COMPREHENDING THE EPISTEMIC HARMS IN GENERATIVE ARTIFICIAL INTELLIGENCE

Toluwalase Ajayi*

Abstract

This paper investigates the knowledge-related harms associated with generative AI and contends that this technology has significantly transformed and will keep altering our lives. The extensive usage of generative AI and its strong incorporation into the economy and society have enhanced effectiveness and generated advantages. However, it has also inevitably affected the prevailing social structure and raised concerns regarding knowledge. This study utilizes a qualitative approach via a literature review to examine the knowledge-based damages in generative AI in form of LLMs. It suggests that the intricate internal mechanisms of generative AI intelligence models complicate the understanding of how they reach conclusions, leading to worries about accountability and the possibility of mistakes or prejudices. The caliber of data used to train generative AI completely dictates its performance, as success relies on the data learned. Therefore, generative AI systems align with John Locke's concept of tabula rasa, which begin without any pre-existing knowledge and depend solely on input information to acquire knowledge and make choices. Nevertheless, poor quality data can lead to erroneous outcomes generative AI, similar to flawed experiences, can result in erroneous human perceptions, which may cause adverse outcomes and knowledge-related challenges. For example, prejudiced training data can produce biased outputs from generative AI, creating epistemic issues. This can facilitate the dissemination of false information and sway beliefs by generating persuasive yet inaccurate content. The paper asserts that establishing epistemic principles and frameworks for the responsible advancement of generative AI is essential to mitigate possible damage and guarantee its beneficial effects on knowledge systems.

Keywords: AI Epistemology, Knowledge-Related Damages, Epistemic Harms, Generative AI, LLMs.

Introduction

In recent years, Artificial Intelligence (AI) has emerged as a potent force capable of reshaping numerous facets of society. Artificial Intelligence has been characterized in various ways over time. The initial definition of Artificial Intelligence was provided by John McCarthy, Computer Science Department at Stanford University, which introduced the term in 1956. McCarthy describes it as the discipline and technology involved in creating intelligent devices. The discipline was established on the assertion that a fundamental characteristic of humans, intelligence, can be articulated with such accuracy that it can be replicated by a machine (Smith, 2006, p. 4).

This work is licensed under a Creative Commons Attribution-Non-Commercial 4.0 International License.
*Toluwalase Ajayi, Ph.D. (Senior Lecturer) Department of Philosophy Faculty of Arts, Olabisi Onabanjo University, Ago – Iwoye, Ogun-State. Nigeria. West Africa. http://ORCID.org/0000-0003-4277-279x

The field of Artificial Intelligence (AI) remains enigmatic. Indeed, Hofstadter (1985, p. 633) remarks: Even after almost 50 years of study in the discipline, there is no consensus on a definition of artificial intelligence. Artificial Intelligence is sometimes defined as the manipulation of symbols for problem solving or by tautological statements such as Artificial Intelligence refers to the aspect of computer science that focuses on creating intelligent computer applications (Waterman, 1986, p. 10). Nevertheless, models grounded in generative AI possess inherent shortcomings that could hinder their utility as instruments for facilitating responsible investigation. (Buchanan and Shortliffe, 1985).

The paper argues that the rise of Generative AI access to information has long influenced how we obtain information, generating worries regarding its epistemic harms, on cognitive biases, transmission and creation of knowledge. In this paper, we focus on how Generative artificial intelligence might need to be understood now that it has entered into conventional society. The fast evolution of Generative AI has raised important epistemic issues throughout every segment of society. The paper posits that Generative AI endangers both epistemic and social trust; for example, one could mistakenly think to be engaging with a person when, in reality, conversing with a chatbot. This can result in unwarranted confidence. Therefore, misplaced trust in the software incorrectly believed to be human can result in knowledge-related and social damage. The paper highlights the significance of epistemic principles in tackling the intricacies of epistemic harms associated with Generative AI in form of LLMs.

Research Problems

The aim of this research is to address the following issues:

a) Reveal the cognitive constraints and shortcomings in AI
b) Examine the impact of algorithmic prejudice in generative AI
c) © Grasp the limited comprehension and processing abilities in generative AI
d) Evaluate the unpredictable outcomes in generative AI
e) Examine the difficulties of interpretability in generative AI in essential applications where transparency is required
f) Assess the effects of algorithmic prejudice in generative AI
g) Examine the idea of tabula rasa within generative AI
h) Examine the discussions related to knowledge concerning simplicity, similarity, and contextual significance in generative AI
i) Evaluate the significance of epistemic principles for generative AI

Research Questions

The subsequent research questions have been formulated

a) What are the cognitive constraints and shortcomings in AI?
b) What are the impacts of algorithmic bias in generative AI?
c) What is the limited comprehension, and handling capabilities in generative AI?
d) What are the erratic results in generative AI?
e) What challenges related to interpretability in generative AI in essential applications where transparency is required?

f) What are the impacts of algorithmic prejudice in generative AI?
g) What does the idea of tabula rasa mean in generative AI?
h) What are the discussions related to knowledge concerning simplicity, similarity and contextual significance in generative AI?
i) What is the importance of epistemic principles for generative AI?

