How does Big Data affect the creation and application of AI models, and what effect does it have on the AI industry's advancements

Table of Contents

1. Introduction

1.1 Research Background

The concept of Artificial Intelligence (AI) was first introduced to the world in 1956 during the Dartmouth Conference. The term "artificial intelligence" was brought forward by Jhon McCarthy in a proposal for the conference, which aimed to explore the possibility of creating computers that could simulate human intelligence. The potential of AI came to the attention of the public in the 1960's, after the development of programs like ELIZA, and GPS which had the potential of AI. The significant resurgence of AI, referred to as "AI Boom" began in early 2010s. Developments in industries such as Machine learning and Deep learning, increased computational power, Big Data, successful applications, and commercial interest and investment collectively contributed to the uprising and widespread of AI technology. The study follows how does Big Data affect the creation and application of AI models and what effect it has on the advancement of AI industry.

Big Data has a considerable impact on the AI industry: from transforming how AI systems are developed and trained, to their deployment. It drives the growth and effectiveness of the AI industry by providing the necessary resources for model training, enabling real-time application, and supporting ongoing research and development.

1.2 Purpose of the Research

The intersection of Big Data and the AI industry affords a complex and unexpectedly evolving landscape, with implications spanning technical, moral, and sector-particular dimensions. The purpose of this study is to find a comprehensive understanding of the multifaceted influences of Big Data on AI models and the enterprise as a whole. Despite the substantial advancements in Big Data and AI, there lies a constant occurring problem in know-how for the nuanced, demanding situations, opportunities, quality of resources, and ethical concerns that emerge from the integration of Big Data on AI. This study seeks to address this problem by delving into the problematic dynamics, aiming to contribute treasured insights to the evolving subject of data-driven artificial intelligence.

1.3 Significance of the Research

The motivation that drew the research was the profound impact of Big Data on modeling artificial intelligence programs and all of AI industry. Big Data fuels the application of AI in various sectors like healthcare, finance, retail, and more. Product and services companies use AI to automate complex tasks. The startups and innovation hubs focused on AI has skyrocketed because of the application of Big Data on AI. But the usage of Big Data requires a massive collection of Data.

This raises questions about the privacy, security, and the etiquette of the data collected on a massive scale. In technical level, the biasness and fairness of the data used to train the AI models raises concerns. The prospect of uncovering how the abundance of facts affects AI model education, ethical considerations, and sector-unique applications serves as a using pressure to discover and make a contribution to this unexpectedly evolving field. Recognizing the present gap in the understanding of the complex interactions between Big Data and AI, this research pursuits to fill that void by way of undertaking a comprehensive evaluation. By identifying and addressing the gaps in cutting-edge information, the study intends to offer a basis for future studies and contribute to the continuing discourse inside the AI industry.

The task is closely related to diverse subjects, including records technological know-how, system gaining knowledge of, and ethics in era. The integration of these subjects is important to getting a holistic expertise of the challenges and possibilities presented via Big Data in the development of AI.

1.4 Research Objectives

01. Assess the Impact of Big Data on AI Model Training & Performance
    - Explore how the provision of Big Data influences the education and overall performance of AI models across various domains.
    - Examine the consequences for accuracy, adaptability, and real-world applications in unique sectors.
02. Identify Challenges and Opportunities in Using Big Data for AI Models
    - Examine the main problems and opportunities related with using Big Data to build AI models.
    - Analyze the effects of those factors on the scalability and efficiency of AI programs, especially in dealing with huge datasets.
03. Examine the Facilitation of Advanced AI Algorithms by using Big Data
    - Explore the ways in Big Data allows the advent of complicated AI algorithms.
    - Identify sectors or programs that revel in sizable advancements and analyze the sensible implications of those developments.
04. Evaluate Ethical Considerations in Personal Data Usage for AI Models
    - Examine ethical issues bobbing up from the usage of sizable quantities of private data to educate AI models.
    - Investigate industry standards and practices in area to deal with and mitigate these ethical issues in the AI industry.

1.5 Research Questions

01. How does the supply of big data impact the education and performance of AI models in various domain names, and what are the implications for accuracy and flexibility in actual global applications?
02. What are the number one challenges and opportunities related to using big data to assemble AI models, and how do these factors impact the scalability and performance of AI packages, specifically in handling huge datasets?
03. In what ways does the combination of big data facilitate the improvement of tricky AI algorithms, and which sectors or packages enjoy the massive improvements, highlighting the sensible implications of these traits?
04. How do the ethical concerns bobbing up from the utilization of considerable quantities of personal statistics for schooling AI fashions impact the AI enterprise, and what industry requirements and practices are in place to cope with and mitigate these ethical concerns?

1.6 Hypothesis

The hypothesis and assumptions used to evaluate the objective of the study are as follows.

When tailoring the questions, it was taken as an assumption that data quality, computational resource constraints, and privacy concerns are the main challenges faced in the case, as preliminary data found in the literature review suggested.

The analysis of all questions and objectives all together seemed to be a bit of a challenge, therefore some objectives like identifying the challenges and evaluating ethical concerns were done in descriptive analysis. However, the applicable part of those were taken into calculations. For instance, how often does one using Big Data to create and apply AI models face the mentioned challenges and how does it affect the advancement of AI industry was taken into account when doing calculations.

The main hypothesis to build the regression model was as follows.

H_0: When applying Big Data into AI models, the volume, accuracy, variety, quality of Big Data, the ethical and privacy concerns regarding its application into AI, the understanding of those ethical and privacy concerns of the public, and the computational resource constraints regarding the application of Big Data affects significantly to the advancement of the AI industry.

H_1: Those variables do not have a significant effect on the advancement of AI industry.

Models will be built using stepwise regression analysis and the hypothesis will be tested later in the study.

2. Literature Review

In A. Klimczuk's article titled "The Big Data World: In the published article titled “Benefits, Threats and Ethical Challenges” in Emerald Open Research 1, an outstanding job is done by over-viewing the big data and artificial intelligence (AI) landscape from various perspectives. The study systematically delves into the enormous advantages that come with big data and AI technologies’ fusion, creating a profound effect on diverse spheres. At the same time, this article critiques connected threats by identifying potential risks and weaknesses inherent in widespread use of these technologies. Discourse has an important ethical component, exploring moral ambiguities and predicaments in the face of ubiquitous utilization of big data and AI. The topic central to the discourse is data ownership and privacy protection. With respect to the field of big data and AI, Klimczuk’s work is a welcome contribution that provides valuable insights in an ongoing conversation about the ethical dimensions emerging technologies. (Iphofen and O’Mathúna, 2021). The article provides a thorough analysis of the diverse terrain surrounding big data and artificial intelligence (AI) Klimczuk’s contribution presents a systematic investigation of multiple benefits resulting from the combination of big data and artificial intelligence technologies, focusing on their disruptive role in different sectors. At the same time, an article critically analyzes embedded dangers present in these technologies since they can be potential threats because of risks and vulnerabilities. A crucial topic of discussion is the ethical implications to massive big data and AI utilization, as major conundrums are addressed in their implementation. In addition, the article touches upon intricate problems of data possession and emphasizes that human rights have to be protected. In this regard, Klimczuk’s contribution offers important ethical insights into emerging technologies and thus makes the work a vital source of information for scholars as well as practitioners in regard to big data AI. (Iphofen and O’Mathúna, 2021)

In his article “how Big Data is Advancing AI at Scale” written by R. Bean in MIT Sloan Management Review Vol 2, the author addresses a critical topic concerning relationship between big data and progress made on artificial intelligence (AI) and machine learning The article touches upon a core role big data plays in deepening AI and ML technologies’ capabilities of scalability for massive application. Through the examination of big data’s fusion with AI, the author reveals how organizations grasp synergy between these technologies to derive strategic insights, aide decision-making process and enhance innovation. In this light, Bean’s work offers an intriguing worldview on the underlying interplay between big data, AI and business outcomes; hence making it a relevant resource to scholars interested in understanding how the new paradigm of developing technologies is transforming everyday lives. (Bean, 2017)

In the article “Privacy-Preserving Machine Learning: Securing Data in AI Systems”, J.Atetedaye has stated that “As data-driven insights continue to shape various sectors, the need to protect individual privacy has become paramount” inclining the necessity of the data privacy when using data in Machine Learning, a fundamental theory of AI. The article dives into the world of privacy-preserving machine learning (PPML), explaining the techniques and methods used to ensure data security while training AI models using machine learning. The author provides a comprehensive understanding of the mechanisms used to ensure data privacy, like homomorphic encryption, federated learning, and secure multi-party computation which enables the reader to learn about the strengths and weaknesses of privacy and ethical aspects of the Big Data usage in machine learning. (J.Atetedaye, 2024)

