Bane and boon of hallucinations in context of generative AI

Hrishitva Patel[1]

[1]Information Technology, University of Texas San Antonio, Carlos Alvarez college of Business

April 01, 2024

Bane and boon of hallucinations in context of generative AI

Hrishitva patel [1]

University of Texas San Antonio, Information Technology, Carlos Alvarez college of Business, San Antonio, Texas

Email: hrishitva.patel@my.utsa.edu

Abstract

The phenomenon of hallucinations takes place when generative artificial intelligence systems, such as large language models (LLMs) like ChatGPT, generate outputs that are illogical, factually incorrect, or otherwise unreal. In generative artificial intelligence, hallucinations have the ability to unlock creative potential, but they also create challenges for producing accurate and trustworthy AI outputs. Both concerns will be covered in this abstract.

Artificial intelligence hallucinations can be caused by a variety of factors. There is a possibility that the model will show an inaccurate response to novel situations or edge cases if the training data is insufficient, incomplete, or biassed. It is common for generative artificial intelligence to generate content in response to cues, regardless of the model's "understanding" or the quality of its output.

The rising body of material will be synthesised in the paper in order to provide a comprehensive approach to AI hallucinations. The purpose of this study is to gain a deeper understanding of artificial intelligence hallucinations and to develop strategies that can reduce the negative impacts of these hallucinations while simultaneously maximising their creative potential. For the purpose of contextualising artificial intelligence hallucinations in generative AI and providing information for the building of more reliable AI systems, the technique will examine models and mind maps from the journals that were referred to. Hallucinations experienced by AI models can be mitigated through the critical analysis of AI outputs and the diversification of data sources.

This paper stresses the significance of high-quality training data, human feedback, transparency, and ongoing quality control in the development of artificial intelligence. This is accomplished through a synthesis of the relevant literature as well as an analysis of models and mind maps.

Contents

1. Introduction

AI has progressed leaps and bounds. It can do a plethora of tasks which includes texts, images, sounds and code. This is generally termed as generative AI. The content these software's produce is indistinguishable from the content that humans can produce. This is possible due to the fact there is a huge amount of training data behind which has been fed to the system, it uses these datasets to create detect patterns and synthesize data in a creative manner. Generative AI uses machine learning models like deep neural networks. The data generated by an AI model is not necessarily accurate, this is where it gets interested because it might occasionally cause problems creating hallucinations. There are different cases in hallucinations, sometimes the information is false and hypothetical and other times the information can be plausible but not genuine. AI has integrated itself into our daily lives and people have begun relying on it for many tasks making it an extremely important topic. Hallucinations can cause the spread of misinformation and harm to trust.

Artificial intelligence hallucinations can be extremely harmful in fact-based areas such as the media, academia, and the legal system. Artificial intelligence has been employed in court filings to generate information containing fake judicial opinions and legal citations, with real-world consequences for the parties concerned. AI hallucinations is a problem which crossed beyond the point of dissemination of incorrect information to ethical problems. There are certain risks attached to the usage of AI and they have negative impacts which can result in spreading of stereotypes and maligning reputations. The research will focus on understanding the causes of AI hallucinations and how can they be mitigated, because it has become deeply ingrained within our lives and its use only widens by the passing moment. The overall goal is to determine the flaws to improve the design and the outputs which AI provides us with so the systems in place can become more dependable and ethical.

1.1 Purpose of the Study:

The purpose of this study is to examine AI models and determine their causes of hallucinations, the study would go on to determine the underlying causes by examining the people who are working on AI and automation. Data would be collected from 200 individuals which would include students, start-ups, and managers to evaluate creative techniques to mitigate the use of AI.

1.2 Background of Study:

When AI began back in the 20th century it was simpler, but it has developed to the complex form that can be seen today, since work on AI began hallucinations also began from a simpler spectrum to a more complex one(Natale & Ballatore, 2020). These nascent AI models were produced from basic logic and minimal datasets which caused them to become susceptible to Hallucinations. Equipped with an insufficient training data these models were prone to mistakes. Due to the simplicity of AI models back in the day, singling out hallucinations was easier thus opening ways to their resolution and detection in an easier manner(Challen et al., 2019). The dynamics of AI changed in the late 20th century when machine learning came to be. **Neuronal Networks** were considered to be a giant, as from the name it can be deduced that these models were based on the model of the human brain which went on to provide flexibility and intelligence in a better way as compared to the era before this innovation. Since, AI came to be on this model, it began resulting in even more complex hallucinations.

To further this after the advent of **Deep Learning**, utilizing deep neural networks with layers after layers which could learn from enormous quantities of information(Sze et al., 2017). Producing data of unprecedented complexity, the **convolutional neural network (CNN**) **and Recurrent Neural Network (RNN)** gained a lot of popularity(Hoffmann et al., 2017). CNN forte was in analysing visible interpretation and analysis while RNN specialized in analysing text-based data making it better at sequential data processing. Diving further into the problem of adequate training data for AI models, there seemed to be not a lack of data butt the lack of proper information due to which it began to create hallucinations. To further enhance the interplay between huge datasets and model architectures a **transformer-based models** were form, this was the most detailed model which processed massive amounts of information(Gillioz et al., 2020), now these hallucinations became even harder to detect and treat. There has been progressive approach towards creating an environment under which hallucinations don't occur, but the complex the model the similar the hallucination it would produce.

1.3 Research Problem:

The research focuses on understanding hallucinations and why are they produced in generative AI systems, the investigation would look into the causes and impacts of hallucinations to design a strategy for mitigating harmful effects all the while harnessing its creative potential.

1.4 Research Questions:

1. What is impact of Data Quality Concerns on Algorithm Opacity in generative AI systems?
2. What is the impact of Data Quality Concerns on AI Hallucination Effects?
3. What is the impact of Algorithm Opacity on AI Hallucination Effects in generative AI systems?
4. Does Algorithm Opacity mediate the association between Data Quality Concerns and AI Hallucination Effects?

1.5 Research Objectives:

1. What is impact of Data Quality Concerns on Algorithm Opacity in generative AI systems?
2. What is the impact of Data Quality Concerns on AI Hallucination Effects?
3. What is the impact of Algorithm Opacity on AI Hallucination Effects in generative AI systems?
4. Does Algorithm Opacity mediate the association between Data Quality Concerns and AI Hallucination Effects?

2. Literature Review

2.1 Cross Disciplinary Perspectives:

To understand AI Hallucination, cognitive science can be sued to understand the human perception and cognition. Humans have a certain fallibility when it comes to interpreting sensory data and recalling events, in a similar fashion AI system can produce outputs which diverge from reality which can be attributed to their programming and training data(McIntosh et al., 2023). AI which is trained on a specific dataset can display confirmation bias, it could lean towards favoring familiar patterns and disregarding outliers or provide new information all together. Having an understanding of human cognitive science can help in balancing and creating precise AI systems. To further understand AI hallucinations psychology can be looked at as it studies human perceptions and errors which aid in comprehending AI hallucinations(McCarthy-Jones, 2011). In AI models data processing perceptual errors can be somewhat parallel to that in humans. An experience known as Pareidolia under which humans often experience patterns which do not exist(Boschetti et al., 2023).

Similarly, AI can generate hallucinatory outputs in images and pattern recognition making connections which might not exist. By analysing psychological parallels, it could become easier for researchers to gain valuable insight into why AI models cause errors and what are the effective strategies for mitigating them. Another important field of study is linguistics as it studies the language and how it is structured which offers insights into AI hallucinations, this holds immense value under Natural Language Processing (NLP). Humans tend to use social and environmental cues in their conversation along with language which offers a clear and concise understanding of what is being communicated while AI cannot make use of these cues rather it reads under what context the language is being used and it conveys an answer in the similar tone(Church & Liberman, 2021). So, AI lacks depth at the moment to interpretation which can lead to errors due to differences in perception.