Literature Review

Generative AI, in the form of Large Language Models (LLMs), is increasingly being regarded as a support for knowledge exploration or as a cognitive instrument. LLMs are progressively being utilized in epistemic situations, specifically in scenarios where an individual employs LLM outputs to achieve their epistemic objectives, like obtaining justification or enhancing their knowledge or comprehension. Depending on a LLM system to attain epistemic objectives (experience, testimony, etc.) carries epistemic dangers. The outputs of LLMs are frequently accepted without scrutiny, which may result in a process of epistemic deskilling. Large Language Models (LLMs) present issues that can be explored from an epistemological perspective.

Floridi (2023) asserts that LLMs and humans are certainly distinct. Although human beings are remarkable "semantic engines" (their comprehension and depiction of reality is based on semantic abilities, i.e., the capability to assign meaning and significance to the world. LLM is an impressive "syntactic engine" that forecasts the most probable sequence of words that follows in a segment of writing in reply to a prompt, functioning on the capacity to handle vast quantities of data and variables through intricate frameworks, such as neural networks, without having an understanding or consciousness of the information being managed. In contrast to the scientific approach, where understanding involves credible forecasts reliant on models and understandings, outputs produced by AI do not have this kind of theoretical basis. Therefore, examining the fundamental operations of LLMs is critical because of its significant implications, respond to human behaviors and the standards for validating them.

The output generated by LLMs might turn out to be precise (when there exists an accepted benchmark or criterion to assess accuracy), but it does not rely on the capacity to assign meaning and collectively confirm its epistemic trustworthiness. From an analytical perspective, LLMs (and, more broadly, Gen AI) lack the capability to perform a semantic elevation. Or descent in relation to artificially produced sentences; namely, they are not able to move from the linguistic to the metalinguistic level to assess the validity of their assertions or advocate for its correctness.

Although epistemology emphasizes a collective method of representation and affirmation, LLMs only offer the resulting output, raising issues about reproducibility and oversight. Specifically, the notion of control is critical, as "whoever controls the questions controls the answers and whoever controls the answers controls reality" (Floridi, 2023, p 5). Variable results emphasize that the LLM's depiction of reality requires systems that guarantee accountability. Additionally, the LLM's incapacity to assign meaning or

confirm the trustworthiness of its outputs poses significant inquiries regarding its function in generating knowledge that underpins decision-making. Exploring the epistemological aspects of LLMs encompasses wider philosophical questions surrounding the processes of knowledge acquisition, representation, and validation. Grasping determining how to harmonize the outputs generated by LLMs and their depiction of reality with human understanding is essential.

Alvarado (2023) Defines AI as an "epistemic technology," where the word epistemic is intended to denote the wider range of activities that pertain directly to the gathering, preservation, application, and generation of knowledge. The epistemic characteristics of AI originate from the kind of material it processes and the kinds of actions it executes on that material, but also for "its own internal capacities. However, referring to it as an epistemic technology does not imply that the information or portrayal of reality it offers is reliable. In other terms, acknowledging AI as an epistemic technology does not necessitate that we also view it as an epistemically trustworthy technology. For these reasons, the necessity to engage in rigorous epistemic and methodological examination is even more urgently needed.

LLMs tend to elide the distinction between what is generated (i.e., which results from an appropriate creative method) and what is simply replicated and reflected (i.e., which arises from the capacity to imitate a creative task). If one focuses solely on the outcome and neglects the process, humans frequently struggle to differentiate, for example, synthetic from non-synthetic outcomes. It is essential not only to assess a body of knowledge but also to manage the process through which that knowledge is developed.

Gui and Toubi (2023) noted that LLMs challenges the human capacity to distinguish between truth and falsehood. Taking into account the statistical aspect of LLM's results, instances of incorrect information (commonly referred to as hallucination) are often produced, despite appearing credible: Readers and consumers of texts will need to adapt to the uncertainty of whether the origin is artificial or human. Likely, they will not perceive or even care. For instance, LLMs offer responses that they cannot automatically qualify in terms of their level of truthfulness or dependability. When asked about this, individuals can instinctively assess their own statements, categorizing them as certain, likely, conceivable, and so forth (utilizing phrases that come before the assertion itself, like "I am confident that," "I think that," "it is very likely that," etc.). Any response provided by an LLM should, in this context, consistently be prefaced with a phrase such as: "It is only likely that …," to indicate that any AI-generated statement cannot inherently assert any claim to truth. Long before the advent of LLMs, the concept of truth in democratic societies reveals itself as the outcome of an inclusive process of understanding.

