchers entered into a well-known medical LLM a complex description of a patient's symptoms, including intermittent fever, joint pain, and abnormalities in specific blood indicators. The LLM quickly generated a detailed diagnostic report, concluding that the patient had systemic lupus erythematosus and offering apparently professional treatment suggestions. The report cited several medical documents and clinical guidelines, and its language resembled that of an experienced physician.

However, when a rheumatology and immunology expert reviewed the report, the expert found that the LLM had undergone serious concept drift in its understanding of the key concept of "antinuclear antibody (ANA) titer." In the cognition of human physicians, ANA titer is not merely a number; it must be judged comprehensively in relation to the patient's age, sex, medical history, and other clinical manifestations, namely structural generation (I). A mildly elevated titer may be a normal physiological phenomenon in a healthy elderly person, but it may be an early sign of disease in a young woman.

The LLM made the wrong diagnosis because its training corpus contains a strong statistical correlation between "elevated ANA titer" and "systemic lupus erythematosus." It flattened the complex and highly contextualized medical concept of ANA titer into a simple statistical trigger. Because it does not face the "problem pressure" (D) that a misdiagnosis may delay treatment or expose a patient to drug side effects, its "diagnosis" is only a probability output without responsibility.

This medical form of concept drift is extremely dangerous. Wearing the appearance of professionalism, it can easily mislead ordinary patients who lack medical knowledge and may even cause inexperienced junior physicians to lower their guard. More seriously, when an AI-generated misdiagnosis causes a medical accident, the traditional system for determining liability in medical harm faces a major challenge: how can an algorithmic model with no subjectivity and no capacity to feel problem pressure bear legal responsibility for medical negligence? This again highlights the crisis of responsibility vacuum brought by concept drift.

6 Further Discussion: Cognitive-Scientific and Neurobiological Perspectives on Concept Drift

6.1 The Contrast between Neuroplasticity and Static Weights

The conceptual network of the human brain is dynamic and highly neuroplastic. When we learn new concepts or use old concepts in different contexts, synaptic connections in the brain undergo real-time physical change. This dynamic remodeling mechanism ensures that human concepts can adapt

to changing environments and experiences. By contrast, the parameters, or weights, of a large language model are frozen after training during the inference stage. They cannot dynamically adjust the internal structure of concepts according to real-time feedback and new experience in the way the human brain does. This static weight network determines that, when AI faces novel or complex contexts, it can only rely on existing probability distributions for mechanical splicing, which easily causes concept drift.

6.2 Cognitive Load and the Allocation of Computational Resources

In human cognition, concept use is constrained by cognitive load. When humans engage in complex conceptual reasoning, they consume substantial mental resources and undergo significant metabolic cost. This resource constraint forces humans to maintain a high degree of focus and rigor when handling concepts, avoiding meaningless semantic slippage. Although AI computation also consumes enormous computing power, this consumption is essentially different from cognitive load. When AI generates text, its computational resources are allocated uniformly and without discrimination. It does not invest more "attention" in a concept because that concept is complex or important. This lack of cognitive focus and priority allocation makes AI seem careless when handling core concepts and makes arbitrary drift more likely.

6.3 Intentionality and Unintentional Symbolic Operation

Duan Weiwen argues, in his discussion of machine intentionality, that although artificial intelligence can display purposive behavior resembling that of humans, this "derived intentionality" cannot be equated with human "original intentionality" and therefore cannot bear moral responsibility [4]. Philosophers Franz Brentano and Edmund Husserl emphasized that human consciousness has intentionality, meaning that consciousness is always about something. When humans use concepts, they always have definite intentions and purposes; concepts are tools through which humans point toward and grasp the world. AI, however, is an unconscious physical system, and its symbolic operations entirely lack intentionality. Although its generated text appears meaningful on the surface, that meaning is assigned by human readers, not expressed by AI's own intention. Because AI lacks internal intentionality, it has no definite direction or aim when combining symbols and can only follow probability. This is the ultimate philosophical cause of concept drift [15].