2.2 Technical Underpinnings of Hallucinations:

It is essential to have a solid understanding of the technological mechanisms that cause AI hallucinations to effectively handle and comprehend this issue in generative AI. These factors, which range from how data is processed to the architectural decisions made by AI models, have a substantial impact on the occurrence of hallucinations as well as the type of these experiences.

The attention mechanisms that are present in neural networks, play a crucial role in deciding how an artificial intelligence model processes and prioritizes various components of the received data(Niu et al., 2021). The attention mechanisms which were initially created to increase the performance of models in tasks such as language translation. These mechanisms enable the model to concentrate on relevant aspects of the input while, discarding less relevant information. This means, that these systems on the other hand, have the capability to be a source of hallucination. If the attention of the model is not properly aligned, it may give an excessive amount of value to aspects of the data that are either irrelevant or misleading(Yu et al., 2019), which will result in outputs that are out of sync with the real content. This also brings us back to the issue of depth of understanding AI and such a mismatch might be especially prominent in complicated activities, where the essential context may not be immediately apparent or may be dispersed throughout a variety of different areas of the input. Breaking down text into smaller units for processing is called tokenization. This is one of the fundamental steps of NLP. The tokenization process is as important as any when it comes to AI-generating hallucinations, tokenization can greatly affect performance by splitting words into smaller and more manageable pieces if the process is not well-tuned then the algorithm might misinterpret the

meaning of certain terms like homonyms or phrases to recognize the significance of idioms or phrases.

The architecture of the intelligence model is one of the crucial factors when it comes to hallucinations(Singh et al., 2020). Under these models it is studies how various components are organized within a system and how do they interact with one another. CNNs, RNNs, and Transformers are all examples of such systems, each having its own design, advantages, and disadvantages. CNNs are very useful when images are being analysed, they process spatial information. They are structured in a way which makes them lead to overfitting particular patterns in the training data they have been given, which results in the creation of hallucinations. This mostly happens when the information fed is ambiguous and can lead to double meanings. When the context has a widespan of coverage of text, RNNs can face difficulty in establishing long-range dependencies which further translates into hallucinations. Similarly, in case of transformers they have the same issue, even though their capacity of managing sequential information is better as compared to RNNs but they face the same problem due to the complication of decision-making processes.

2.2.1 The Role of Training Data and Algorithms:

The importance of training data cannot be understated when it comes to generative AI, the quality and composition of the said data can be a critical factor when it comes to biases leading to skewed outputs. When there is missing data than the generative AI model tries to fill in the missing information and this might not be done appropriately(Fuchs, 2018). This makes the model prone to generation of hallucinatory content because the model is not adequately regularized to prioritize certain aspects of data which might be incomplete within the algorithm.

2.3 Ethical and Societal Considerations:

When hallucinations are considered in the context of Generative AI, there are not only technical underpinning but there are ethical and societal considerations which need to be looked at which touch the fundamental aspects of trust, responsibility, and impact in various domains of society. AI has deepened its roots within our daily lives, similarly the impact of the content it generates has become increasingly critical. This creates a necessity of a deep ethical reflection and a proactive approach. Below are some examples of studies conducted on the ethical challenges which the world is faced with.

Ethical challenges	Challenges	Issues	References
Harmful or inappropriate content	Content produced by generative AI	Could be violent, offensive, or erotic	(Zhou et al., 2021)
Bias	Training data representation	Representing only a fraction of the population may create exclusionary norms	(Zhou et al., 2023)
	Training data language	In one single language (or few languages) may create monolingual (or non-multilingual) bias	(Weidinger et al., 2021)
	Cultural sensitivity	Cultural sensitivities are necessary to avoid bias	(Malik et al., 2023)
	Employment decision-making	Bias exists in employment decision-making by generative AI	(Chan, 2022)
Over-reliance	User verification	Users adopt answers by generative AI without careful verification or fact-checking	(Iskender, 2023) (Van Dis et al., 2023)
Misuse	Plagiarism in academia	Plagiarism for assignments and essays using texts generated by AI	(Cotton et al., 2023)
	Cheating in academia	Generative AI can be used for cheating in examinations or assignments	(Susnjak, 2022)
Privacy and security	Information disclosure	Generative AI may disclose sensitive or private information	(Fang et al., 2017) (Siau & Wang, 2020)

Digital divide	First-level digital divide	For people without access to generative AI systems	(Bozkurt & Sharma, 2023)
	Second-level digital divide	In which some people and cultures may accept generative AI more than others	(Malik et al., 2023)

2.3.1 Impact on Journalism and Spread of Misinformation:

AI has offered us with a great degree of efficiency and scalability, this has a great potential to revolutionize content creation. This is a sensitive field under which the reliability and integrity of information has always been in the line of fire, even before the advent of Generative AI systems. AI has the capability to generate articles which can have biased interpretations which in the end contributes to the misinformation(Flores Vivar, 2019). This has a two-pronged affect, one being the trust of public in media will dwindle and second is the implications for democracy and informed public discourse. AI is no doubt helpful, but the ethical dilemma here is that editorial standards need to increase to corroborate the factual accuracy of the data being produced.

2.3.2 Legal and Healthcare Decisions:

In legal and healthcare sector, AI has a wide range of applications which deal with high stake decisions. These decisions require accurate AI systems to generate a thorough analysis such as incase of case outcome predictions. Using AI can be helpful, but it can cause hallucinations and generate inaccurate legal advice or maybe even biased assessments which compromise fairness and justice(Li, 2023). Furthermore, in healthcare AI is used in treatment recommendations, diagnosis and patient management which is all very sensitive work any inaccuracies might lead to misdiagnosis to inappropriate treatments(Umapathi et al., 2023). This will end up causing harm to the patient. AI application in such cases should be highly scrutinized furthermore, it is necessary for the development of AI that great care be applied to stop perpetuation of harm.

2.3.3 Public Perception and Trust:

Managing the public perception in AI technology is important, the hallucinations in AI have led to a public distrust in the technology which can lead to scepticism. This creates a gap in the overall understanding and usability of AI, for which the developers need to engage with public, policy makers and other stakeholders involved to build a deep understanding and trust

in AI technology, there are certain challenges which need to be tackled within the technology against which the users need to be sensitized this could be achieved by clear communication and of the capabilities.

Technology Challenges	Challenges	Issues	References
Hallucination	Contents generated	May be nonsensical or incorrect	(Alkaissi & McFarlane, 2023),(Azamfirei et al., 2023),
	Generative AI output	May provide fictitious information or information embedded with factual errors	(Ji et al., 2023)
Quality of Training Data	Data sufficiency	Difficult to obtain enough training data to ensure quality	(Gozalo-Brizuela & Garrido-Merchan, 2023), (Su & Yang, 2023)
Explainability	Interpretability of output	Difficult to interpret and understand the outputs of generative AI	(Malik et al., 2023)
	Error identification	Difficult to discover mistakes in the outputs of generative AI	(Rudin, 2019)
	Trust in AI	Users are less or not likely to trust generative AI	(Burrell, 2016)
	Regulatory challenges	Difficulty for regulatory bodies in judging fairness or bias	(Rieder & Simon, 2017)
Authenticity	Content manipulation	Manipulation of content (such as images and videos) causes authenticity doubts	(Gragnaniello et al., 2022)
Prompt Engineering	Importance in digital literacy	Prompt engineering becomes vital for the full utilization of generative AI	(Liu & Chilton, 2022)

	Improvement techniques	Techniques like brute-force trial with the prompt needed to improve AI-generated content quality	

2.4 Regulatory and Policy Landscape:

The topic of artificial intelligence (AI) regulation and policy is one that is fast changing as a result of the rising awareness of AI's significant social influence and the need to weigh its advantages against any potential concerns, including the possibility of AI hallucinations. As artificial intelligence (AI) technologies progress and become increasingly interwoven into many industries, governments and global organizations are facing the difficulty of creating structures that guarantee the ethical advancement and use of AI.