The article “Regulatory Compliance and Ethical Considerations: Compliance challenges and opportunities with the integration of Big Data and AI” by K. Hubert, and F. Oluyinka talks about the multifaceted landscape of compliance challenges, encompassing data privacy, security, and algorithmic transparency, alongside the evolving ethical considerations in AI and Big Data. It is stated that data quality, privacy, security, resource constraints like biasness and fairness are the main challenges faced when integrating Big Data in AI. Ensuring Proper Data Encryption and Storage, Minimizing Data Breaches and Unauthorized Access, Obtaining and Managing User Consent for Data Processing, Ensuring Transparency in AI Decision-Making Processes, and Establishing Accountability Frameworks for AI Systems were given as solutions to the ethical and privacy concerns regarding the usage of Big Data in AI industry. It was said that the benefits of using Big Data on AI outweighs the negative side of it. (K. Hubert, and F. Oluyinka, 2024)

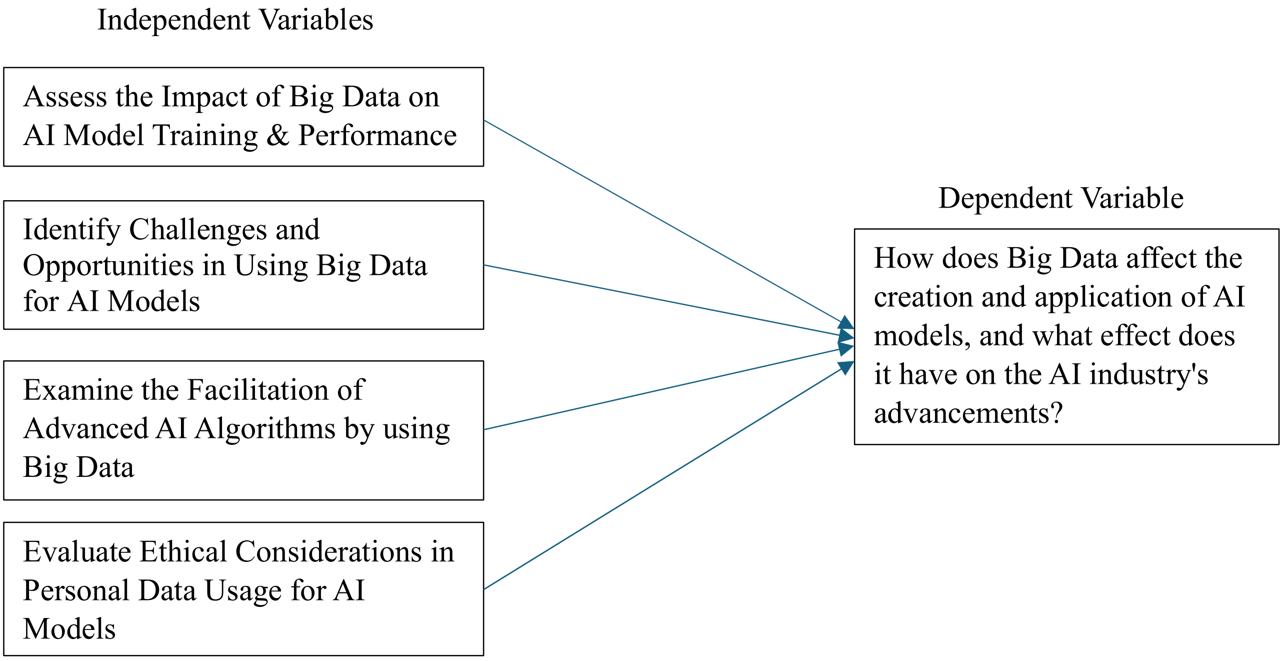
In the article "Privacy and Security of Big Data in AI Systems: In the third article by M. A. Alsheikh titled “A Research and Standards Perspective,” published on IEEE Xplore 3, the author delves into a focused discussion of critical areas around big data privacy and security issues within AI systems Alsheikh’s work is critical to analyze the issues and concerns regarding maintaining confidentiality alongside integrity of large datasets which are an integral part of AI applications. The paper not only discusses the technical aspects but also moves to ethics, specifically touching upon moral implications of using personal data for training AI models. The author takes research, and standards approach to the scholarly debate, providing some information that is valuable in terms of understanding and reducing potential privacy threats as well as security challenges present at this conjunction. Alsheikh’s work becomes a major resource for researchers and practitioners to arrive at the ethical landscape as well as set strong rules in AI systems that need large datasets. (Dilmaghani et al., 2019)

The MDPI 1 compilation, “Big Data and Artificial Intelligence”, is a treasured collection of research papers investigating the significant influence that big data has on AI models. The collection of topics covered in this great book supports a wide variety, including enabling technologies methodologies frameworks and models pertaining to AI big data research. The repository also investigates the best practices and real life scenarios, offering an in-depth analysis of modern trends observed. In addition, the consideration of ethical issues associated with big data and AI reflects that they are ensuring responsible and conscientious use such technologies. Therefore, as a comprehensive source of information MDPI 1’s collection turns into an informative and multifaceted reference for the scholars; practitioners who are interested in understanding all aspects associated with big data and AI development. (www.mdpi.com, n.d.)

The web resource Understanding Data Bias & Impact on AI/ML Decision Making made available by Progress 2 is a thought-provoking exploration of the ethical aspects in connection with artificial intelligence (AI) and machine learning(ML). This resource under discussion is focused on ethical implications of AI and ML systems. Data bias, and the subsequent effect it has on how decisions are made within these technologies remains a central focus in this discussion. As the website goes into more detail about data bias, it manages to contribute towards a nuanced understanding of potential pitfalls and ramifications concerning fairness and equality under AI/ML applications. The resource of Progress2 is an appropriate reference for the researchers and practitioners who want to learn about ethical aspects that should be taken into account in order to eliminate bias in AI as well as machine learning, which has recently emerged.(Progress Blogs, 2023)

IEEE Xplore 3, ‘The Impact of Big Data on AI’, offers an interesting analysis in the area of big data and its utility for applying techniques based on artificial intelligence (AI) methods. The webpage takes an in-depth look at the revolution big data has brought upon how AI models are built and used. The resource not only identifies opportunities for improving the performance of AI systems through large datasets but also discusses the challenges related to big data in building such models. Through a subtle analysis of the synergies, challenges and possibilities that occur at the intersection between big data and AI, IEEE Xplore 3 becomes an essential source for researchers or practitioners who want to delve deeper into this ever-changing field in which new methodologies are driven by Big Data. (ieeexplore.ieee.org, n.d.)

2.1 Conceptual Framework


Independent Variables
Assess the Impact of Big Data on AI Model Training & Performance
Identify Challenges and Opportunities in Using Big Data for AI Models
Examine the Facilitation of Advanced AI Algorithms by using Big Data
Evaluate Ethical Considerations in Personal Data Usage for AI Models
Dependent Variable
How does Big Data affect the creation and application of AI models, and what effect does it have on the AI industry's advancements?

The conceptual framework diagram shows how different factors (independent variables) influence the main question of the study (dependent variable).

Independent Variables:

- Assess the Impact of Big Data on AI Model Training & Performance:
  - This helps us to have an understanding on how having Big Data helps or affects how AI models are trained and how well they perform.

- Identify Challenges and Opportunities in Using Big Data for AI Models:
    - This helps us explore the upsides and downside of using big data to build AI models, like issues in data quality or how it can improve accuracy of AI models.
- Examine the Facilitation of Advanced AI Algorithms by Using Big Data:
    - This helps us to examine how big data helps create more complex and advanced AI algorithms than those that are already in use.
- Evaluate Ethical Considerations in Personal Data Usage for AI Models:
    - This helps us get an understanding on the ethical issues of using large amounts of personal data to train AI models and how this affects public trust on AI industry.

Dependent Variable:

- How Does Big Data Affect the Creation and Application of AI Models, and What Effect Does It Have on the AI Industry's Advancements?
    - This addresses the main question in the study, understanding the overall impact of big data on developing and using AI models and how it influences the AI industry, by analyzing the independent variables with the dependent variable.

This framework helps us organize the study by showing how different factors related to big data influence the development and use of AI models.

3. Methodology

3.1 Research Onion

In its most basic version, Saunders' research onion outlines the various choices you'll have to make while creating a research technique, be it for a thesis, dissertation, or other official research endeavor. You'll encounter a variety of decisions as you work your way inside the onion from the outside in. These decisions range in complexity from high-level and philosophical to tactical and practical. This follows the overall format of the chapter on methodology as well. Saunders' research onion is a helpful tool for considering technique in a comprehensive way, despite its obvious flaws. It aids in understanding the choices you must make about your study design and methods, at the very least.

