**EXPLORING THE ETHICAL IMPLICATIONS OF BIG DATA AND DATA MINING**

SHAILENDRA CHOURASIA PANKAJ KUMAR JAIN

Assistant Professor, CSE, GGCT Assistant Professor, CSE, GGCT

shailendrachourasia@ggits.org pankajjain@ggct.co.in

**ABSTRACT**

Big data analytics is a rapidly developing phenomena that is influenced by interactions between people, businesses, and society. The ethical ramifications for these stakeholders are still little understood and experimentally underexplored. The output of digital data is constantly increasing as a result of the advancement of new technologies, the Internet, and social networks. The heterogeneous mass of digital data produced by businesses and individuals is referred to as "Big Data" because of its properties (big volume, variety of forms, and speed of processing), which call for specialised and highly advanced computer storage and analysis tools. We outline the fundamental ideas behind the ethical concerns raised by big data analytics and data mining. Then, in order to support the ethical use of big data analytics, we investigate those notions using stakeholder theory and discourse ethics and make suggestions for how to strike a balance between interactions between people, companies, and society. This article aims to define the principles, issues, and applications of big data and data mining as well as the significance of big data and data mining analytics.

Key words: Big Data, Data Mining, Big data Analytic, Ethical Implication.

**INTRODUCTION**

To find patterns, correlations, and other insights from massive and complicated data sets, big data analytics use algorithms (Martin, 2015). Big data analytics may be used in many ways to increase economic and social value, including healthcare, public safety, and service innovations. However, it has recently been under fire for having unsavoury effects on a number of stakeholders. Concerns about privacy violations, substantial individual profiling, or customer discrimination have been made public (Zwitter, 2014).These issues highlight a contradiction between stakeholder values where organisations' objectives and incentives do not line up with those of people and society. As a result, we are unsure of how to fairly divide the costs and advantages of big data across different stakeholder groups.

We utilise a stakeholder perspective (Mitchell, Agle, & Wood, 1997) to examine the interrelationships among diverse stakeholders in order to address the value conflict inherent in big data analytics applications. For two reasons, we do this. First, they urged discourse ethics to be used by IS researchers to answer moral conundrums.According to discourse ethics, morality develops through fair discussions among parties involved in an ideal speech environment (i.e., where everyone is on an even playing field). Second, by emphasising stakeholders, we address Markus' (2015) call for IS researchers to investigate the effects of big data analytics for various stakeholders as well as researcher suggestions that understanding the effects of corporations' "non-responsible" use of big data on people and society represents a key research priority. Despite the fact that recent research has brought up ethical concerns with big data analytics, it has not taken a stakeholder viewpoint or examined the relationships among big data analytics stakeholders.

The industry with the fastest growth over the past four years has been information technology (IT), which is expected to expand between $340 million and $1 trillion over the next two years. Millions of individuals work as IT professionals in fields related to computers. However, the IT sector has a negative name for unethical and unprofessional behaviour. The area of IT that is the most specialised is big data and data mining. Globally, digital data is rapidly expanding, from 150 exabytes in 2005 to 1200 exabytes in 2010. In the upcoming years, data growth is projected to be 40%. From 2007 through 2020, the amount of digital data is projected to grow 44 times, doubling roughly every 20 months (Chowdhury, 2014). These industries are crucial for establishing professionalism and moral behaviour because of this. With the aforementioned issue in mind, I conducted research for the following question –

Why professionalism and ethics are important in Information Technology and its specific area of Big Data Analytics and Data Mining?

Individuals' personal information is used by big data and data mining analysts. Businesses gather customer data using a variety of technologies and use it for data analysis, particularly for prediction analysis. However, it's crucial to consider how they use customer data. The research question for this paper is in favour of the United Nations Economic Commission. It will contribute to further study on "Big Data for Development: Assessing the Fitness in Monitoring and Evaluation (M&E) System" for any work that raises the question, "How might Big Data help monitor more precisely and timely economic, social, and environmental phenomena?" I offered scholarly and expert perspectives on IT professionalism and ethical issues with an emphasis on the Big Data and Data Mining domain in order to take into account this genuine need.