2.4.1 Challenges in AI Regulation

The governance of AI has several difficulties, one of which is keeping up with the quick advancement of technology. Rules must be adaptable enough to take into account new developments while maintaining the core moral and security standards. The worldwide character of AI development presents another difficulty, necessitating international norms and collaboration for the efficient management of cross-border concerns.

Challenges	Issues	References
Copyright	Using AI-generated content may violate copyright	(Pavlik, 2023)
	Controversies exist over AI authorship	(Bridy, 2012),(Murray, 2023), (Sallam, 2023)
Governance	Lack of human controllability over AI behavior	(Taeihagh, 2021)
	Data fragmentation and lack of interoperability	(Taeihagh, 2021)
	Information asymmetries between tech giants and regulators	(Kroll, 2015), (Taeihagh et al., 2021)

2.5 Learning from AI Errors:

Analysing AI errors offers insights into the learning process, akin to the importance of mistakes in human learning. Studying AI hallucinations helps researchers understand model design better. This process is similar to human learning from mistakes, gradually improving

skills and knowledge. The ethical and practical implications of AI hallucinations and human creativity should be carefully considered(Chanda & Banerjee, 2022). Unlike human errors, which are bounded by individual responsibility and social norms, AI-generated errors can have widespread and unforeseen consequences, particularly in critical applications like healthcare or autonomous vehicles. Therefore, while embracing the creative potential of AI errors, it is crucial to ensure that these systems are designed and deployed responsibly.

2.6 Role of AI in Self-Regulation and Error Mitigation:

The initial phase of AI self-regulation involves the creation of self-diagnostic capabilities(Molenaar et al., 2023). AI systems should possess the ability to evaluate their outputs for accuracy and differentiate reliable information from hallucinations. Implementing self-diagnosis necessitates advanced algorithms for comparing outputs with standards or expectations.

In NLP, an AI system can assess text using linguistic rules and contextual relevance(Youvan, 2023). If the output deviates significantly from expected norms, the system identifies it as a potential error or hallucination. Iterative learning is a continuous improvement process in which AI systems enhance performance by learning from their own outputs(Liang et al., 2023). AI models can be trained to learn from errors and adjust parameters to minimise future mistakes in error mitigation. This approach emulates human learning, where experience and reflection enhance future performance. Reinforcement learning improves AI's iterative learning through feedback for behaviour adjustment. Real-time error correction is crucial for AI self-regulation. This pertains to AI systems autonomously identifying and rectifying errors in real-time, without human intervention.

Real-time error correction necessitates AI systems with rapid response and immediate corrective capabilities. Timeliness and accuracy are crucial in applications like autonomous vehicles and real-time medical diagnostics, as any delays or inaccuracies can have serious consequences. The incorporation of AI for self-regulation and error mitigation poses various challenges and considerations. One key challenge is to prevent self-regulation mechanisms from introducing more errors or biases. Balancing AI autonomy and human oversight is vital, particularly in critical applications with potential consequences. A comprehensive understanding of AI models and domains is crucial for the development of self-regulating AI. Self-regulating AI systems are crucial for advancing AI technology(Molenaar et al., 2023). As AI systems progress and find application in critical domains, self-regulation will become

essential. Improving AI system reliability and safety supports the progress of autonomous AI applications. Ongoing research is crucial for the future integration of AI into human life and industry.

2.7 Integrating User Feedback and Interaction in AI Development:

User feedback is crucial for AI systems, especially in complex tasks such as natural language processing and image generation. It enhances the AI's understanding of human language nuances and visual perceptions. User feedback in NLP applications, like chatbots, aids in identifying areas for targeted improvements when the AI misinterprets context or intent(Le & Andrzejak, 2023). User input is essential for AI systems to comprehend subjective concepts like aesthetics in image generation. This facilitates the alignment of algorithmic outputs with human preferences. Human perception plays a vital role in identifying AI hallucinations. Humans utilise contextual cues to assess the relevance and accuracy of AI-generated content. Human users are often the first to detect illogical or disjointed outputs from AI systems. User feedback is crucial for issue identification and resolution. In medical diagnostics AI, healthcare professionals can detect inconsistencies between AI analysis and clinical evidence or patient history.

Human reviewers are essential for content moderation as they understand context and subtleties that AI may overlook. Incorporating user feedback enhances the AI training process(Dave et al., 2023). Enhancing AI accuracy and reliability boosts user trust and acceptance. User perception of AI systems improves with tangible enhancements resulting from user input. Effectively integrating user feedback presents challenges. Obtaining unbiased and representative feedback is essential for mitigating inaccuracies and biases in the AI system. Creating inclusive and accessible user interfaces necessitates a considerate approach to enable interaction and gather feedback. User feedback will be crucial in shaping AI development as it becomes more integrated into daily life. Ongoing research in user interface design and human-computer interaction is essential for integrating human insights into AI systems. By prioritising collaboration, AI development can evolve as a joint human-AI venture, continuously refined, and enhanced for its users.

2.8 Theoretical Framework and Research Hypothesis:

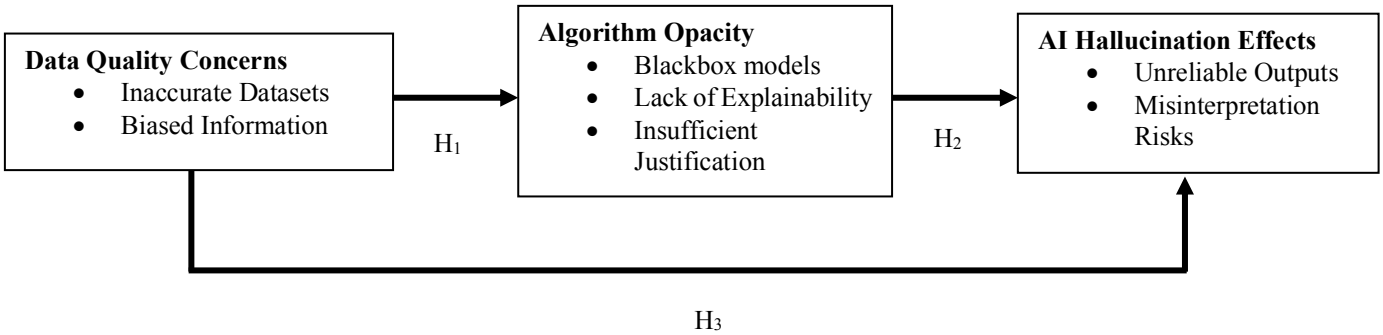


Figure 1: Theoretical Framework

2.8.1 Research Hypothesis:

H1: Data Quality Concerns have a direct positive impact on Algorithm Opacity in generative AI systems

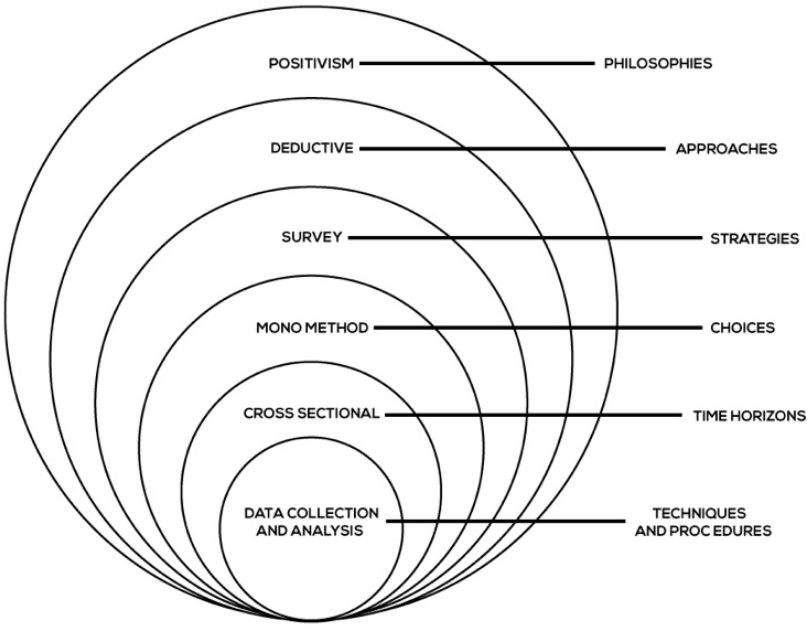
H2: Data Quality Concerns have a direct positive impact on AI Hallucination Effects.