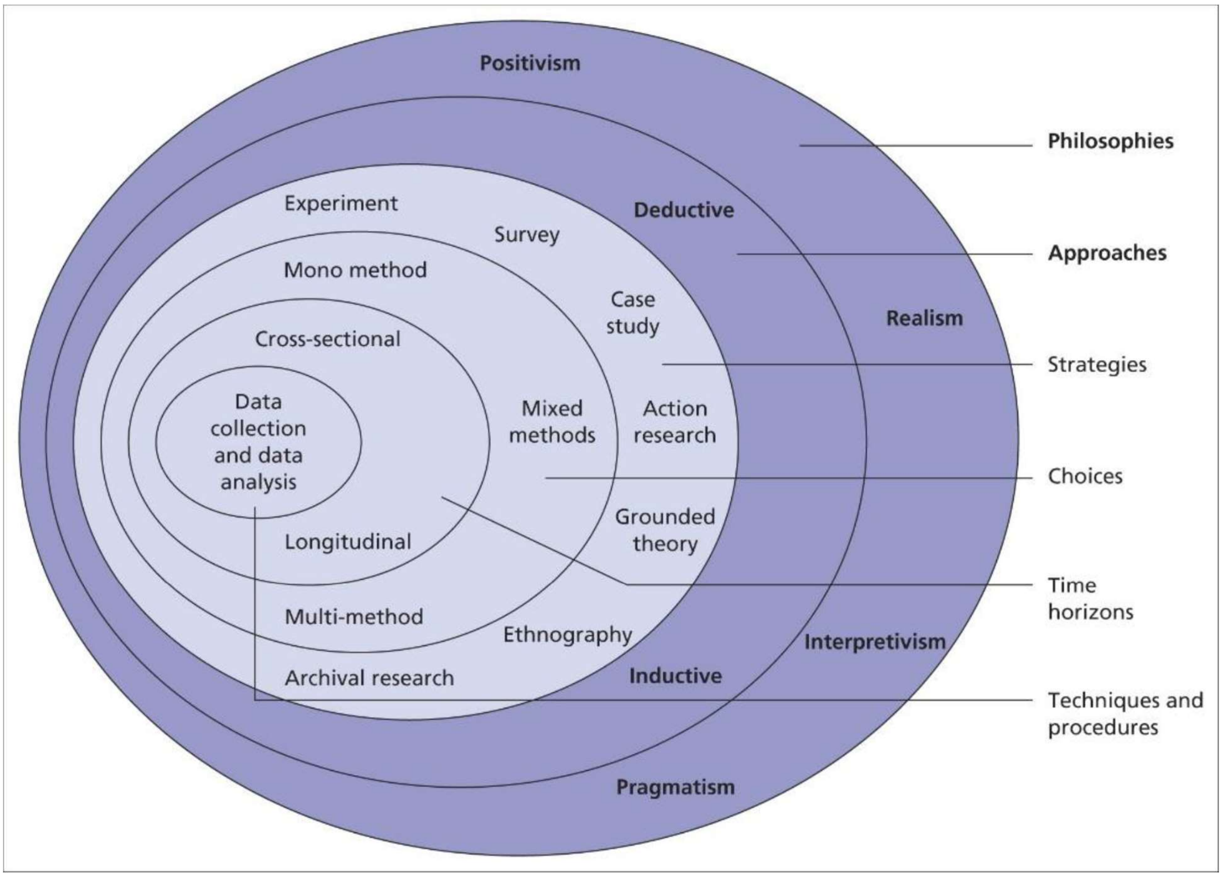


Figure 3.1.1: Research Onion

3.1.1 Research Philosophies

The research philosophy is the outermost layer of the onion. If we talk about what this research philosophy actually is: it is the foundation of any research, which outlines the set of assumptions the investigation is based on. One might characterize research philosophy from an ontological or an epistemological perspective. It addresses the “what” and “how” of what we know.

Ontology is, to put it simply, the "what" of what we know; that is, what is the nature of reality, and how much of it we actually know and can comprehend. Conversely, epistemology deals with "how" we may learn and comprehend the reality; that is, how we can determine what reality is and the boundaries of our knowledge. Although this is a severe oversimplification, it serves as a helpful beginning point.

Positivism

Research from a positivist's perspective holds that knowledge is not limited to the subject under study. To put it another way, the study can only be conducted objectively; opinions or personal ideas cannot be included. The researcher only observes; they do not interpret. According to positivism, knowledge can only be obtained by empirical inquiry that uses measurement and observation as its foundation. Put otherwise, all information is considered to be posteriori knowledge, or knowledge that comes from study of a subject rather than from human thinking and interpretation.

Interpretivism

Conversely, interpretivism places greater focus on the impact that social and cultural elements may have on a subject. This perspective is centered on people's ideas and thinking in relation to the sociocultural context. According to the interpretivist theory, the researcher must take an active part in the study in order to get a comprehensive understanding of the subject and its function, behaviors, and purpose.

Pragmatism

Pragmatism emphasizes how crucial it is to examine phenomena with the best instruments available. Pragmatism's primary goal is to approach research from a practical perspective, where information is continuously questioned and interpreted rather than fixed. Because of this, pragmatism involves subjectivity and the researcher, especially when it comes to making judgments based on the choices and answers of participants. Put otherwise, pragmatism is not restricted or dedicated to any one particular ideology.

Chosen Research Philosophy with Justification

The collection and analysis of the data have to be done in a quantitative matter, independent of opinions of any kind. So, **positivism** aligns with the objective measurement and analysis required to assess the impact of Big Data on AI in a quantitative manner, since it emphasizes using quantitative methods, ensuring rigorous statistical evaluation to draw meaningful conclusions from information. Although the fourth objective of the study doesn't seem to be aligned with positivism, since positivism encourages research is replicable, using empirical research to explore the subject of how ethical and privacy concerns affect the public trust on using Big Data on AI, we can apply positivism in there also. Even when taking the assumptions on the second objective, for evaluating the effect of challenges in using Big Data on AI, empirical research helped in finding the general areas of which challenges occur. Especially, when creating the questionnaire required, the existing research will be used to take a comprehensive understanding on how to align them with the objectives. This helps the research to be independent of the personal ideas of the researcher, and all information be posterior knowledge.

Overall, by adopting a positivist research philosophy, the study can raise awareness on quantifiable and measurable factors of the impact, supplying a scientific and goal analysis of the way Big Data affects the introduction and alertness of AI models.

3.1.2 Research Approach

The research approach refers to the larger technique one will use to conduct research, whether it is inductive or deductive. Inductive techniques include developing hypotheses via research rather than beginning a project with a theory as a foundation. Deductive techniques, on the other hand, begin with a hypothesis and seek to expand upon (or test) it through study.

Chosen Research Approach with Justification

For reading the impact of Big Data on AI within the context of a quantitative research methodology and positivist philosophy, the advocated studies approach is survey research based on deductive techniques. The study will conduct a survey to test whether the predetermined hypothesis (mentioned under the chapter Introduction: Hypothesis) are correct or not through quantitative analysis of collected data.

Data Collection: Surveys permit the gathering of quantitative information from a huge pattern of respondents. This is mainly beneficial whilst studying opinions, theories, or practices related to the effect of Big Data on AI.

Quantitative Analysis: Survey responses are inherently numerical, allowing for statistical evaluation. This aligns properly with the positivist philosophy and the need for goal, measurable information.

Generalizability: Surveys give the opportunity to generalize findings to a broader populace, presenting insights into how various stakeholders understand and revel in the effect. It also makes it easier as an undergraduate to approach potential individuals and companies. Survey also enables a time-green way to collect statistics from a huge wide variety of participants, taking into account a complete have a look at within an affordable time frame.

3.1.3 Research Strategies

This layer of the research onion describes the various ways that research can be carried out according to the study's objectives. Keep in mind that these tactics are known as study designs outside of the onion. There are several research strategies.

Experimental research: which entails modifying one variable (the independent variable) and seeing how this affects another one (the dependent variable), carried out in a controlled setting.

Action research: which is carried out outside of controlled contexts like labs, in real-world situations like classrooms, hospitals, workplaces. And the researcher can learn about issues or shortcomings pertaining to interactions in the actual world by doing action research.

Case study research: which is a comprehensive, in-depth analysis of a particular subject, where the topic is examined in-depth to provide a comprehensive grasp of problems in a practical context. This kind of research is typically qualitative and inductive in character.