In a society ravenous for information, and constantly generating representations, the 'truth' of these takes on a unique importance. This becomes increasingly pertinent in light of the phenomenon known as the feedback loop Large Language Models acquire from the material that is provided to them, yet as the content becomes more entirely or partially generated by AI, it will begin to learn from its own outputs, resulting in a feedback loop. The issue of epistemic trust highlights the difficulty of humans depending on the

responses and depiction of reality expressed by LLMs, which restates patterns recognized from its training data, lacking genuine comprehension and knowledge of the outside world. As has been noted more broadly concerning AI-produced content, this can encourage educators and learners to have a degree of confidence in the results. Indeed, Gen AI does not derive knowledge from real-world observations or other essential elements of the scientific method, nor is it in harmony with human or societal values. For these reasons, it cannot generate authentically original material regarding the actual world, items and their connections, individuals and social interactions, human-object connections, or human-technology interactions. The acceptance of the seemingly innovative content produced by Gen AI models as scientific knowledge is debated.

Furthermore, it is important to highlight that concerning the discussion on scientific and objective knowledge, the outcomes of LLMs might not align with the key attributes and tenets of the open science policies. This influences not just epistemic trust but also, due to the inaccuracy, lack of transparency, and complexity of its model, social acceptance. From this viewpoint, it is essential to emphasize that, evaluating the probability or correctness of such results will rely on how effectively we believe the model interpreted the training data. Therefore, the epistemic reliability of the analytical methods will be pertinent to the trust we can assign to AI. Generated text, search.

Alvarado, (2023) asserts that knowledge constraint is also a significant factor contributing to trust problems concerning Gen AI. When users do not comprehend how a Gen AI system reached a certain result, they are less inclined to embrace or utilize it. This has significant implications for how responsibility is assigned. In an increasingly intertwined scenario of human and artificial collaboration in content creation. New methods for the distribution of responsibility for the creation of semantic artifacts will likely be necessary. Reflect on, in this context, the escalating discussion within the realm of scientific inquiry, which concerns not just reproducibility and reliability in their outcomes and findings, but also regarding the actual authorship of the material generated by LLMs. Like a contaminated source upstream, defective AI-generated material can not only disseminate tainted content downstream, which is closely intertwined with reliable and authentic content, but ultimately it also creates dependency in end users, reducing their resistance.

There is nothing more challenging in the quest for understanding than formulating the relevant and appropriate inquiries, from which any genuine exploration can commence. Nonetheless, for an exploration to be productive and fulfill our desire for inquiry, it is essential that the response be found within the limits established by the question, but it is also crucial that the answer be relevant necessary that, within those limits, the response offers something different, which is not already part of the question's prompt; otherwise, the inquiry stays tautological. Furthermore, any genuine response, which provides us with insight into something new, generates additional questions, meaning the potential to examine reality from various perspectives, promoting diversity in research and understanding: Every segment of semantic information is an answer to a question, which, collectively, prompts further inquiries questions regarding itself, which necessitate the proper exchange of information to obtain satisfactory responses, via a suitable web of connections with certain sources of information.

Floridi, (2013) noted that LLMs generate insights by forecasting likelihoods of linguistic patterns present in data derived from an unchanging training dataset that illustrates reality as it exists, presuming our understanding of the visible world. This is probably to reinforce the superiority of the familiar over receptiveness to new ideas and pluralism in a relatively covert manner. Consequently, the understanding and depiction of reality generated by LLMs and Gen AI broadly do not aid in envisioning new and alternative worlds but instead in reinforcing worlds already rooted in conventional perspectives and norms. Gen AI replicates prevailing perspectives in its results and diminishes minority and diverse viewpoints. If human societies are to thrive, it is crucial to understand that Gen AI can never serve as a definitive source of understanding on any subject it interacts with.

Locke (1959), introduced a notion that would fundamentally influence our comprehension of human thought for ages ahead. "Tabula rasa," which translates from Latin to "blank slate," suggested that individuals arrive in the world devoid of inherent knowledge, their minds akin to unfilled pages ready to be written upon by experience. This groundbreaking concept contested existing beliefs that some knowledge is innate, proposing instead that all comprehension originates from sensory perception and environmental interaction. Locke's tabula rasa aligns with AI systems that possess no inherent knowledge and depend solely on input data to acquire knowledge and make choices. AI systems begin without any existing understanding of the issue. Much like Locke's tabula rasa, their knowledge is developed completely through training data and algorithms, devoid of any preconceptions or assumptions. The effectiveness of an AI system is wholly influenced by the quality of the data used during its training, as its performance relies entirely on this.