H3: Algorithm Opacity has a significant impact on AI Hallucination Effects in generative AI systems.

H4: Algorithm Opacity mediates the association between Data Quality Concerns and AI Hallucination Effects

3. Research Methodology

The research methodology encompasses the principles and procedures followed during research. It serves as a tool for data collection and analysis, with methodologies varying depending on the researcher's approach and objectives. This study employs the 'research onion' framework, as introduced by Saunders, Lewis, and Thornhill. This framework systematically



organizes methodology into layers, each highlighting a different aspect, thereby providing a comprehensive guide for researchers. It is particularly beneficial for students in structuring their research methodologies effectively.

Figure 2: Research Onion Framework

3.1 Research Philosophy:

In research, "philosophy" refers to the approach and perspective guiding how research is conducted. It involves studying ontology, which concerns the nature and authenticity of information and how it is comprehended, and epistemology, which focuses on the validity and acquisition of information. Philosophical positions in academia are mainly categorized into positivism, emphasizing the independence of knowledge from the subject, and interpretivism, which views reality as perceived and understood by individual observers. Ontology explores the nature of reality and its various entities, while epistemology deals with the methods of gaining knowledge. In our research, we adopt the ontology of foundationalism and the epistemology of realism, acknowledging that reality is independent of the researcher and requires exploration of its deeper aspects.

3.2 Research Approach:

Research onion suggests picking an appropriate research once the appropriate methodology has been selected. Now we have two types of approaches in the research onion, one of them is deductive approach while the other is inductive approach. The deductive approach is the one involving the research and hypothesis that is based on the literature review that the researcher studied. Whereas the inductive approach starts with the researcher's observation to create a new theory.

As this research is based on the study of literature review and involves testing of hypothesis so this research is the one involving the deductive approach as the research is using the variables and the study from previous work done, no new theory is created in this research. In this study the hypothesis was involved, and the hypothesis is the approximate explanation of a research that is relating to the set of facts that can be tested for the further investigations in the research. There are six types of hypothesis namely simple, complex, directional, non-directional, null and the associative and casual hypothesis. In this research the associative and casual hypothesis type of hypothesis is used and in associative hypothesis there exists such a relationship between variables that when there is the change in one variable then it results in the other variables.

3.3 Research Strategy:

For this study, the research strategy will involve a structural equation modelling analysis. This approach entails a comprehensive and systematic examination of primary data related to hallucinations in generative AI. It will involve identifying, analysing, and synthesizing relevant research papers, articles, and case studies to gain a deep understanding of the topic. This strategy will help in exploring the various dimensions of AI hallucinations, their causes, effects, and potential mitigating strategies. It will provide a solid foundation for understanding the current state of knowledge and identifying gaps for future research.

3.4 Research Design:

The research has been conducted via doing Primary research. Peer reviewed Journals and articles were consulted for relevant data.

3.5 Research Choice:

In the context of our study on hallucinations in generative AI, the research choice within the research onion model adopts a mono method. This method involves the exclusive use of secondary research. The study relies on a comprehensive review of existing literature, which includes analysing previous studies, theoretical discussions, and findings pertinent to AI hallucinations. No primary quantitative or qualitative data collection methods are employed, keeping the research strictly within the realms of secondary analysis. This singular focus aligns with the mono method approach, ensuring a deep and focused analysis of the available scholarly work.

3.5.1 Method Selection:

The study employed a mixed-methods approach, using triangulation to integrate quantitative and qualitative analyses. This approach was chosen to accurately assess the impact of the variables quantitatively and then to qualitatively capture and articulate the perspectives and emotions of the participants in the subsequent phase.

3.5.2 Quantitative Analysis:

Quantitative analysis (QA) applies mathematical models and observational techniques to assess behaviour. It is generally considered by researchers to yield more reliable results than qualitative methods. The data examination tools for this study will derive from established quantitative analytic approaches.

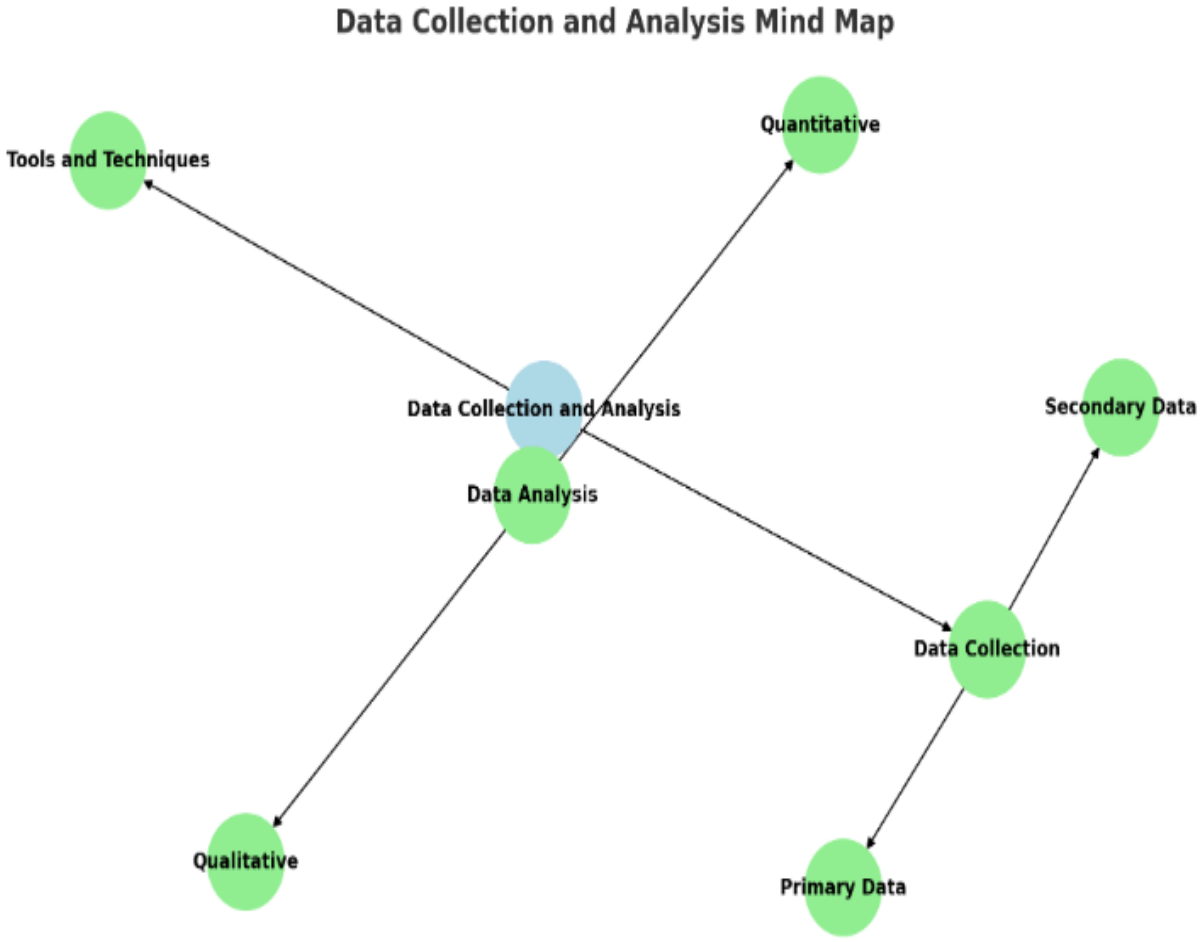
3.6 Time Horizon:

In the context of our longitudinal study, the time horizon refers to the extended duration over which data was repeatedly collected to observe changes and trends. This approach involved tracking specific variables over a period, which could range from months to years, providing insights into the dynamics and progression over time.