Cross-Sectional Study: A move-sectional have a look at includes gathering information from individuals at a certain point in time. This strategy allows for the inclusion of a various range of participants, representing different industries, roles, and reports, presenting a complete snapshot of the current state of perceptions and practices concerning the subject of matter.

Grounded theory: where the researchers use the found data to inform the development of a new theory, model, or framework. Thus, grounded theory is particularly helpful for research into issues that have not been thoroughly studied or are entirely new.

Ethnography: the study of individuals in their natural settings with the goal of interpreting cultural relations. Seeing the world through the participants' subjective eyes and capturing their experiences is the aim of ethnography.

Archival research: An analysis of the existing data establishes meaning when an archive research technique makes use of pre-existing resources. This approach works especially effectively for historical research since it may utilize resources like records and manuscripts.

Chosen Research Strategy with Justification

For analyzing the effect of Big Data on AI with a focus on survey research inside a quantitative and positivist framework, the advocated studies strategy is a Cross-Sectional Study. Given the fast evolution of technology, this is suitable for researching the current situation of perceptions and practices concerning the effect of Big Data on AI. This strategy is also time-efficient and price-effective, making it appropriate for acquiring a wide review of the subject inside the constraints of the studies assignment. Using cross-sectional research, the study can efficaciously seize a wide range of perspectives, imparting a precious picture of how exceptional stakeholders understand and experience the effect of Big Data on AI in the modern context. This aligns with the quantitative, positivist method and the survey research method formerly recommended.

3.1.4 Research Choices

This layer is about selecting the number of qualitative or quantitative data categories you will utilize in your study. Three alternatives are available: mixed, multi-method, and mono. As the names suggest, If you decide to employ a mono approach, you will only use one kind of data, either quantitative or qualitative. Using a mixed-methods strategy would entail using both quantitative and qualitative data. This may be accomplished by using a survey to gather quantitative data, which could then be statistically analyzed to produce quantitative results in addition to your qualitative ones. When using multi method, the researcher would employ a greater variety of approaches, rather than merely one quantitative and one qualitative approach. For instance, you could utilize two qualitative techniques (such content and theme analysis) in a study looking at archives from a particular subject, and you could also employ quantitative techniques to analyze numerical data.

Chosen Research Choice with Justification

For discovering the effect of Big Data on AI through a move-sectional survey with a quantitative and positivist method, an appropriate studies preference is to pay attention on Key Stakeholder Surveys. Targeting surveys at key stakeholders in the AI industry, consisting of facts scientists, developers, software engineers, enterprise leaders, and policymakers ensures that diverse views from individuals influencing or impacted by way of the usage of Big Data in AI are taken into consideration. By surveying a whole lot of stakeholders, the studies can collect complete insights of the subject at hand. This choice enables us to design survey inquiries to cope with the worries, challenges, and possibilities aligned with the subject. Focusing on key stakeholders guarantees that the survey research captures nuanced insights from those with direct involvement and impact inside the subject, aligning with the objectives of the cross-sectional observe and the wider research framework.

3.1.5 Time Horizon

The number of points in time at which you intend to collect your data is simply described by the time horizon. There are two possibilities: the longitudinal and cross-sectional time horizons.

A longitudinal time horizon might be used instead if you wanted to examine the evolution or growth of a subject before and after an event or a particular time period. This could be a quantitative, qualitative, or a combination of the two, but the data will be collected on at least two different time frames on the same sample or population.

A cross-sectional time horizon, on the other hand, would focus on a certain point of time and collect data to analyze the behavior of a subject at that point in time.

Chosen Research Time Horizon with Justification

For studying the impact of Big Data on AI thru key stakeholder surveys and using a move-sectional layout inside a quantitative and positivist technique, a suitable research time horizon is Short-Term, which is a cross-sectional time horizon.

Given the hastily evolving nature of era and the dynamic panorama of the AI industry, a short-term time horizon lets in the research to seize the instantaneous impact of Big Data on AI practices and perceptions without being overshadowed by means of prolonged adjustments. Short-time period studies are well-ideal for supplying timely insights into the cutting-edge situation, making sure that the studies remain relevant and reflective of the most current tendencies inside the subject. By adopting a short-term time horizon, the study can correctly capture the cutting-edge state of affairs and offer actionable suggestions based totally on the on-the-spot effect of Big Data on AI practices as perceived by means of key stakeholders.

3.1.6 Techniques and Procedures

The innermost layer of the research onion, where one really need to apply the research to make decisions on methods and approaches. Here, the researcher will select sampling strategies, choose what kind of analysis will be done, arrange resources, and make preparations for data collection. It is crucial to remember that these methods and approaches must be in line with every other layer of the research onion, including the choices, time horizon, research philosophy, research approaches, and research strategy.

Chosen Research Techniques and Procedures with Justification

For engaging in a brief-term pass-sectional survey on the effect of Big Data on AI thru key stakeholder perspectives, an appropriate studies method is Online Questionnaires. It provides a convenient approach for contributors, making the collectable data larger. This is also time effective and allows standardized responses, making sure consistency in survey management and reaction recording.

Procedure for the Online Questionnaire:

The objectives of the research (mentioned above) were clearly studied and searched for similar empirical research for guidance. After that well established questions were tailored to address each objective and to help analyze the respective hypotheses for the objectives. Then it was shared among a few individuals in the Big Data and AI industry for a pilot test and find and deal with any issues associated with the survey shape, wording, or drift.

The purpose of the questionnaire was clearly mentioned as well as declared participant and response confidentiality. It was declared that the data will only be used for the research purpose and will not be shared with any other parties for any purpose. Then the questionnaire was shared with the respondents in the related field using convenience sampling and snowball technique. The target population was the employees in the Big Data and AI industry in Sri Lanka.

The data was collected through google forms from 60 respondents and was stored in google drive. Then a copy (backup) was downloaded into the personal computer as a CSV file and analyzed using the statistical software SPSS. Some data had to be transformed into numerical using the recoding function available in SPSS before analyzing. A descriptive analysis was done to get an overview of the collected data and clean the dataset. After that a correlation analysis was conducted to find the relationship between the data. Then a regression analysis was done to check the relationship between the dependent variable and the independent variables. Four regression methods were acquired using the stepwise regression method, and checking partial F-values the best model was selected. The results will be visualized graphically, and the outcome will be interpreted in the chapters below.

By utilizing online questionnaires, the research can correctly gather quantitative information from key stakeholders, considering a well-timed and comprehensive understanding in their views on the effect of Big Data on AI inside the brief-time period research time horizon.

3.2 Ethical issues of the research study

The data was collected anonymously and no information that would lead to identifying the respondents was collected. Most of the questions in the questionnaire were selected from empirical studies, reducing the ethical concerns of the questionnaire. The data will not be provided to any parties for any purpose and will only be used to the study of the research.

4. Results and Discussion

4.1 Descriptive analysis

The background of the respondents was as follows

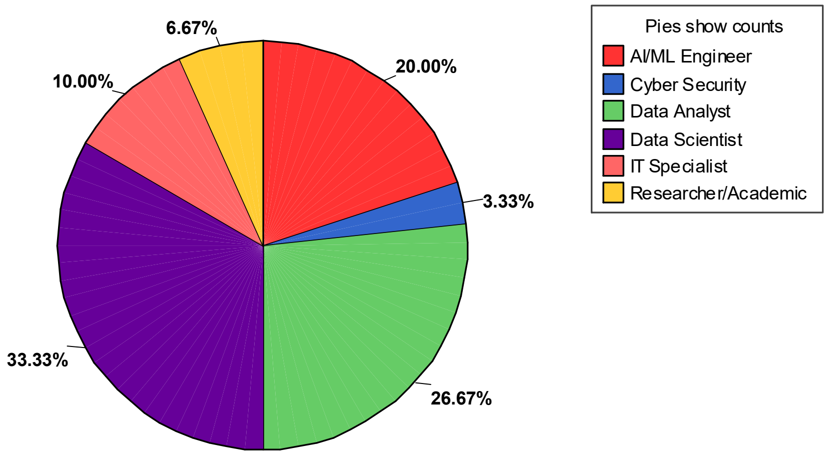


Figure 4.1 : The job roles of the respondents

This shows from which roles has the sample (respondents) been taken. As a convenience and snowball technique was used, all the participants are from the technological industry.