7 Regulatory Paths: Reconstructing the Boundary of Human-Machine Cognition

7.1 Epistemological Sobriety: Rejecting the Trap of Anthropomorphism

Xu Yingjin points out that generative artificial intelligence has fundamental epistemological defects, and that its statistical prediction based on big data cannot be equated with human rational cognition based on causal relations [18]. In the face of concept drift in AI-generated content, we must first remain epistemologically sober and reject anthropomorphism. We must recognize that AI-generated text is only the result of probabilistic matching, not an expression of genuine understanding. When AI uses concepts such as "I," "think," and "understand," it has not undergone the psychological states corresponding to these concepts. Therefore, when reading and using AI-generated content, we must always maintain critical suspicion and must not blindly trust its surface logical coherence and professionalism.

7.2 Technical Constraints: Introducing Knowledge Graphs and Logical Reasoning Modules

At the technical level, in order to mitigate concept drift, we can consider introducing stricter semantic constraint mechanisms into LLMs. For example, knowledge graphs can be deeply integrated with large models, using explicit entity relations and logical rules in the graph to restrict semantic slippage during generation. In addition, symbolic reasoning modules can be embedded into neural networks to construct neurosymbolic systems, so that when models combine concepts, they follow not only probability distributions but also hard constraints of formal logic.

7.3 Institutional Regulation: Establishing the Principle of Human-in-the-Loop

Facing what Floridi calls the "Fourth Revolution," namely the comprehensive reshaping of human reality by the infosphere, we must reexamine the ethical boundaries of human-machine relations [5]. At the institutional level, we must establish the core principle of human-in-the-loop. Especially in high-risk fields such as law, medicine, and finance, tasks involving core conceptual judgments and value decisions must never be completely handed over to AI. AI can only serve as an auxiliary tool, providing preliminary textual drafts or data analysis. The final confirmation of concepts, logical review, and responsibility must be undertaken by human experts with professional knowledge and ontological roots. At the same time, a strict algorithmic accountability mechanism should be established, requiring developers and deployers of AI systems to bear corresponding legal responsibility for serious concept drift and misleading consequences caused by generated content. Chen Xiaoping emphasizes that the core of solving AI ethical problems lies in embedding ethical principles into the entire process of system design and application, thereby realizing deep integration of technology and ethics [3].

7.4 Educational Response: Improving Public AI Literacy

Finally, at the educational level, it is urgent to improve public AI literacy. We should not only teach students how to use AI tools, but also cultivate their ability to identify concept drift and logical fallacies in AI-generated content. The focus of education should move from memorizing and repeating knowledge to critical thinking, interdisciplinary synthesis, and deep value judgment. Only when humans themselves possess strong conceptual mastery and a solid cognitive foundation can they maintain subjectivity in collaboration with

AI and avoid being swallowed by algorithmically generated conceptual facades.

8 Conclusion

In summary, hallucination in AI-generated content is not merely a technical failure, but a concept drift that necessarily occurs once AI is detached from human ontological roots. By introducing the theoretical perspective of *Knowing and Saying*, this paper has shown that human concepts remain stable because they are deeply anchored in survival pressure (D), embodied experience, and social interaction. AI, as a pure symbol-processing system, inevitably falls into disordered semantic slippage because of symbol suspension, context flattening, and the probabilistic nature of its generative mechanism.

This conclusion has important philosophical and practical significance. Philosophically, it reminds us that we must clearly recognize the essential difference between human and machine cognition. AI can simulate human linguistic output and may even surpass human computational capacity in certain tasks, but it can never replace humans in bearing the survival weight and value meaning behind concepts. Practically, it requires us to establish a strict human-in-the-loop mechanism when applying generative AI. We cannot completely entrust to AI core tasks involving value judgment, causal reasoning, and responsibility, that is, tasks involving strong problem pressure (D). AI can only serve as an auxiliary tool that provides probabilistic reference. The final "conceptual confirmation" and "meaning assignment" must be completed by human subjects with ontological roots. Only on the basis of this clear philosophical boundary can safe and trustworthy human-machine collaboration truly be achieved.