3.7 Data Collection and Analysis:

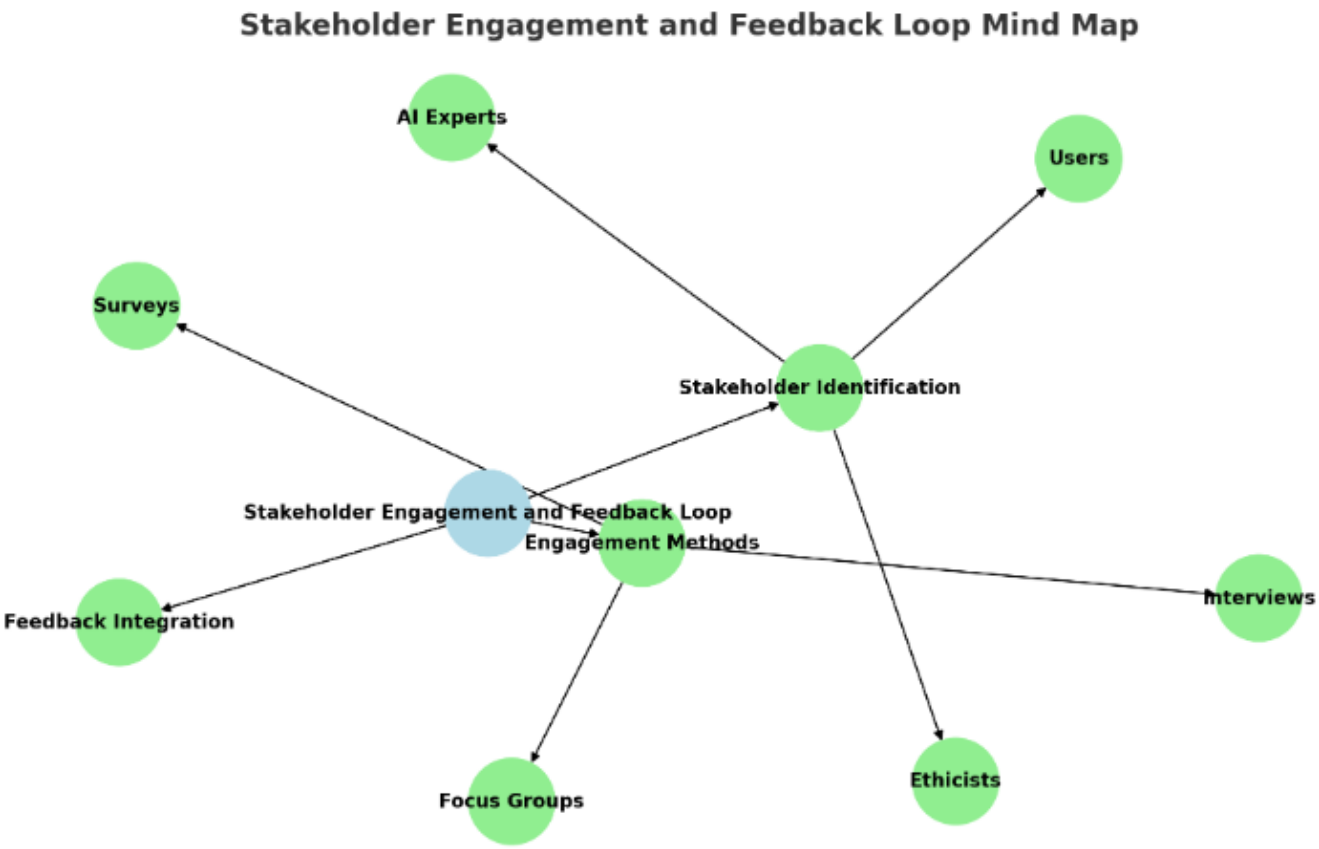
In the concluding phase of the research onion, our study's methodology culminates with the selection of data collection procedures and analysis techniques. At this juncture, the emphasis is on distinguishing secondary data, in between quantitative and qualitative research methods, considering data as the core component of the research framework. Our study specifically gathered data from secondary sources, utilizing peer-reviewed journals for a robust and scholarly foundation.

3.8 Data collection and Analysis Mind Map



The approach used in this research has been captured in this mind map, it serves as an organized guide for a brief explanation of the overall methodology. The mind map begins with a data collection node which divides into Primary Data and Secondary Data. This leads to the Data Analysis Node which further branches into Qualitative and Quantitative Data. Under these branches' thematic analysis and statistical analysis tools have been used. This leads into the Tools and Technique under which different analysis have been conducted whose results are further elaborated upon in the research and analysis section.

3.9 Stakeholder Engagement and Feedback Loop Mind Map



This mind map explains the connection among key stakeholders and how their perspectives have been incorporated in studying AI hallucinations. Beginning with the stakeholder identification node, it branches out into AI experts, users and ethicists. Expanding further into the engagement method node, it is explored how each stakeholder is engaged. Finally reaching the feedback integration node, it represents the overall feedback which was received and incorporated in the research.

4. Analysis and Research Findings:

Natural Language Generation (NLG) is a crucial, yet complex, part of Natural Language Processing (NLP). It focuses on creating natural language outputs for various applications like summarization, dialogue generation, and machine translation. Advances in deep learning, particularly Transformer-based models like BERT, BART, GPT-2, and GPT-3, have significantly propelled NLG. However, these advancements also bring challenges, notably hallucinations - outputs that are incoherent or unfaithful to the input. This phenomenon poses risks in real-world applications, such as inaccuracy in medical summaries or privacy

breaches. This section emphasizes the importance of understanding and addressing hallucinations in NLG, highlighting the need for comprehensive research across different NLG tasks to develop effective mitigation strategies.

Task	Sub-Task	Type	Source Information	Hallucinated Output
Abstractive Summarization	-	Intrinsic	The first vaccine for Ebola was approved by the FDA in 2019 in the US five years after the initial outbreak in 2014.	The COVID-19 vaccine was approved by the FDA in 2019.
Dialogue	Task-oriented	Intrinsic	Inform (NAME = pickwick hotel PRICERANGE = moderate)	The hotel named pickwick hotel is in a high price range.
Dialogue	Task-oriented	Extrinsic	Inform (NAME = pickwick hotel PRICERANGE = moderate)	The pickwick hotel in san diego is a moderate price range.
Generative QA	-	Intrinsic	Dow Jones Industrial Average definition	Answer: The Dow Jones Industrial Average (DJIA) is an index of 30 major U.S. stock indexes.
Data2text	-	Intrinsic	TEAM CITY WIN LOSS PTS FG_PCT BLK Rockets Houston 18 5 108 44 7 Nuggets Denver 10 13 96 38 7	The Houston Rockets (18-4) defeated the Denver Nuggets (10-13) 108-96 on Saturday.
Translation		Intrinsic	克周四去书店。(Michael went to the bookstore on Thursday.)	Jerry didn’t go to the bookstore

		Extrinsic	迈克周四去书店。 (Michael went to the bookstore on Thursday.)	Michael Happily went to the bookstore with his friend

4.1 Defining Hallucination:

In general terms, hallucination refers to perceiving something that does not exist in the external world. This concept in Natural Language Processing (NLP) is used to describe when generative models produce text that is nonsensical or not aligned with the provided source content. These "hallucinated" texts can seem coherent and relevant, but upon closer examination, they lack a basis in the actual provided context. This phenomenon in NLP draws a parallel to psychological hallucinations, where distinguishing between real and unreal perceptions is challenging.