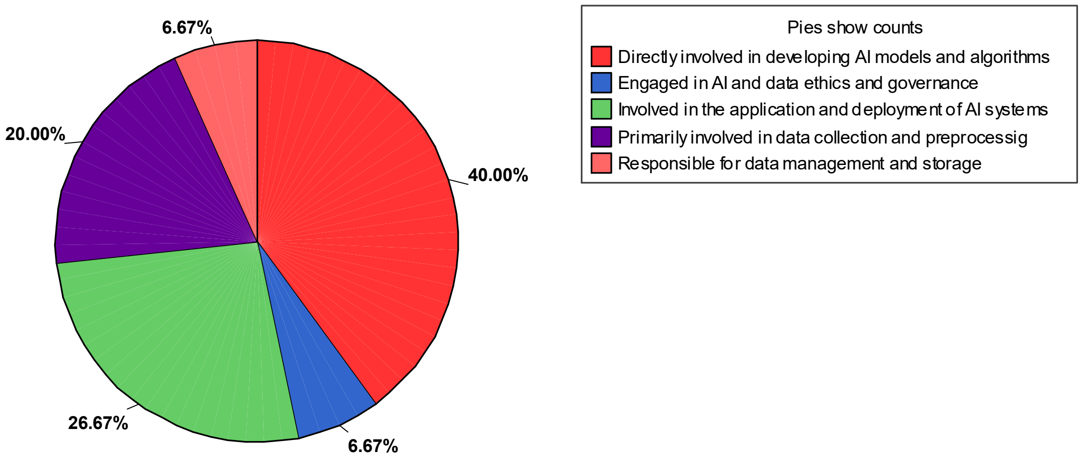


Figure 4.2 : Relationship of the respondents to the Big Data and AI industry

As Figure 4.2 shows, most of the respondents are directly involved in the AI or data industry.

To get a brief understanding of the subject of the respondent's ideals, which comes from experts in the industry, some questions related to the objectives of the study were addressed.

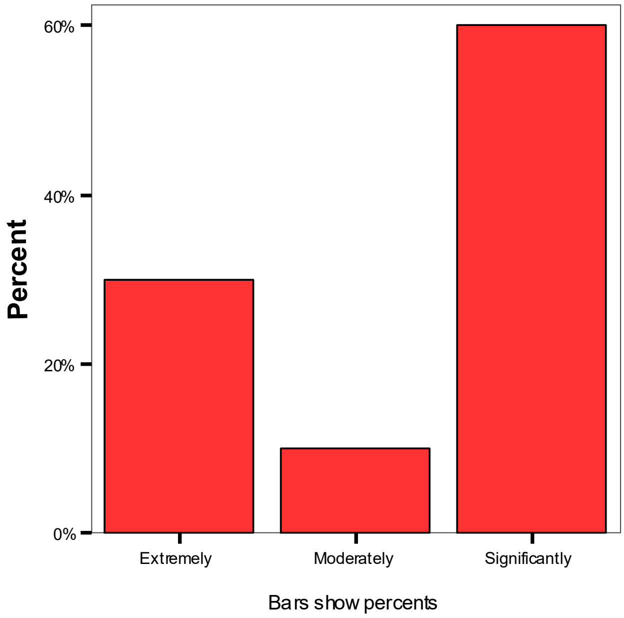


Figure 4.3 : The extent of the impact of Big Data on AI model building and training

The options to select "Not at all" or "Slightly" were given, but no respondent selecting that, and from the answers given we can assume that all the experts find there's a considerable impact on AI industry from Big Data.

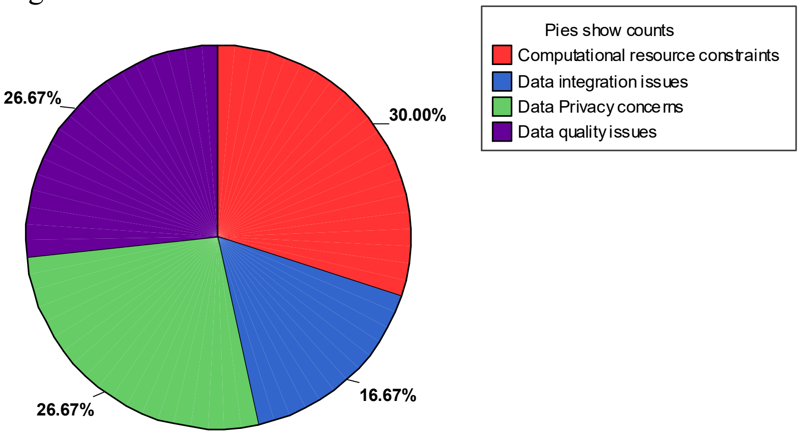


Figure 4.4 : Challenges in using Big Data on AI

As figure 4.4 indicates, the main challenges in using Big Data on AI industry are computational resource constraints, data quality issues, data privacy concerns, and data integration issues. This addresses the second objective of the study.

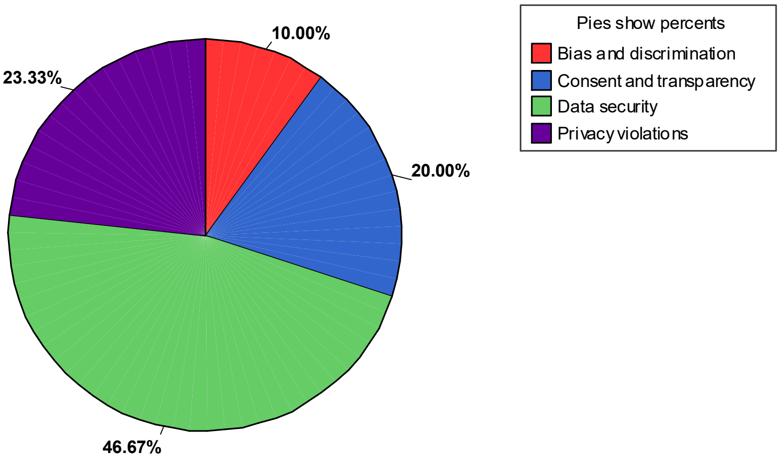


Figure 4.5 : The ethical issues of using Big Data on AI industry

Delving into the ethical issues of the subject, it was found that data security, privacy concerns, consent and transparency, and bias and discrimination were the major ethical issues regarding the use of Big Data on AI industries.

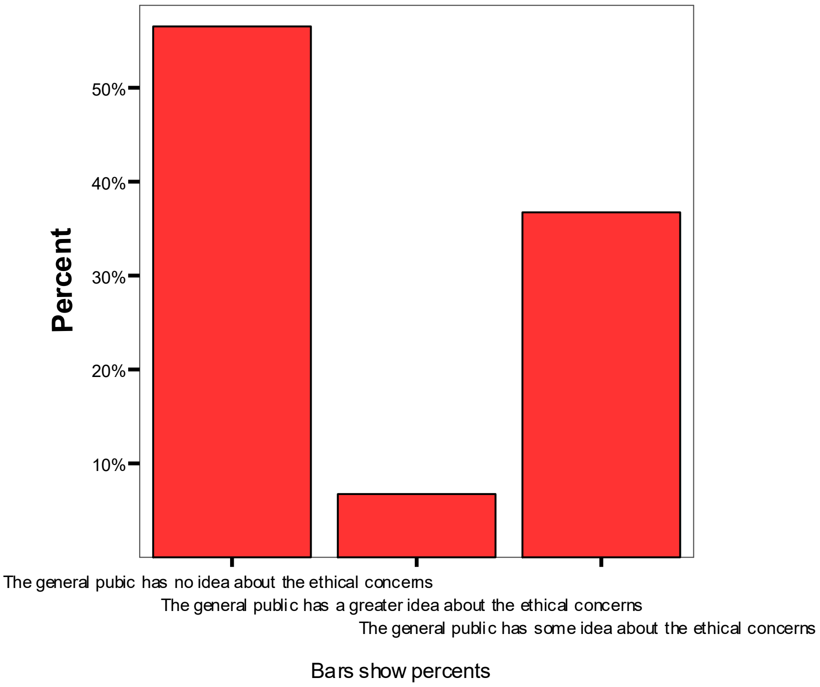
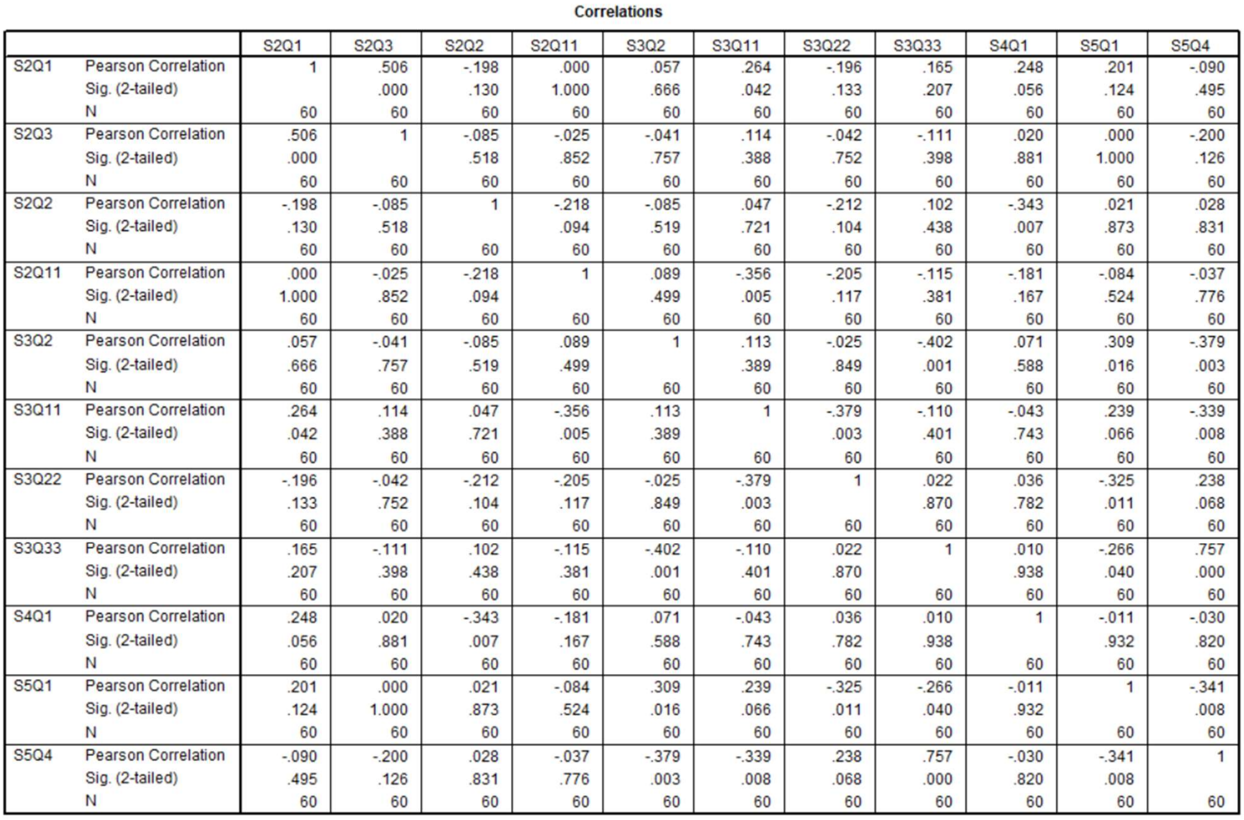