Future research needs to explore further how technical architectures can better simulate the anchoring mechanism of human concepts, for example by using embodied AI and multimodal learning to mitigate symbol suspension. This requires us to go beyond the purely textual modality and integrate visual, auditory, tactile, and other multidimensional perceptual information into model training, thereby establishing a richer experiential basis. At the same time, interdisciplinary cooperation across philosophy, cognitive science, computer science, and law will be the necessary path for solving concept drift and building responsible artificial intelligence. While pursuing technological progress, we must maintain respect for the ontological foundations of human beings and ensure that artificial intelligence remains a tool for enhancing human cognition and serving human welfare, rather than a blind force that dissolves the human world of meaning. In the face of the challenge of concept drift brought by AI-generated content, we need neither fall into technological pessimism nor blindly embrace technological utopianism. Only under the guidance of philosophical reflection and with clear boundaries for human-machine cognition can we preserve the dignity of knowledge and the weight of meaning in the intelligent age.

References

[1] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2021, pp. 610–623.

[2] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, N. Pinto, and J. Turian, "Experience grounds language," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 8718–8735. [Online]. Available: https://aclanthology.org/2020.emnlp-main.703/

[3] X. Chen, "Ethical problems of artificial intelligence and their solution paths," *Philosophical Research*, no. 1, pp. 116–125, 2021, english title translated from the source manuscript; metadata not independently confirmed from an open authoritative record.

[4] W. Duan, "Machine intentionality and the moral status of artificial intelligence," *Philosophical Trends*, no. 5, pp. 98–105, 2019, english title translated from the source manuscript; metadata not independently confirmed from an open authoritative record.

[5] L. Floridi, *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Hangzhou: Zhejiang People's Publishing House, 2016, chinese translation by Wang Wenge cited in the source manuscript; original English edition published by Oxford University Press in 2014.

[6] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020.

[7] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1–3, pp. 335–346, 1990.

[8] M. Heidegger, *Being and Time*. Beijing: SDX Joint Publishing Company, 2006, chinese translation by Chen Jiaying and Wang Qingjie cited in the source manuscript; original German work published in 1927.

[9] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023. [Online]. Available: https://arxiv.org/abs/2311.05232

[10] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[11] J. Li, "Embodied cognition and the developmental predicament of artificial intelligence," *Studies in Philosophy of Science and Technology*, vol. 37, no. 2, pp. 1–6, 2020, english title translated from the source manuscript; metadata not independently confirmed from an open authoritative record.

[12] M. McLuhan, *Understanding Media: The Extensions of Man*. Nanjing: Yilin Press, 2011, chinese translation by He Daokuan cited in the source manuscript; original English work published by McGraw-Hill in 1964.

[13] M. Merleau-Ponty, *Phenomenology of Perception*. Beijing: Commercial Press, 2001, chinese translation by Jiang Zhihui cited in the source manuscript; original French work published in 1945.

[14] J. R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–424, 1980.

[15] J. R. Searle, *Mind, Language and Society: Philosophy in the Real World*. Shanghai: Shanghai Translation Publishing House, 2001, chinese translation by Li Bulou cited in the source manuscript; original English edition published by Basic Books in 1998.

[16] T. Wang, "Semantic understanding in large language models and its philosophical limits," *Philosophical Analysis*, vol. 15, no. 1, pp. 112–124, 2024, english title translated from the source manuscript; metadata not independently confirmed from an open authoritative record.

[17] L. Wittgenstein, *Philosophical Investigations*. Shanghai: Shanghai People's Publishing House, 2001, chinese translation by Chen Jiaying cited in the source manuscript; original work first published in 1953.

[18] Y. Xu, "The epistemology of big data and the ethical challenges of generative artificial intelligence," *Studies in Dialectics of Nature*, vol. 39, no. 8, pp. 45–51, 2023, english title translated from the source manuscript; metadata not independently confirmed from an open au-

thoritative record.

[19] H. Zhang, "Statistical fitting and causal explanation: Two dimensions of artificial intelligence epistemology," *Journal of Dialectics of Nature*, vol. 44, no. 6, pp. 33–39, 2022, english title translated from the source manuscript; metadata not independently confirmed from an open authoritative record.

[20] X. Zhang, "Knowing and saying: An ontological investigation of human cognition," PSSXiv (Philosophy and Social Sciences Preprint Server), may 2026, pSSXiv:202605.04152V1; CSTR:32012.36.PSSXiv.202605.04152. [Online]. Available: https://zsyyb.cn/abs/202605.04152