Features of Large Language Models (LLMs)

Natural Language Processing

Given a collection of texts as input, a language model analyzes statistical attributes of language from these texts, including the occurrence rates and likelihoods of words and their contextual surroundings, which can subsequently be applied for various objectives such as natural language understanding (NLU), generation (NLG), reasoning (NLR), and more generally, processing (NLP) (Deong, et. al., 2019). LLMs has been incremental since the late 2010s and has occurred in surges. After a wave that brought in embedding models like a word2vec for a concise depiction of words as embeddings, the initial significant surge occurred with the advent of LLMs developed upon the Transformer architecture. (Mikolov, et. al. 2013). An LLMs is generally a model that has been pre-trained on extensive large-scale datasets, requiring considerable computational resources and time; however, these models can later be adjusted to focus on specific domains with reduced effort (Bakker, et.al., 2022). LLMs can also capture sensitive details extracted from its training data. Although this information is stored in embeddings that are not interpretable by humans, it has been discovered that a malicious user can deconstruct those embeddings to retrieve the sensitive details, potentially leading to harmful ramifications for the affected individuals.

Synthesizing Knowledge across Texts

LLMs are multifunctional cognitive tools, the philosophy of technology has aimed at enhancing comprehension of cognitive support characteristics and roles of artifacts. (Cassinadri, 2024). Such artifacts are physical items or constructions that effectively assist in carrying out a cognitive activity such as, for instance, recalling, navigating, reasoning, or searching for information-seeking (Heersmink, 2013). LLMs are not just simple language controllers, LLMs, have become essential to different facets of everyday life and are becoming more significant in many professional workflows. LLMs perform an important function in accurate special reasoning-driven areas like robotic object repositioning and text-to-image creation. However, existing techniques encounter difficulties due to a scarcity of real-world instances, managing various layout descriptions, and differing degrees of detail. To tackle these challenges, a new framework called Spatial Knowledge Enhanced Layout (SKE-Layout) has been proposed. SKE-Layout combines various spatial knowledge resources, utilizing both authentic and fabricated data to improve spatial contexts. Charniak, and Dermott (1985). It employs a range of representations tailored to specific tasks and utilizes contrastive learning and multitask learning methods for precise spatial information retrieval. For instance, a task involving object rearrangement could entail organizing particular items on a table, whereas a task related to image generation typically requires constructing a visually coherent scene based on descriptions of where objects are located. Both situations highlight the importance of precise spatial comprehension to guarantee correct positioning and avoid mistakes like misalignment or omissions.

Perform tasks such as Summarization and translation

Large language models have shown extraordinary potential in improving summarization methods. In the age of extensive data, the vast amount of written information has highlighted the significance of effective text summarization methods. Summarizing text, the process of condensing lengthy documents or articles into brief, coherent summaries while maintaining the central meaning and key information, is highly valuable in numerous fields. Large language models have achieved notable progress in code translation activities. Nonetheless, guaranteeing both accuracy and comprehensibility of translated code continues to be a hurdle, restricting their successful use in practical software development. Code translation entails transforming code from one programming language to another, a process often necessary for application porting or software migration (Nguyen, et.al. 2013). LLMs like GPT-4 have transformed numerous programming activities, encompassing code creation and software debugging including with their outstanding performance. (Fan et al. 2023), LLMs have greatly progressed the area of code summarization, offering advanced techniques for producing brief and precise summaries of source code.

Models developed utilizing deep learning algorithms

LLMs are advanced algorithms based on deep learning that can identify, retrieve, condense, forecast, and produce text grounded in insights acquired from extensive

training on substantial datasets. They are also a subset of a broader technology known as language models. Al language models share a common feature: they are capable of analyzing and producing text that resembles natural language. This is referred to as executing tasks associated with natural language processing (NLP). All language models have the capability to carry out NLP tasks, but they vary in other attributes, such as their dimensions. In contrast to other models, LLMs are classified as large due to two factors: (1) they are trained on extensive datasets (2) They consist of a vast quantity of learnable parameters (i.e. representations of the underlying structure of training data that can help models perform tasks on new or never-before seen data). The magnitude of a LLM is characterized by the quantity of parameters. (Kamath et.al. 2024). LLMs exhibit strong abilities in representing general language and display exceptional performance in tackling intricate tasks. Advanced learning algorithms are capable of identifying intricate images, written content, audio signals, and various data patterns to generate precise insights and forecasts.