Figure 4.6 : The extend of the understanding of ethical concerns of the general public

Although there were many ethical issues regarding the use of Big Data on AI industries, the experts pointed out that the general public is mostly unaware of these concerns. The fourth objective is adressed here.

With this overview of the subject at hand, a correlation analysis between the used variables were conducted.



Correlations

		S2Q1	S2Q3	S2Q2	S2Q11	S3Q2	S3Q11	S3Q22	S3Q33	S4Q1	S5Q1	S5Q4
S2Q1	Pearson Correlation	1	.506	-.198	.000	.057	.264	-.196	.165	.248	.201	-.090
	Sig. (2-tailed)		.000	.130	1.000	.666	.042	.133	.207	.056	.124	.495
	N	60	60	60	60	60	60	60	60	60	60	60
S2Q3	Pearson Correlation	.506	1	-.085	-.025	-.041	.114	-.042	-.111	.020	.000	-.200
	Sig. (2-tailed)	.000		.518	.852	.757	.388	.752	.398	.881	1.000	.126
	N	60	60	60	60	60	60	60	60	60	60	60
S2Q2	Pearson Correlation	-.198	-.085	1	-.218	-.085	.047	-.212	.102	-.343	.021	.028
	Sig. (2-tailed)	.130	.518		.094	.519	.721	.104	.438	.007	.873	.831
	N	60	60	60	60	60	60	60	60	60	60	60
S2Q11	Pearson Correlation	.000	-.025	-.218	1	.089	-.356	-.205	-.115	-.181	-.084	-.037
	Sig. (2-tailed)	1.000	.852	.094		.499	.005	.117	.381	.167	.524	.776
	N	60	60	60	60	60	60	60	60	60	60	60
S3Q2	Pearson Correlation	.057	-.041	-.085	.089	1	.113	-.025	-.402	.071	.309	-.379
	Sig. (2-tailed)	.666	.757	.519	.499		.389	.849	.001	.588	.016	.003
	N	60	60	60	60	60	60	60	60	60	60	60
S3Q11	Pearson Correlation	.264	.114	.047	-.356	.113	1	-.379	-.110	-.043	.239	-.339
	Sig. (2-tailed)	.042	.388	.721	.005	.389		.003	.401	.743	.066	.008
	N	60	60	60	60	60	60	60	60	60	60	60
S3Q22	Pearson Correlation	-.196	-.042	-.212	-.205	-.025	-.379	1	.022	.036	-.325	.238
	Sig. (2-tailed)	.133	.752	.104	.117	.849	.003		.870	.782	.011	.068
	N	60	60	60	60	60	60	60	60	60	60	60
S3Q33	Pearson Correlation	.165	-.111	.102	-.115	-.402	-.110	.022	1	.010	-.266	.757
	Sig. (2-tailed)	.207	.398	.438	.381	.001	.401	.870		.938	.040	.000
	N	60	60	60	60	60	60	60	60	60	60	60
S4Q1	Pearson Correlation	.248	.020	-.343	-.181	.071	-.043	.036	.010	1	-.011	-.030
	Sig. (2-tailed)	.056	.881	.007	.167	.588	.743	.782	.938		.932	.820
	N	60	60	60	60	60	60	60	60	60	60	60
S5Q1	Pearson Correlation	.201	.000	.021	-.084	.309	.239	-.325	-.266	-.011	1	-.341
	Sig. (2-tailed)	.124	1.000	.873	.524	.016	.066	.011	.040	.932		.008
	N	60	60	60	60	60	60	60	60	60	60	60
S5Q4	Pearson Correlation	-.090	-.200	.028	-.037	-.379	-.339	.238	.757	-.030	-.341	1
	Sig. (2-tailed)	.495	.126	.831	.776	.003	.008	.068	.000	.820	.008	
	N	60	60	60	60	60	60	60	60	60	60	60

Figure 4.7 : Correlation Analysis of the Used Variables

The S*Q* numbers represents the respective variables taken from the questions.

- S2Q1: The advacement of AI modeling and training using Big Data.
- S2Q3: The effect of the variety of Big Data used, on AI industries.
- S2Q2: The effect of the volume of Big Data used, on AI industries.
- S2Q11:The effect of the usage of Big Data on AI models accuracy.
- S3Q2: How often one face challenges regarding data privacy and security.
- S3Q11:The effect of the quality of Big Data used, on AI industries.
- S3Q22:How the computational resource constraints affect the scalability and efficiency of AI models.
- S3Q33:The effect of privacy concerns on the scalability of AI industry.
- S4Q1: The effect of Big Data on the complexity and personalification of AI models.