**RESEARCH BACKGROUND**

Massive Data Analysis The three Vs of technology—volume, variety, and velocity—or the capability of its underlying algorithms to produce insights—have been used most frequently to conceive big data analytics up until this point. Our understanding of big data analytics as a socio-technical phenomenon that affects all stakeholders is, however, constrained by such a technology focus. We augment the technological view by identifying three social processes that aim at and have an impact on people based on an analysis of the developing literature: Data sourcing, data sharing, and algorithmic decision-making are the first two. First, many big data applications use people as a means of collecting data. The "catch-all-you-can" strategy is used by businesses and government agencies to get as much data as possible from people. This method measures people's daily activities, especially for the organisation performing the analytics. Second, until its value is exhausted, data collected from individuals is transferred from one entity to another. This logic has created a secondary market for organizations to sell or share customer data. Individuals cannot see the purpose behind data-driven services to extract and share customer data. Third, organizations use algorithms to profile individuals— sometimes inadvertently based on their race, ethnic group, gender, and social and economic status—and restrict their options and choices, which raises wider ethical questions about how markets operate in a fair and free manner and questions that pertain to freedom of choice for individuals. Or, to put it another way, how can people be free if they are being watched over and controlled by algorithms that want to have a say in what they do?

In this article, we emphasise the use of big data analytics by businesses in their interactions with consumers, especially when those businesses employ big data analytics to provide consumers with services and goods. We contend that when organisations gather, examine, distribute, and/or sell people's data without those people's informed or sincere consent, ethical problems result.

\_\_\_\_\_\_\_\_\_\_ yaha tak ho gaya hai

**Discourse Ethics**

Ethics is the study of how people ought to behave and what defines truthful conduct. There are several conventional methods of thinking about ethics (such as utilitarian, kantian, and aristotelian), yet each has drawbacks. The utilitarian approach, for instance, makes it difficult to predict the effects of one's actions in the modern world, and in human societies, choosing the common good of the majority may be discriminatory towards minority. Here, we examine the ethics of big data analytics using discourse ethics as a framework. introduced discourse ethics to the IS literature and argued in favour of its ability to address moral concerns that face IS practitioners and scholars. Discourse ethics, which has its roots in research, is a more contemporary theory that integrates older theories of ethics, particularly Kantian and utilitarian views. Discourse ethics encompasses the idea of universalism and mostly focuses on morality. According to universalism, moral standards apply to activities that are equally good for everyone, regardless of setting or community.

Discourse ethics is centred on the discourse process, or the process of communicative activity. According to Habermas, a deliberative process is the most effective way for stakeholders to arrive at pragmatist, ethical, and moral norms. Only those norms, he contends, can be said to be valid, in the eyes of everyone who is affected and who is a participant in a practical conversation.In this approach, ethical standards cannot already exist or be enforced; rather, they must be fairly debatable among relevant parties. The stakeholders' real discussions will shape the ethical discourse, which they should continually revise over time. This necessitates the existence of the ideal speech environment, which calls for equal participation from all parties and the freedom to speak, claim, and dispute their opinions.

We propose that ethical big data analytics will emerge from stakeholders' engagement with and creation of ethical discourse and use discourse ethics as our overarching ethical framework. However, the ability of stakeholders to develop such a discourse depends on their level of equality and happiness with the ideal speech environment. Stakeholder theory (Mitchell et al., 1997) is used to identify, categorise, and analyse the stakeholders and their inter-relationships because discourse ethics does not explain how to identify stakeholders and their salience.

Some IT professionals work with specific Big Data management frameworks. Big Data is a subset of the 3Vs, which stand for volume, velocity, and variety. It involves the examination of real-time data, particularly to find trends over time or historical analysis. explains that "Data Mining" is an interdisciplinary field that combines techniques from artificial intelligence, machine learning, statistics, and database systems. Furthermore, according to Anonymous (1998), data mining is the process of retrieving, summarising, and summarising data from various databases. Data originates from consumers or from persons in both scenarios. Laws governing privacy, copyright, patents, and trademarks can prevent unethical and unprofessional activity in the IT sector, particularly in the fields of big data and data mining.

**DESCRIBE BIG DATA AND DATA MINING**

**What is Big Data**

The term "Big Data" refers to the evolution and use of technologies that provide the right user at the right time with the right information from a mass of data that has been growing exponentially for a long time in our society. The challenge is not only to deal with rapidly increasing volumes of data but also the difficulty of managing increasingly heterogeneous formats as well as increasingly complex and interconnected data.