Opaque inner Working LLMs

The internal working large language models (LLMs) are not well comprehended, making it challenging to clarify their problem-solving approaches. LLMs like GPT-4 are powerful tools, but opaque "black boxes", they serve as the driving force for everything from customer service chatbots to sophisticated research instruments. The majority of prominent LLMs are created by large technology firms such as Google, Meta, and Open AI. These models are proprietary systems whose comprehensive specifics are not disclosed to the public. The companies regard their LLMs as trade secrets, revealing merely restricted details regarding model design, training dataset, and decision-making processes. Huang and Siddarth (2003). This lack of clarity hinders the ability to conduct independent evaluations of the systems for biases, errors, or knowledge-related problems. In the absence of transparency, it becomes unfeasible to ascertain whether unsafe data is utilized in training or if the models display unjust biases. Rich, (1983). The general public must naively rely on the creators' reassurances. This deficiency in transparency allows for the implementation of possibly troublesome systems with minimal responsibility. It also prevents recognizing and tackling problems with LLMs. For instance, LLMs can produce imprecise or incorrect forecasts. The absence of clarity in their decision-making creates challenges in pinpointing the underlying issues. Although LLMs may have access to vast amounts of training data, the way they analyze and use that data is influenced by their model structures and algorithms. (Roose, 2023). Their understanding is fixed, trapped at the moment of training. However, their reasoning may reveal fluid and erratic biases influenced by those architectural factors. Nevertheless, the distinguishing feature of GPT-3 that positioned it as the first contemporary LLM is its capability for few-shot learning, which allows it to acquire a new skill from a limited number of samples in one exchange, and chain-of-thought reasoning, which is the ability to articulate its reasoning for difficult tasks upon request. (Kojima, et.al. 2022). Calls for technology firms to allow examination of their AI systems are increasing. As LLMs become more capable of using human language and human concepts, they also become more adept at acquiring the generalizations we desire.

EPISTEMIC HARMS RESULTING FROM THE USE OF LLMs

AI has achieved extraordinary milestones, from excelling in intricate games to powering voice-activated helpers. Nevertheless, even with these progressions, AI systems frequently lack common sense – the capability to comprehend and reason about reality. (Chollet, 2019). Every human and animal show the common sense essential for survival, illustrating fundamental intelligence through their engagement with their surroundings. The spread of misinformation has already been demonstrated to be an issue in social media and online communities, where inaccurate information quickly circulates among users.

Human-LLMs engagement brings about both epistemic harms and possibilities. Epistemic harms in form of knowledge related risks can be classified into five groups: 1. Epistemic damages on the deterioration of embodied cognition, 2. Epistemic detriments to critical thinking and creativity, 3. Erroneous epistemic trust and contentment, 4. Epistemic Impact on Truth and Accuracy, 5. Difficulties in Epistemic Misalignment.

Epistemic Damages on the Deterioration of Embodied Cognition

Embodied cognition is the concept that our physical body, sensations, and bodily experiences are crucial for our comprehension of the world, the development of conceptual understanding, and the creation of meaning. Frank, et.al. (2024) argue that firsthand human experience is a vital prerequisite for the potentiality of science. LLMs exhibit a form of linguistic proficiency that theories of embodied and enactive cognition have long considered unattainable for systems devoid of the significant viewpoint of a sentient being, which is sense-making. There are essential distinctions between human thought processes and AI, especially LLMs.

Epistemic Detriments to Critical Thinking and Creativity

Critical thinking and creativity represent another facet of epistemic harm linked to LLMs. The paper contends that the epistemic effects of Gen AI, encompassing LLMs, on students' cognitive abilities and creativity exhibit drawbacks, such as cognitive fixation and diminished creative confidence. (Habib, et.al. 2024). It is essential to implement LLMs in educational settings in a way that enhances their advantages while reducing their drawbacks, allowing AI to serve as a facilitator instead of a limitation on creative capabilities. The research highlights the importance of LLM-driven learning platforms that promote source validation and independent reasoning, guaranteeing that students engage in critical learning rather than passive absorption.

Erroneous Epistemic Trust and Contentment

Cognitive shortcomings in LLMs consist of the black box issue and illusions. Biases influence the decisions of LLMs, and misinformation presents a significant challenge to trust. Confidence in AI, particularly its Generative LLMs models, is tenuous. There are substantial worries regarding AI bias, privacy, and data security. Public doubt about AI has markedly increased, mainly because of the bias in LLMs and their hallucinated

outputs, which contribute to misinformation. Political knowledge and principles are at risk due to the use of LLMs and AI technologies by politically authoritative figures, owing to their lack of transparency. Trust holds significant value; by placing our trust, we subject ourselves to potential risks of harm and betrayal. In order to reframe trust and satisfaction in LLMs, it is essential to thoughtfully examine the intricate relationship between the human aspect, namely the trustor, and the agent aspect, referred to as the trustee, in the technology of LLMs that is being relied upon. Irzik and Kurtulmus (2019) assert that epistemic trust can be characterized as accepting another individual's testimony or assertions as justification for believing that statement and having faith that the individual conveys that statement honestly, and has good reasons or justification for it.