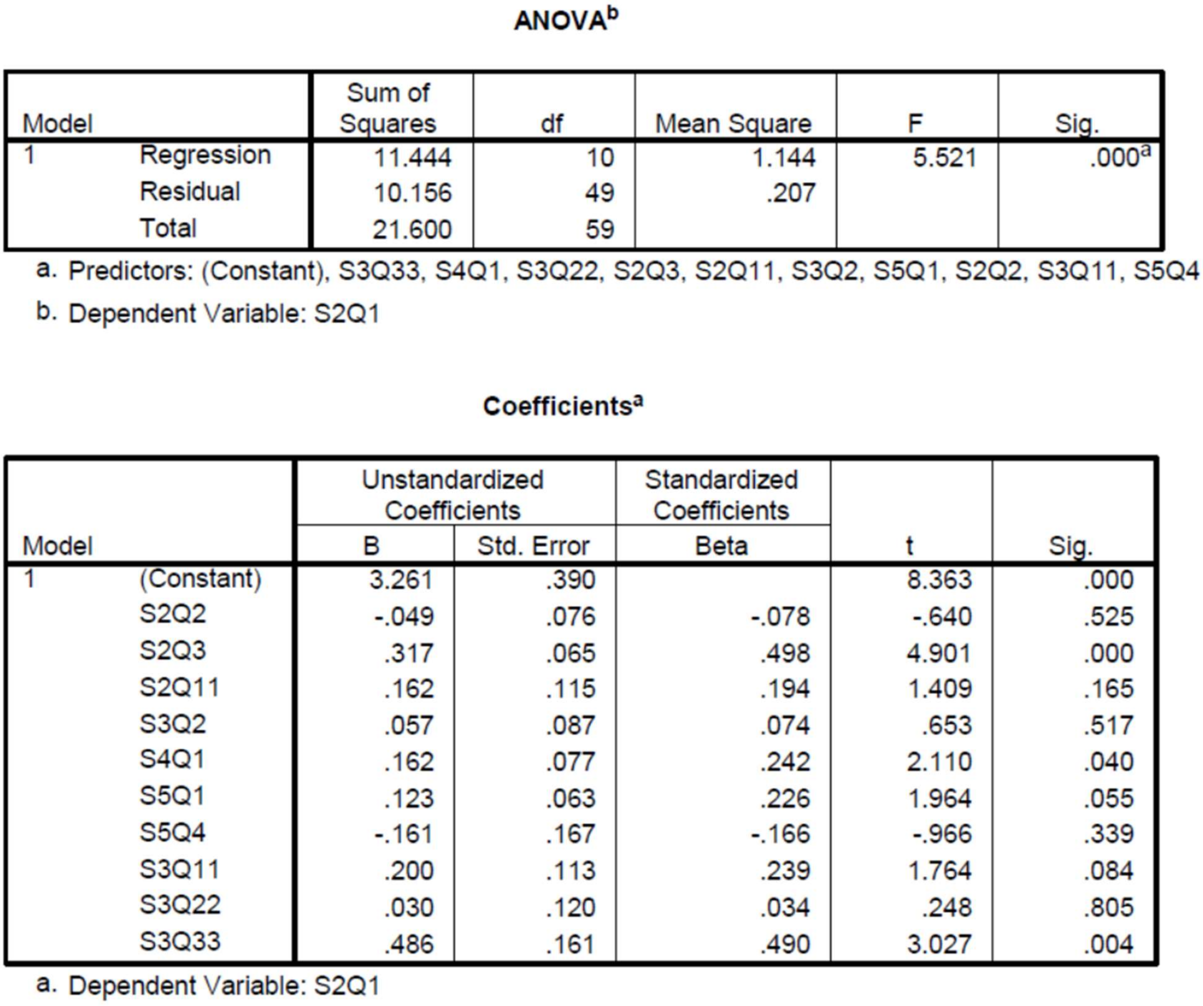
- S5Q1: The effect of ethical concerns regarding the use of Big Data on AI.
- S5Q4: How the understanding of the general public on the ethical concerns affect the AI industry.

The relationship between each variable seems to differ on a large scale over the variables.

4.2 Regression Analysis

A stepwise regression analysis was done using the SPSS software to analyze the hypothesis.

Initial model:



ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11.444	10	1.144	5.521	.000[a]
	Residual	10.156	49	.207		
	Total	21.600	59			

a. Predictors: (Constant), S3Q33, S4Q1, S3Q22, S2Q3, S2Q11, S3Q2, S5Q1, S2Q2, S3Q11, S5Q4

b. Dependent Variable: S2Q1

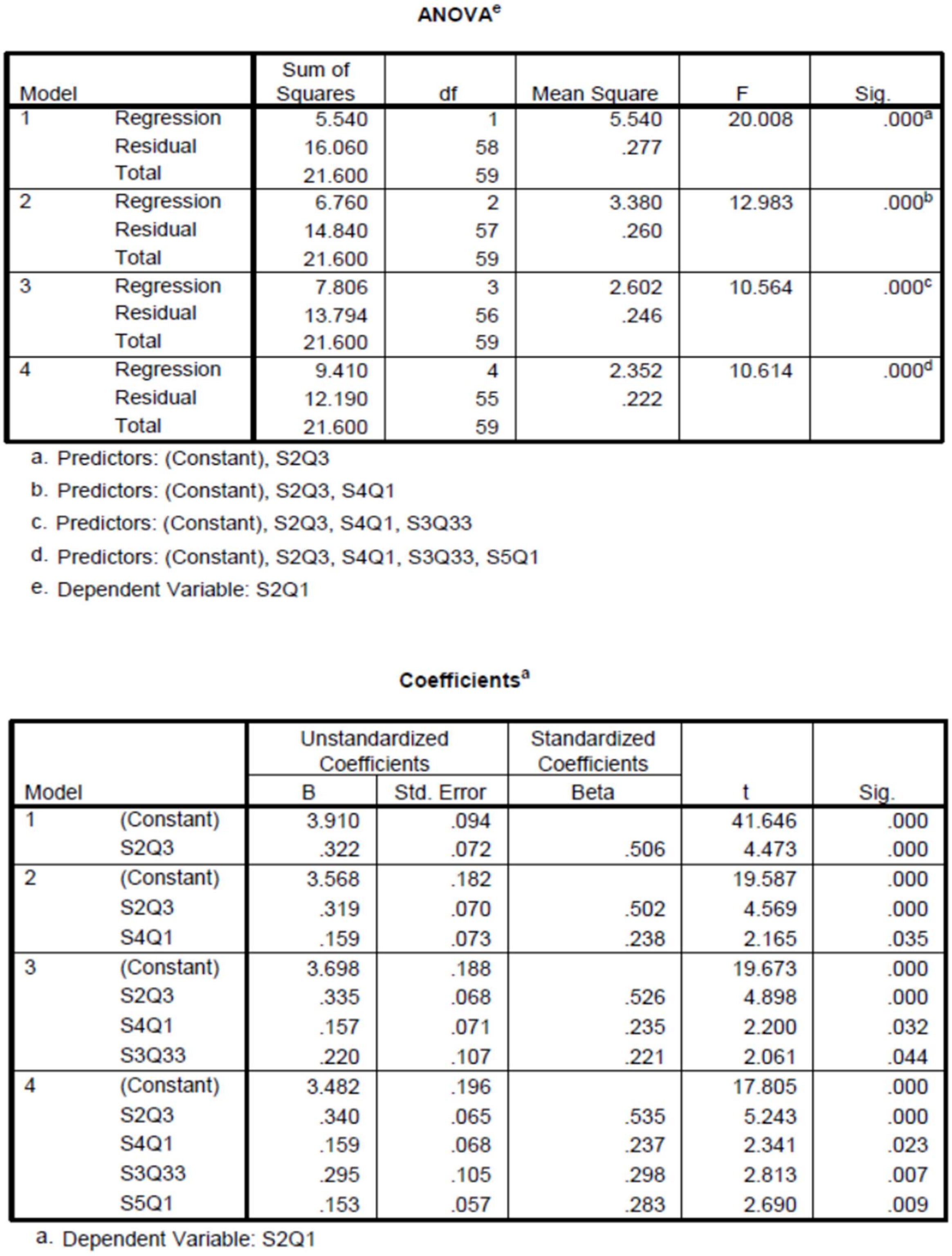
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.261	.390		8.363	.000
	S2Q2	-.049	.076	-.078	-.640	.525
	S2Q3	.317	.065	.498	4.901	.000
	S2Q11	.162	.115	.194	1.409	.165
	S3Q2	.057	.087	.074	.653	.517
	S4Q1	.162	.077	.242	2.110	.040
	S5Q1	.123	.063	.226	1.964	.055
	S5Q4	-.161	.167	-.166	-.966	.339
	S3Q11	.200	.113	.239	1.764	.084
	S3Q22	.030	.120	.034	.248	.805
	S3Q33	.486	.161	.490	3.027	.004

a. Dependent Variable: S2Q1

Figure 4.8 : Regression analysis output of initial model

Reduced model:



ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.540	1	5.540	20.008	.000[a]
	Residual	16.060	58	.277		
	Total	21.600	59			
2	Regression	6.760	2	3.380	12.983	.000[b]
	Residual	14.840	57	.260		
	Total	21.600	59			
3	Regression	7.806	3	2.602	10.564	.000[c]
	Residual	13.794	56	.246		
	Total	21.600	59			
4	Regression	9.410	4	2.352	10.614	.000[d]
	Residual	12.190	55	.222		
	Total	21.600	59			

a. Predictors: (Constant), S2Q3

b. Predictors: (Constant), S2Q3, S4Q1

c. Predictors: (Constant), S2Q3, S4Q1, S3Q33

d. Predictors: (Constant), S2Q3, S4Q1, S3Q33, S5Q1

e. Dependent Variable: S2Q1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.910	.094		41.646	.000
	S2Q3	.322	.072	.506	4.473	.000
2	(Constant)	3.568	.182		19.587	.000
	S2Q3	.319	.070	.502	4.569	.000
	S4Q1	.159	.073	.238	2.165	.035
3	(Constant)	3.698	.188		19.673	.000
	S2Q3	.335	.068	.526	4.898	.000
	S4Q1	.157	.071	.235	2.200	.032
	S3Q33	.220	.107	.221	2.061	.044
4	(Constant)	3.482	.196		17.805	.000
	S2Q3	.340	.065	.535	5.243	.000
	S4Q1	.159	.068	.237	2.341	.023
	S3Q33	.295	.105	.298	2.813	.007
	S5Q1	.153	.057	.283	2.690	.009

a. Dependent Variable: S2Q1

Figure 4.9 : Regression analysis output of reduced model

To check which model is the most suitable, a partial F-value test was conducted.