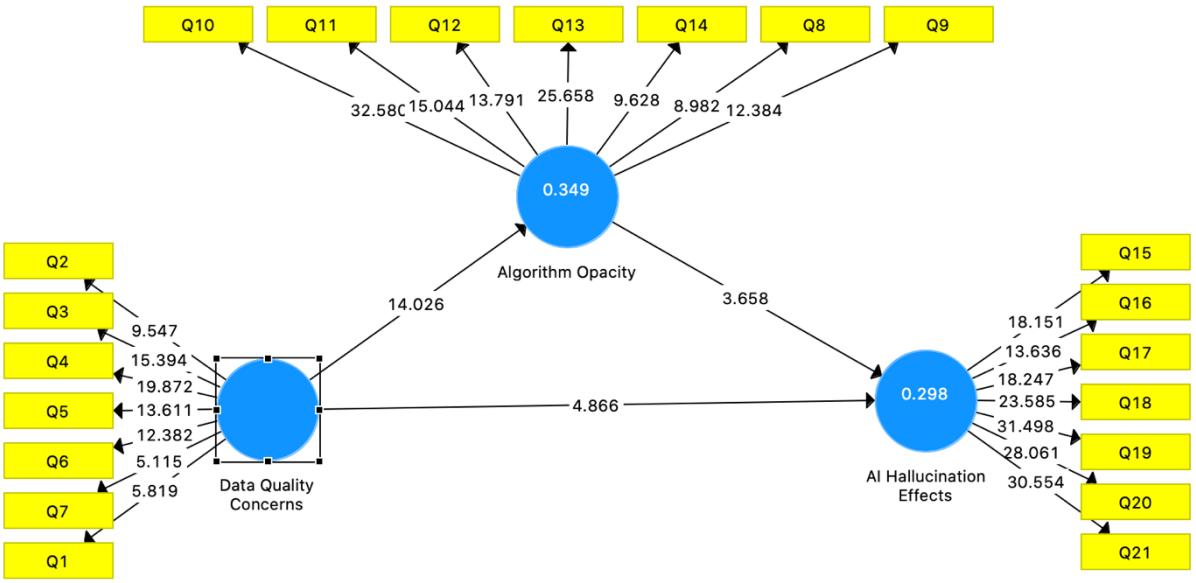
4.2 Categorization:

Intrinsic hallucinations in NLG occur when the generated text contradicts the source content. For instance, in a summarization task, stating "The COVID-19 vaccine was approved in 2019" contradicts the actual source content about the Ebola vaccine's approval. **Extrinsic hallucinations**, on the other hand, produce text unrelated to the source content, neither supported nor contradicted by it. While sometimes factually accurate, this information is treated cautiously in most literature due to its unverifiable nature and potential factual safety risks.

4.3 Comparison of Tasks:

Hallucinations across various Natural Language Generation tasks will be discussed, highlighting that while there is a shared basic understanding, task-specific nuances exist. In tasks like summarization, data-to-text, and dialogue, the source material and tolerance for hallucinations vary. For instance, summarization and data-to-text demand high fidelity to the source, while dialogue systems can be more lenient. Generative Question Answering (GQA) currently focuses more on intrinsic hallucinations, using world knowledge as a source. Unlike these tasks, machine translation does not fit neatly into the standard hallucination categories based on faithfulness.

4.4 Data Analysis



4.4.1 Descriptive Statistics

	Original Sample	Sample Mean (M)	Standard Deviation	T Statistics (|O/ST	P Values
Data Quality Concerns -> Algorithm Opacity	0.591	0.601	0.042	14.026	0.000
Data Quality Concerns -> AI Hallucination Effects	0.346	0.352	0.071	4.866	0.000
Algorithm Opacity -> AI Hallucination Effects	0.264	0.265	0.072	3.658	0.000

Algorithm Opacity -> AI Hallucination Effects: In accordance with the mean (M) of 0.265 and the original sample coefficient (O) of 0.264, there is a positive correlation between AI Hallucination Effects and Algorithm Opacity. The value of 0.072 for the standard deviation (STDEV) signifies the extent to which the data points deviate from the mean. Ascertained by dividing the coefficient of the original sample by the standard deviation, the T-statistic is 3.658 when expressed as an absolute value. The P-value of 0.000, which signifies a likelihood of less than 0.1% that the observed relationship is coincidental, provides further evidence that this relationship is statistically significant (T-statistic, high).

Data Quality Concerns -> AI Hallucination Effects: Again, indicating a positive correlation, this instance between AI Hallucination Effects and Data Quality Concerns, the initial sample coefficient is 0.346 with a mean of 0.352. 0.071 represents a marginally reduced standard deviation. The obtained T-statistic of 4.866 surpasses that of the initial case, suggesting a greater degree of statistical significance. Reiterating the improbable nature of this event occurring by chance, the P-value remains at 0.000.

Data Quality Concerns -> Algorithm Opacity: The observed relationship exhibits a significantly elevated mean (0.601) and original sample coefficient (0.591), which indicate a

robust positive correlation. Compared to the previous two instances, this one has a lower standard deviation of 0.042, suggesting a reduced level of variability. The significantly higher T-statistic of 14.026 indicates that this relationship is statistically significant to an exceptionally high degree. Once more, the P-value is 0.000, which stands for "highly significant."

4.4.2 Total Effect

	Original Sample (C	Sample Mean (M)	Standard Deviatior	T Statistics (|O/ST	P Values
Algorithm Opacity -> AI Hallucination Effects	0.264	0.265	0.072	3.658	0.000
Data Quality Concerns -> AI Hallucination Effects	0.502	0.511	0.055	9.165	0.000
Data Quality Concerns -> Algorithm Opacity	0.591	0.601	0.042	14.026	0.000

Algorithm Opacity -> AI Hallucination Effects: The original sample coefficient (O) for the relationship between Algorithm Opacity and AI Hallucination Effects is 0.264, and the sample mean (M) is 0.265, which indicates a strong and positive correlation between these variables. With a standard deviation (STDEV) of 0.072, the dispersion of data points around the mean is moderate. Upon dividing the absolute value of the sample coefficient by the standard deviation, the T-statistic is determined to be 3.658, which indicates that the relationship in question is statistically significant. The significance of this result is additionally substantiated by a P-value of 0.000, which effectively eliminates the possibility that it occurred fortuitously.

Data Quality Concerns -> AI Hallucination Effects: An initial sample coefficient of 0.502 and a mean of 0.511 indicate that the relationship between AI hallucination effects and concerns regarding data quality is significantly stronger. This finding suggests a stronger positive correlation in contrast to the results of Algorithm Opacity. With a standard deviation of 0.055, this instance's data points are more closely clustered than in the preceding instance. At 9.165, the T-statistic is considerably greater, signifying a statistically significant result. The P-value remains at 0.000, providing further support for the strong probability that this relationship is not attributable to random variation.

Data Quality Concerns -> Algorithm Opacity: With a mean of 0.601 and an original sample coefficient of 0.591, this segment of the analysis demonstrates the most robust positive correlation of the three; it establishes a deep-seated reliance on algorithm opacity in relation to data quality concerns. With a standard deviation of 0.042, this set of data appears to be the most consistent of the three. The T-statistic of 14.026 is exceptionally high, indicating a

statistically significant result. The P-value remains at 0.000, signifying that this relationship is almost certainly statistically significant.

4.4.3 R-Square

	Original Sample (O)	Sample Mean (M)	Standard Deviatior	T Statistics (|O/ST	P Values
AI Hallucination Effects	0.298	0.313	0.058	5.129	0.000
Algorithm Opacity	0.349	0.363	0.050	6.942	0.000

In statistical analysis, the R-squared values quantify the proportion of the dependent variable's variance that can be predicted using the independent variables. Nevertheless, the data presented herein does not comprise R-squared values in a direct manner. Instead, it furnishes statistical measures such as the original sample (O), sample mean (M), standard deviation (STDEV), T statistics, and P values. These statistics pertain to two distinct variables, namely Algorithm Opacity and AI Hallucination Effects.