Epistemic Impact on Truth and Accuracy

There are knowledge-related worries regarding truth and precision, as LLMs do not consistently provide us with accurate information. At times, LLMs generate false information, which serve as means for spreading misinformation. Marcus and Davis (2023) state that LLMs frequently generate incorrect or deceptive content; they are designed to create responses that appear credible but can actually be erroneous. LLMs are also utilized on purpose to disseminate false information or to influence individuals. For example, AI is being utilized for political manipulation by partnering with a data analysis company, which entails using data from social media to deliver customized messages to particular voter groups in an effort to sway their political views. In other words, LLMs could pose a threat to democracy.

When LLMs are employed for data examination and information compilation, the realm of scientific publishing faces the potential for inaccuracies or falsified data, which could breach the reliability of scientific publishing numerous esteemed journals, like *The* National Academy of Sciences have implemented guidelines limiting the use of LLMs, permitting their use solely for enhanced clarity and writing style. (King, 2023). There is an agreement in the scientific community regarding that, LLMs cannot qualify for co-authorship. However, LLMs are designed to mimic human behavior, and their training allows them to excel in numerous tasks, when individuals are less informed about specific areas of expertise, whether in their career or general understanding, the process of enhancing knowledge similarly decreases.

Difficulties in Epistemic Misalignment

LLMs are progressively utilized as a means for acquiring knowledge, yet users struggle to clearly indicate their preferences for information presentation. Misalignment between internal certainty and external expression can lead to users being misled by overconfident or under confident statements, posing significant risks in high stakes domains such as legal advice, medical diagnosis, and misinformation direction. (Feng, et.al. 2023). For example, when users ask LLMs to reference credible sources, they find that the existing interface lacks the capability to do so structured way to articulate these preferences. The main method by which we carry out research is by asking questions, and the eventual success or failure of an intellectual exploration largely depends on the choice and level of quality of these questions. De Ridder (2022) states that conducting research in the digital

era poses extra challenges, as the vast amount of information delivered through unclear discovery methods, like web searches and recommendation systems, can create false senses of comprehension, leading users to believe they have conducted a comprehensive inquiry, while their approaches are flawed or superficial. When individuals possess limited understanding of knowledge areas, whether related to their field of work or general awareness, the effort to enhance that knowledge equally decreases.

Epistemic standards pertain to the methods by which individuals assess the origins and the condition of their convictions as well as their interrelations regarding the evidence at hand. In the optimal epistemic situation, reasoning individuals strive for conceptual comprehension and validation of their beliefs in relation to the available evidence. When learning agents delegate their understanding and utilization of a knowledge area to an LLM, they relinquish the journey of developing individual expertise. Consequently, this leads to a heightened dependence on LLMs for understanding and executing tasks. This dependence further on ally can erode dispositions and habits of learning practices given the long-term reliability of LLMs.

Methodology

In order to examine the epistemic harms connected to generative artificial intelligence, this study uses a qualitative technique through a review of the literature. It highlights societal norms, potential prejudices, and legal requirements—all of which are essential for epistemic and ethical AI development. The article asserts that despite growing public concerns and numerous instances of misuse, the discourse around the rapid advancement of generative AI lacks a conceptual framework that makes a clear connection between epistemic concerns and multiple instances of misappropriations.

Epistemic Principles for Large Language Models

Although Large Language Models act as instruments for gathering knowledge, users are unable to clearly indicate their preferred format for information delivery. However, providers have somewhat tackled certain issues; however, they failed to establish adequate mechanisms for defining epistemic preferences. Drawing from social epistemology, we proposed the subsequent epistemic principles:

Epistemic Alignment

Large language models have arisen as formidable knowledge tools however their adaptability prompts the inquiry of how to guarantee they provide information in a manner that aligns with personal preferences regarding knowledge quality, evidence standard and perspective diversity. (Gal, et. al. 2024). For AI developers, epistemic alignment provides clear direction for facilitating various methods of understanding, while for users, it strives to present information that caters to their individual requirements instead of relying on uniform solutions.

User Preference Based knowledge

The preferences of LLMs should correspond with human expectations and needs, for improved service. The anticipation is for the LLMs agent to behave in a manner that renders its actions or strategy understandable to the human, while the extra interaction in the form of clarifying conversation can assist in closing the divide between the individual's expectations and the agent's actions (Zhang, et.al. 2017). The individual in this situation might be a specialist in the field and has understanding about which actions can be taken, as well as the outcomes of those actions. But does not know how the LLMs agent computes its plan or policy. Feng, et.al. (2023) describes preference alignment as an emerging subject within the LLMs community that seeks to integrate human preferences for fine-tuning the LLMs during the training phase. The goal of preference alignment is to guide the model in producing content that aligns with human preferences.