The following formula is used to prove the test statistic

$$F_{Partial} = \frac{(SS_{Regression}^{Full} - SS_{Regression}^{Reduced})/(k-P)}{SS_{Residual}^{Full}/[n-(k+1)]}$$

Then the F table value of the model will be calculated using

$$F_{Table} = F_{0.05,DF1,DF2}$$

And the Partial F value and table value will be compared to select suitable models.

Model 1:

H_0: The model is adequete

H_1: The model is inadequate

$$F_{Partial} = \frac{(11.444 - 5.54)/(10-1)}{10.156/[60-(10+1)]} = 3.1650$$

$$F_{Table} = F_{0.05,49,58} = 1.56$$

$F_{Partial} > F_{Table}$

Thus, we reject the null hypothesis. Therefore: the model 1 is inadequate.

Model 2:

H_0: The model is adequete

H_1: The model is inadequate

$$F_{Partial} = \frac{(11.444 - 6.76)/(10-1)}{10.156/[60-(10+1)]} = 2.5110$$

$$F_{Table} = F_{0.05,49,57} = 1.57$$

$F_{Partial} > F_{Table}$

Thus, we reject the null hypothesis. Therefore: the model 2 is inadequate.

Model 3:

H_0: The model is adequete

H_1: The model is inadequate

$$F_{Partial} = \frac{(11.444 - 7.806)/(10-1)}{10.156/[60-(10+1)]} = 1.9502$$

$$F_{Table} = F_{0.05,49,56} = 1.57$$

$F_{Partial} > F_{Table}$

Thus, we reject the null hypothesis. Therefore: the model 3 is inadequate.

Model 4:

H_0: The model is adequete

H_1: The model is inadequate

$$F_{Partial} = \frac{(11.444 - 9.410)/(10 - 1)}{10.156/[60 - (10 + 1)]} = 1.0903$$

$$F_{Table} = F_{0.05,49,58} = 1.58$$

$F_{Partial} < F_{Table}$

Thus, we fail to reject the null hypothesis. Therefore: the model 4 is adequate.

Variables used in model 4:

- S2Q3: The effect of the variety of Big Data used, on AI industries.
- S3Q33:The effect of privacy concerns on the scalability of AI industry.
- S4Q1: The effect of Big Data on the complexity and personalification of AI models.
- S5Q1: The effect of ethical concerns regarding the use of Big Data on AI.

As of the result of model testing, we can say which independent variables has the most impact on the dependent variable. So the initial model which consists of all the variables will be reduced to the following reduced state and accepted.

Initial Hypothesis:

H_0: When applying Big Data into AI models, the volume, accuracy, variety, quality of Big Data, the ethical and privacy concerns regarding its application into AI, the understanding of those ethical and privacy concerns of the public, and the computational resource constraints regarding the application of Big Data affects significantly to the advancement of the AI industry.

Reduced Hypothesis:

H_{0new}: When applying Big Data into AI models, the variety of used data, the privacy concerns regarding its application into AI, the effect of Big Data on the complexity and personification of AI models, and the effect of ethical concerns regarding the usage of Big Data on AI affects significantly to the advancement of the AI industry.

From the four models acquired thorugh stepwise regression analysis, the fourth model, which was accepted consists some parts of all the four objectives of the study. Meanwhile those models which were rejected had missed atleast one objective.

5. Conclusions and Recommendations

5.1 Conclusions

It is obvious that the usage of Big Data on AI model training and Performance has made a significant positive impact on the AI industry, increasing the accuracy and robustness of the models. Also the usage of Big Data from different varieties has made the trained models applicable in most situationa and it has also enabled more personalized application of AI programs. For example, if a developer was to train an AI program to help a stroke patient to keep in check using Big Data of healthcare industry all over the world, it would enhace the decision-making capabilities and the efficiency of the model. This implicates both enhanced performance and applicability of AI models using Big Data. The results showed that all the independent variables align to support the dependent variable, ensuring that the research hypothesis was correct.

Considering the objective that couldn't be covered by the regression analysis, as discribed in the descriptive analysis, it was obvious that the main ethical and privacy concerns reagrding the subject were computational resource constraints, data quality issues, data privacy concerns, and data integration issues. Not having a buff server or a supercomputer puts a great restraint on training AI models using Petabytes of Big Data. Also, the quality of the data becomes a problem as it gets larger. The current Big Data storage system uses a network of super servers across continents to store data, which could raise data privacy concerns. Integrating Petabytes of data into a learning model proves to be a challenge too. Ethical considerations like privacy violations and data security seem to arise sometimes. But as mentioned before, the public is mostly unaware of most of these situations, making the impact of ethical concerns minimal.

5.2 Recommendations

To the improvement of the AI industry using Big Data, it is highly recommended to reduce the computational resource constraints and further increase the variety of Big Data. Using Big Data to create more complex algorithms is also recommended.

On the field of ethical and privacy concerns, it is recommended to be more transparent and inform the public about the ethical considerations regarding personal data usage. More closed server usage for data storage is recommended.

For future studies, it is important to either study deep into the integration of Big Data on AI models or delve deep into the subject of ethical and privacy concerns.

5.3 Limitations

The population and sample were limited to a small region of the world, there could be new and unknown data and information on the vast geographical horizons.

The chosen area was much vaster than expected and it was hard to delve deep into one part of the subject.

It was hard to do a proper analysis of some of the objectives without much more deep insights of the subject and had to settle on a surface level of conclusion based solely on the ideals of the experts on the field.

6. References

Iphofen, R. and O'Mathúna, D. eds., (2021). Ethical Issues in Covert, Security and Surveillance Research. Advances in Research Ethics and Integrity. Emerald Publishing Limited. doi:https://doi.org/10.1108/s2398-6018202108.

Bean, R. (2017). How Big Data Is Empowering AI and Machine Learning at Scale. [online] MIT Sloan Management Review. Available at: https://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale/.

Dilmaghani, S., Brust, M.R., Danoy, G., Cassagnes, N., Pecero, J. and Bouvry, P. (2019). Privacy and Security of Big Data in AI Systems: A Research and Standards Perspective. [online] IEEE Xplore. doi: https://doi.org/10.1109/BigData47090.2019.9006283.

Progress Blogs. (2023). Understanding Data Bias & Impact on AI/ML Decision Making. [online] Available at: https://www.progress.com/blogs/understanding-data-bias-impact-ai-ml-decision-making [Accessed 27 Jan. 2024].

ieeexplore.ieee.org. (n.d.). The impact of Big Data on AI | IEEE Conference Publication | IEEE Xplore. [online] Available at: https://ieeexplore.ieee.org/document/9458076 [Accessed 27 Jan. 2024].

https://gradcoach.com/saunders-research-onion/

Phair, D. and Warren, K. (2021). Saunders' Research Onion: Explained Simply. [online] Grad Coach. Available at: https://gradcoach.com/saunders-research-onion/.

GDPR (2019). Complete guide to GDPR compliance. [online] GDPR.eu. Available at: https://gdpr.eu/.

IEEE Standards Association. (n.d.). Autonomous and Intelligent Systems (AIS). [online] Available at: https://standards.ieee.org/initiatives/autonomous-intelligence-systems/.

Privacyinternational.org. (2019). Privacy International. [online] Available at: https://privacyinternational.org/.

Partnership on AI. (n.d.). Home. [online] Available at: https://partnershiponai.org/.

IEEE (n.d.). IEEE Xplore Digital Library. [online] Ieee.org. Available at: https://ieeexplore.ieee.org/Xplore/home.jsp.

Science Direct (2022). ScienceDirect.com | Science, Health and Medical journals, Full Text Articles and books. [online] Sciencedirect.com. Available at: https://www.sciencedirect.com/.

F.Oluyinka and K.Hubert, (2024) Regulatory Compliance and Ethical Considerations: Compliance challenges and opportunities with the integration of Big Data and AI. https://www.researchgate.net/publication/377330616

J.Atetedaye, (2024) Privacy-Preserving Machine Learning: Securing Data in AI Systems. https://www.researchgate.net/publication/380711820

7. Appendix