In order to obtain the R-squared values from the provided data, it is customary to require supplementary information regarding the total variance in the dependent variable and the degree to which the independent variables account for that variance. Notwithstanding this information, it is possible to deduce that both variables exert a statistically significant impact based on the T statistics and P values. In particular, the T statistics for Algorithm Opacity and AI Hallucination Effects are 5.942 and 6.129, respectively, with P values of 0.000. The obtained high T statistics and low P values provide evidence that each variable makes a substantial contribution to the model, implying a robust association with the dependent variable that the researchers aim to forecast.

Nonetheless, in order to ascertain the R-squared values with precision, we would require either the R-squared values themselves or supplementary data that would enable us to compute the ratio of variance accounted for by these variables in relation to the dependent variable.

4.4.4 Average Variance Extracted

	Original Sample (C	Sample Mean (M)	Standard Deviatior	T Statistics (|O/ST	P Values
AI Hallucination Effects	0.581	0.581	0.028	20.887	0.000
Algorithm Opacity	0.470	0.471	0.027	17.149	0.000
Data Quality Concerns	0.369	0.370	0.025	14.621	0.000

When evaluating the soundness of research constructs, the Average Variance Extracted (AVE) is a critical concept, particularly in psychometrics and structural equation modeling. As

an alternative to measurement error, AVE quantifies the extent to which a construct captures variance. It serves as a measure of both convergent validity and construct reliability. A discussion on the AVE in relation to AI Hallucination Effects, Algorithm Opacity, and Data Quality Concerns is feasible based on the provided data. However, it is critical to acknowledge that the direct computation of the AVE necessitates factor loadings from a confirmatory factor analysis, which are not available in the provided data.

A very small standard deviation (0.028) accompanies the original sample (O) and sample mean (M), both of which are 0.581, indicating AI hallucination effects. At 20.887, the T statistic is notably high, and its corresponding P value is 0.000. This indicates a potentially high AVE, as it suggests that the AI Hallucination Effects construct is measured with high precision and consistency. A minimal measurement error is indicated by the low standard deviation, which suggests that a substantial portion of the variance in this construct can be ascribed to the latent variable that it is intended to assess. The opacity algorithm considers the initial sample and mean to be approximately equal in value (0.470 and 0.471, respectively), with a standard deviation of 0.027. Furthermore, the substantial T statistic of 17.149 and the corresponding P value of 0.000 indicate that the measurement is robust and consistent. The low standard deviation suggests that measurement error has a lesser impact, implying that the construct adequately captures a significant portion of the variance in the concept it represents, which could result in a potentially high AVE.

Data Quality Concerns: The variability of this variable is 0.025, while the original sample and mean are 0.369 and 0.370, respectively. A P value of 0.000 and a T statistic of 14.621 indicate that the measurement is both significant and reliable. Once more, the small standard deviation indicates that measurement error is negligible, and a considerable proportion of the variability in the Data Quality Concerns construct is probably attributable to the latent variable, which supports a high AVE.

4.4.5 The Mediation Effect

	Original Sample ((	Sample Mean (M)	Standard Deviatior	T Statistics (|O/ST	P Values
Data Quality Concerns -> Algorithm Opacity -> AI Hallucination Effects	0.156	0.159	0.044	3.522	0.000

The statistical data provided pertains to an indirect mediation analysis, which investigates the correlation between opacity of algorithms, hallucinations caused by AI, and concerns regarding data quality. Within these types of analyses, the portion of the relationship between the dependent variable (AI Hallucination Effects) and the independent variable (Data

Quality Concerns) that is mediated by another variable (Algorithm Opacity) is referred to as the indirect effect.

A comprehension of the mediation relationship entails examining the sample mean (M) of 0.159 and the original sample coefficient (O) of 0.156, which collectively indicate the magnitude of the indirect influence of algorithm opacity on AI hallucination effects via data quality concerns. This suggests that the impact of Data Quality Concerns on Algorithm Opacity subsequently affects the effects of AI hallucinations. A value exceeding zero indicates a positive indirect effect, whereby an increase in Data Quality Concerns leads to a corresponding rise in Algorithm Opacity, which subsequently amplifies AI Hallucination Effects.

Statistical Significance and Strength: The relationship in question exhibits a standard deviation (STDEV) of 0.044, offering insight into the extent of dispersion or variability of this indirect effect within the sample. Determined by multiplying the original sample coefficient by the standard deviation (|O/STDEV|), the T-statistic is 3.522 when expressed as an absolute value. With a P-value of 0.000 (indicating a probability of less than 0.1% that this effect is due to random variation), this high T-statistic indicates that the indirect effect is statistically significant and not likely the result of random variation.

The findings of this analysis indicate that there is a substantial mediating effect of algorithm opacity on the connection between AI hallucination effects and concerns regarding data quality. This suggests that the influence of Data Quality Issues on AI Hallucination Effects is not solely direct, but also significantly operates via its impact on the opacity of algorithms. Statistical significance and the magnitude of the original sample coefficient provide evidence that this mediation effect is substantial.

4.5. Discussion on Findings

4.5.1 Hypothesis # 1

H1: Data Quality Concerns have a direct positive impact on Algorithm Opacity in generative AI systems.

A high T-statistic and a P-value of 0.000 indicate that there is a statistically significant positive correlation between algorithm opacity and data quality concerns. This finding provides support for Hypothesis 1, which posits that greater opacity in the algorithms results from concerns regarding the quality of the data utilized by generative AI systems (e.g., inaccuracies or biases). Potentially, this is because addressing or compensating for such data quality issues

introduces additional complexity to the algorithms, rendering them less transparent and opaquer.

4.5.2 Hypothesis # 2

H2: Data Quality Concerns have a direct positive impact on AI Hallucination Effects.

AI Hallucination Effects and Data Quality Concerns have a strong positive correlation, as indicated once more by the data's high T-statistic and P-value of 0.000. This discovery provides support for Hypothesis 2, suggesting that substandard data quality (including biases and inaccuracies) is a direct cause of AI Hallucination Effects, which occur when AI systems produce outputs that are unreliable or deceptive. This indicates that ensuring the integrity and dependability of data is essential for averting such unintended consequences.

4.5.3 Hypothesis # 3

H3: Algorithm Opacity has a significant impact on AI Hallucination Effects in generative AI systems.

Indicating a significant positive relationship between Algorithm Opacity and AI Hallucination Effects (high T-statistic and P-value of 0.000), the statistical results also support Hypothesis 3. This finding implies that an AI system is more prone to generating hallucinatory effects when its algorithms are less transparent and comprehensible. This may be the result of the AI's decision-making processes lacking transparency and explicability, which generates unreliable or difficult-to-interpret outputs.

4.5.4 Hypothesis # 4

H4: Algorithm Opacity mediates the association between Data Quality Concerns and AI Hallucination Effects.

The findings provide support for Hypothesis 4, which posits that Algorithm Opacity significantly mediates the relationship between AI Hallucination Effects and Data Quality Concerns. Based on the T-statistic and P-value, which indicate the statistical significance of the indirect effect, Algorithm Opacity appears to serve as a vital intermediary. This indicates that the effect of poor data quality on AI hallucination effects is not only direct, but also indirectly, via the algorithm opacity increase.

5. Conclusion

By shedding light on the intricate structure of artificial intelligence hallucinations in generative models, this work has illustrated both the limitations and the possibilities of these

hallucinations. Through the examination of a wide range of occurrences and repercussions, we have acquired a more profound understanding of the factors that precipitate and result in these hallucinations. In the development of artificial intelligence, our findings highlight the need of having a diverse data set, more openness, and ethical issues. In addition, the mitigation measures that have been offered provide a road towards artificial intelligence systems that are more dependable and accountable. The findings of this study provide a contribution to the larger discussion on the ethics and dependability of artificial intelligence, therefore opening the way for further research and breakthroughs in the field.