Prompt Engineering of Hermeneutical Epistemic Justice

Prompt engineering of Hermeneutical epistemic justice can transform how we engage with LLMs, ranging from complex summarization to generation dynamic, customized replies. However, it is crucial to guarantee that the information individuals receive is accurate and impartial, particularly since these systems can perpetuate societal biases due to their training data potentially being unrepresentative of the population or excessively generalized. When data is gathered meticulously, it may still possess historical and structural prejudices, resulting in frameworks that overlook the requirements of marginalized social communities. Shubham and Harsh (2024) noted that there remains deficiency in grasping how prompt engineering can be utilized to diminish hermeneutical epistemic injustices in LLMs' responses.

However, prompt crafting of interpretive epistemic fairness in gendered connections and male-to-female the pronoun distribution in gendered LLM text responses can be effectively managed through careful design. Prompt engineering includes formula ting natural language directives to assist LLMs in generating relevant responses.

Epistemic Transparency

It is paramount to pursue new approach for a responsible LLMs openness. Openness is essentially concerned with fostering suitable human comprehension, and this comprehension is pursued by various stakeholders with distinct objectives in diverse settings LLMs models have been observed to biases, posing a danger of spreading detrimental discrimination, stereotypes, and exclusion on a large scale. It is well-documented that they can fabricate information generating results that seem believable – even persuasive but inaccurate. (Bender, et. al. 2021). LLMs can generate sometimes harmful content, such as hate speech and offensive language, or disclose sensitive information that jeopardizes privacy or security. Considering the expected influence of LLMs on both our everyday lives and society as a whole, it is essential that LLMs and applications incorporating LLMs are created and implemented in a responsible manner.

A key element of responsible AI development and deployment is transparency. Epistemic clarity is essential to allow pertinent stakeholders to develop an accurate comprehension of a model or system's functionalities, restrictions, operation, and how to utilize or manage its outcomes. Creators of LLMs are unable to troubleshoot their models or properly evaluate them without at least some grasp of their actions and effectiveness across various tasks. End users need to develop a sufficiently precise comprehension of applications integrated with LLMs to influence the application's conduct and attain suitable degrees of confidence and dependence. (Samuel, 2023). Epistemic transparency is a mechanism to build trust, but achieving a certain degree of trust is essential for end users to fully utilize the advantages of LLMs systems without over depending on inaccurate LLMs results.

Epistemic Calibration

Calibration confidence is crucial in numerous high-stakes scenarios where inaccurate predictions can pose significant issues. (Zhang, et. al. 2022). Intuitively the predictive confidence of a well calibrated model might reflect the true probability of correctness. For instance, if a calibrated model predicts a collection of input x is assigned to classy with a 40% likelihood, so we anticipate that 40% of the inputs actually belong to that class. Formally, a model is considered to be perfectly calibrated. Although enhancing the calibration accuracy of pre-trained language models and attaining similar task performance, our explanation-driven calibration approach is still constrained by the dependability and precision of interpretable techniques. Jiang, et. al. (2023) state that large LLMs typically attain both improved accuracy and enhanced calibration performance. Calibration refers to the alignment between a probability evaluation, represented as the subjective probability, of the happening of a specific event.

Conclusion

Generative AI including LLMs, are swiftly turning into essential components of our existence via smartphones and personal computers, LLMs are increasingly integrated into the fabric of societal operations. LLMs are designed to identify important words and connections found within the ever-changing context of human language. There is a significant risk of misinformation in human workflow due to misguided confidence and a sense of satisfaction with quick, fluid, but possibly incorrect responses. In contrast to human connection, trust is developed gradually and depends more on the output's outward appearance than on its underlying accuracy. Nevertheless, the knowledge-related harms occasionally generated by LLMs mirror the characteristics of the data sets on which they were trained, rather than the internal motivations we recognize in conscious beings. Epistemology holds a distinctive role in the study of knowledge, representing a methodical and theoretical comprehension comparable to scientific understanding. Epistemology continued to be a philosophical exploration of the ways in which knowledge is characterized, obtained, and validated.

References

Alvarado, R, (2023). "AI as an Epistemic Technology", https://link.springer.com(Accessed 24/4/2024).

Bakker, M. Chadwick, M.Sheahan, H. Tessler, M. Campbell-Gillingham, L. Balaguer, J. et. al. (2022). Fine-tunning Language Models to Find Agreement among Humans with Divers Preferences, *Advance Neural Information Processing Sysyem,* 35, 38176-38189.

Bender, E. Gebru, T. Major, A. Shmitchell, S. (2021). "On The Dangers of Stochastic Parrots: Can Language Models be Too Big"? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT),* p. 610-623.

Buchanan, B. G., and E. H. Shortliffe (1985). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, M. A: Addison-Wesley.

Cassinadri, (2024). "GhatGPT and the Technology-Education Tension: Applying Contextual Virtue Epistemology to a Cognitive Artifact", *Philosophy and Technology*, 37, 14.