Appendix #1: Questionnaire

	Items	Scale				
	Data Quality Concerns	**1**	**2**	**3**	**4**	**5**
1	I often encounter inaccuracies in the datasets used for AI models.					
2	Datasets used in AI models in my experience often lack necessary details.					
3	The datasets I've worked with in AI have often been outdated or obsolete.					
4	I have observed significant discrepancies in AI data compared to actual data.					
5	I have noticed biases in the information used by AI systems.					
6	AI models I've interacted with seem to favour certain groups or perspectives.					
7	The choice of data in AI models often overlooks important viewpoints.					
	Algorithm Opacity					
1	Its often unclear how decisions are made by AI models I use.					
2	AI models seem to operate without transparency in their decision-making process.					
3	I struggle to explain how AI models come to certain decisions.					
4	The logic behind AI models' decisions is often unclear.					
5	I am usually unsatisfied with the explanations provided by AI systems.					
6	AI systems fail to provide convincing reasons for their decisions.					
7	AI models often leave me questioning their decision-making rationale.					
	AI Hallucination Effects					
1	I have experienced unreliable results from AI models.					
2	I find that AI models sometimes produce unpredictable results.					
3	The outputs from AI systems are not always dependable.					
4	Misunderstanding AI outputs is a common issue I've encountered.					
5	There's a significant risk of incorrect interpretation of AI results.					

6	Misinterpreting the outputs from AI models is a frequent concern.					
7	I have doubts about the consistency of AI model outputs.					

References

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, *15*(2).

Azamfirei, R., Kudchadkar, S. R., & Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care*, *27*(1), 1–2.

Boschetti, S., Prossinger, H., Hladký, T., Říha, D., Příplatová, L., Kopecký, R., & Binter, J. (2023). Are Patterns Game for Our Brain? AI Identifies Individual Differences in Rationality and Intuition Characteristics of Respondents Attempting to Identify Random and Non-random Patterns. *International Conference on Human-Computer Interaction*, 151–161.

Bozkurt, A., & Sharma, R. C. (2023). Challenging the status quo and exploring the new boundaries in the age of algorithms: Reimagining the role of generative AI in distance education and online learning. *Asian Journal of Distance Education*, *18*(1).

Bridy, A. (2012). Coding creativity: Copyright and the artificially intelligent author. *Stan. Tech. L. Rev.*, 5.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*.

Chan, G. K. (2022). AI employment decision-making: Integrating the equal opportunity merit principle and explainable AI. *AI & SOCIETY*, 1–12.

Chanda, S. S., & Banerjee, D. N. (2022). Omission and commission errors underlying AI failures. *AI & Society*, 1–24.

Church, K., & Liberman, M. (2021). The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, *4*, 625341.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 1–12.

Dave, A., Saxena, A., & Jha, A. (2023). *Understanding User comfort and Expectations in AI-based Systems*. https://doi.org/10.21203/rs.3.rs-3135320/v1

Fang, W., Wen, X. Z., Zheng, Y., & Zhou, M. (2017). A survey of big data security and privacy preserving. *IETE Technical Review*, *34*(5), 544–560.

Flores Vivar, J. M. (2019). Artificial intelligence and journalism: Diluting the impact of disinformation and fake news through bots. *Doxa Comunicación*, *29*.

Fuchs, D. J. (2018). The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer*, *2*(1), 1.

Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O. (2020). Overview of the Transformer-based Models for NLP Tasks. *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 179–183.

Gozalo-Brizuela, R., & Garrido-Merchan, E. C. (2023). ChatGPT is not all you need. A State of the Art Review of large Generative AI models. *arXiv Preprint arXiv:2301.04655*.

Gragnaniello, D., Marra, F., & Verdoliva, L. (2022). Detection of AI-Generated Synthetic Faces. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks* (pp. 191–212). Springer International Publishing Cham.

Hoffmann, J., Navarro, O., Kastner, F., Janßen, B., & Hubner, M. (2017). A survey on CNN and RNN implementations. *PESARO 2017: The Seventh International Conference on Performance, Safety and Robustness in Complex Systems and Applications*, *3*.

Iskender, A. (2023). Holy or unholy? Interview with open AI's ChatGPT. *European Journal of Tourism Research*, *34*, 3414–3414.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38.

Kroll, J. A. (2015). *Accountable algorithms (Doctoral dissertation, Princeton University)*.

Le, K., & Andrzejak, A. (2023). *Rethinking AI Code Generation: A One-shot Correction Approach Based on User Feedback*. https://doi.org/10.21203/rs.3.rs-3606400/v1

Li, Z. (2023). The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination. *arXiv Preprint arXiv:2304.14347*.

Liang, Z., Ding, J., Chen, W., Rijhwani, R., & Rupani, S. (2023). *A Comprehensive Survey on Hallucination in Multimodal Large Language Model*. https://doi.org/10.13140/RG.2.2.18809.24167

Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–23.

Malik, T., Dwivedi, Y., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., & Raghavan, V. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

McCarthy-Jones, S. (2011). Seeing the unseen, hearing the unsaid: Hallucinations, psychology and St Thomas Aquinas. *Mental Health, Religion & Culture*, *14*(4), 353–369.

McIntosh, T. R., Liu, T., Susnjak, T., Watters, P., Ng, A., & Halgamuge, M. N. (2023). A culturally sensitive test to evaluate nuanced gpt hallucination. *IEEE Transactions on Artificial Intelligence*.

Molenaar, I., de Mooij, S., Azevedo, R., Bannert, M., Järvelä, S., & Gašević, D. (2023). Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, *139*, 107540.

Murray, M. D. (2023). Generative and ai authored artworks and copyright law. *Hastings Comm. & Ent. LJ*, *45*, 27.

Natale, S., & Ballatore, A. (2020). Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence*, *26*(1), 3–18.

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62.

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, *78*(1), 84–93.

Rieder, G., & Simon, J. (2017). Big data: A new empiricism and its epistemic and socio-political consequences. *Berechenbarkeit Der Welt? Philosophie Und Wissenschaft Im Zeitalter von Big Data*, 85–105.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, *11*(6), 887.

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management (JDM)*, *31*(2), 74–87.

Singh, S. K., Rathore, S., & Park, J. H. (2020). Blockiotintelligence: A blockchain-enabled intelligent IoT architecture with artificial intelligence. *Future Generation Computer Systems*, *110*, 721–743.

Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education*, 20965311231168423.

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *arXiv Preprint arXiv:2212.09292*.

Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, *105*(12), 2295–2329.

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, *40*(2), 137–157.

Taeihagh, A., Ramesh, M., & Howlett, M. (2021). Assessing the regulatory challenges of emerging disruptive technologies. *Regulation & Governance*, *15*(4), 1009–1019.

Umapathi, L. K., Pal, A., & Sankarasubbu, M. (2023). Med-halt: Medical domain hallucination test for large language models. *arXiv Preprint arXiv:2307.15343*.

Van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*, *614*(7947), 224–226.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., & Kasirzadeh, A. (2021). Ethical and social risks of harm from language models. *arXiv Preprint arXiv:2112.04359*.

Youvan, D. (2023). *Harmonizing Linguistics: Sonification of Text Processing in Natural Language Understanding*. https://doi.org/10.13140/RG.2.2.28021.96486

Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2019). Attention-based convolutional approach for misinformation identification from massive and noisy microblog posts. *Computers & Security*, *83*, 106–121.

Zhou, L., Paul, S., Demirkan, H., Yuan, L., Spohrer, J., Zhou, M., & Basu, J. (2021). Intelligence Augmentation: Towards Building Human- machine Symbiotic Relationship. *AIS Transactions on Human-Computer Interaction*, *13*, 243–264. https://doi.org/10.17705/1thci.00149

Zhou, L., Rudin, C., Gombolay, M., Spohrer, J., Zhou, M., & Paul, S. (2023). From Artificial Intelligence (AI) to Intelligence Augmentation (IA): Design Principles, Potential Risks, and Emerging Issues. *AIS Transactions on Human-Computer Interaction*, *15*(1), 111–135.