Charniak, E., and D. V. McDermott (1985). *Introduction to Artificial Intelligence*. Reading, MA: Addison-Wesley.

Chollet, F. (2019). *On the Measure of Intelligence.* https://arxiv.org/abs/1911.01547 (Accesed: 10/1/2025).

Deong, L. Yang, N. Wang, W. Wei, F. Liu, X. Wang, Y. et.al. (2019) "Unified Language Model Pre-Training for Natural Language Understanding and generation", in *Advances in Neutral Information Processing Systems,* New York: Curran Associates Inc.

De Ridder, J. (2022). *Online Illusion of Understanding Social Epistemology,* ISSN14645297oi:10.1080/02691728.2022.2151331.

Fan, Z. Gao, X. Mirchev, M. Roychoudhury, A. and Tan, S.H. (2023). "Automated Repair of Programs from Large Language Models", *Proceedings of the 2023 IEEE/ACM 45TH International Conference on Software Engineering (ICSE),* p. 1469-1481.

Feng, S. Sorensen, T. Liu, Y. Fisher, J. Park. C. Choi, Y. Tsvetkov, y. (2023). "Modula Pluralism: Pluralistic Assignment via Multi LLMs Collaboration", https://arxiv.org/abs/2406.15951.

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and* opportunies, Oxford: Oxford University Press

Frank. A, Gleiser, M.Thompson, E. (2024). *The Blind Spot: Why Science Cannot Ignore Human Experience.* USA: MIT Press.

Gal, Y. Roee, A. and Mor, G. (2024): Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words? 5, 2024URL http://arxiv.org/abs/2405.16908.

Gui, G. and Toubi, O. (2023). The Challenge of Using LLMs to Simulate Human Behaviour: A Causal Inference Perspective, https://arxiv.org (Acessed7/5/2025

Hofstadter, D. R. (1985). *Metamagical Themas: Questing for the Essence of Mind and Pattern.* New York: Basic Books.

Huang S, Siddarth, D. (2023) *Generative AI and the digital commons*. ArXiv. https://arxiv.org/abs/2303.11074 (Accesed: 10/6/2025).

Irzil, G, and Kurtulmus, T. (2019). What is Epistemic Public Trust in Science? *The British Journal for the Philosophy of Science,* 70 (4): 1145-66.

Kamath, U. Keenan, K. Somers, G. and Sorenson, S. (2024). Large Language Models: A Deep Dive: Bridging Theory and Practice (1st ed.). Springer Nature Switzerland, https://doi.org/10.1007/978-3-031-65647-7

King, M.R. (2023). "A Place for Large Language Models in Scientific Publishing, apart from credited authorship". *Cell MoI.* Bioeng. 16-19, 95-98.

Kojima, T. Gu, S. S. Reid, M. Matsuo, Y. and Iwasawa, Y. (2022). "Large Language Models are Zero-Shot Reasoners", *ICML 2022 Workshop o Knowledge Retrieval and Language Models.* (Accessed 7/1/2025).

Locke, J. (1959). *An Essay Concerning Human Understanding, Vol.1*. New York: Dover Publications Inc.

Locke, J. (1959). *An Essay Concerning Human Understanding, Vol. 11*. New York: Dover Publications Inc.

Marcus, G. and Davis E. (2023). Large Language Models like ChatGPT say The darnedest
things. Blog@ACM *Communication of ACM,*
https://cacm.acm.org/blogcacm/large-
language-models-like-chatgpt-say-the-darnedest-things/

Mikolov, T. Cohen, K. Corrado, G. and Dean , J. (2013). "Efficient Estimation of Word Representations in Vector Space", arXiv.doi: 10.48550/arXiv.1301.3781.

Rich, E (1983). *Artififecial Intelligence*, New York: McGraw-Hill

Roose, K. (2023). "A Conversation with Bing's Chatbot Left Me Deeply Unsettled". *New*
York Times, URL https://www.nytimes.com/2023/02/16/technology/bing-
chatbot-microsoft-chatgpt.html. (Accessed 17/2/2025).

Samuel, R. (2023). "Eight Things to Know about Large Language Models", *CoRR* arXiv:23004.00612.

Shubham, V. and Harsh, D. (2024). "A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks". arXiv

Smith, C (2006): *The History of Artificial Intelligence,* USA: University of Washington.

Waterman, D. A. (1986). *A Guide to Expert Systems*, *Reading,* MA: Addison-Wesley.

Zhang, J. Sreedharan, S. Kulkami, A. Chakraborti, T. Zhu, H. and Kambhampati, S. (2017).
"Plan Explicability and Predictability for Robot Task Planning", *2017IEEE International Conference on Robotic and Automation (ICRA)* p.1313-1320.

Zhang, L, Deng, Z. Kawaguchi, K. ZOU, J (2022). When and How Mix up Improves Calibration, ICML. 4, p. 26135-26160