Being a complex polymorphic object, its definition varies according to the communities that are interested in it as a user or provider of services. Invented by the giants of the web, the Big Data presents itself as a solution designed to provide everyone a real-time access to giant databases. Big Data is a very difficult concept to define precisely, since the very notion of big in terms of volume of data varies from one area to another. It is not defined by a set of technologies; on the contrary, it defines a category of techniques and technologies. This is an emerging field, and as we seek to learn how to implement this new paradigm and harness the value, the definition is changing.

**Characteristics of Big Data**

The term Big Data refers to gigantic larger datasets (volume); more diversified, including structured, semi-structured, and unstructured (variety) data, and arriving faster (velocity) than

before. These are the 3V.

*** 3V:***

**Volume:** represents the amount of data generated, stored and operated within the system. The increase in volume is explained by the increase in the amount of data generated and stored, but also by the need to exploit it.

**Variety:** represents the multiplication of the types of data managed by an information system. This multiplication leads to a complexity of links and link types between these data. The variety also relates to the possible uses associated with a raw data.

**Velocity:** represents the frequency at which data is generated, captured, and shared. The data arrive by stream and must be analyzed in real time.

**5V:**

To this classical characterization, two other "V"s are important:

**Veracity:** level of quality, accuracy and uncertainty of data and data sources.

**Value:** the value and potential derived from data.

**What Is Big Data Analytics**

Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be applied to analyze and extract patterns from large-scale data. The analysis of structured data evolves due to the variety and velocity of the data manipulated. Therefore, it is no longer enough to analyze data and produce reports, the wide variety of data means that the systems in place must be capable of assisting in the analysis of data. The analysis consists of automatically determining, within a variety of rapidly changing data, the correlations between the data in order to help in the exploitation of it.

Big Data Analytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different from the usual ones, more complex, and of a large enormous scale. It mainly focuses on solving new problems or old problems in better and effective ways.

 **Types of Big Data Analytics**

**Descriptive Analytics :** It consists of asking the question: What is happening?

It is a preliminary stage of data processing that creates a set of historical data. Data mining methods organize data and help uncover patterns that offer insight. Descriptive analytics provides future probabilities and trends and gives an idea about what might happen in the future.

**Diagnostic Analytics :** It consists of asking the question: Why did it happen?

Diagnostic analytics looks for the root cause of a problem. It is used to determine why something happened. This type attempts to find and understand the causes of events and behaviors.

**Predictive Analytics :** It consists of asking the question: What is likely to happen?

 It uses past data in order to predict the future. It is all about forecasting. Predictive analytics uses many techniques like data mining and artificial intelligence to analyze current data and make scenarios of what might happen.

**Prescriptive Analytics :** It consists of asking the question: What should be done?

 It is dedicated to finding the right action to be taken. Descriptive analytics provides a historical data, and predictive analytics helps forecast what might happen. Prescriptive analytics uses these parameters to find the best solution.

Data mining refers to the activity of going through big data sets to look for relevant or appropriate information. The idea is that businesses collect massive sets of data that may be homogeneous or automatically collected. Decision makers need access to smaller, more specific pieces of data from those large sets. They use data mining to uncover the pieces of information that will inform leadership and help chart the course for a business.

Data mining can involve the use of different kinds of software packages such as analytics tools. It can be automated, where individual workers send specific queries for information to database. Generally, data mining refers to operations that involve relatively sophisticated search operations that return targeted and specific results. For example, a data mining tool may look through dozens of years of accounting information to find a specific column of expenses or accounts receivable for a specific working year . The importance of Big Data does not mean how much data we have but what would you get out of that data. We can analyze data to reduce cost and time, smart decision making.

**Data Mining And Big Data**

Data mining, also known as data discovery or knowledge discovery, is the process of analyzing data from different viewpoints and resulting it into useful information. This information is used by businesses to increase their revenue and reduce operational expenses. The software programs used in data mining are amongst the number of tools used in data analysis. The software enables users to analyze data from different point of views, classify it and make a summary of the data trends identified. Technically, data mining involves the process of discovering patterns or relationships in large areas of related databases. The actual data mining task is the automatic or semi-automatic analysis of large datasets. This is done to assist in the extraction of previously unknown and unusual data patterns. These include detecting abnormalities in records, cluster analysis of data files and sequential pattern mining. Database techniques like spatial indices are commonly used in these processes.

After these processes, the patterns can be seen as the summary of the input data and can be used in further analysis like predictive analytics or machine learning. For instance, multiple groups of data can be identified through data mining steps. This is the process of analyzing larger data sets with the aim of uncovering useful information. Examples of this information include market trends, customer preferences, hidden patterns and unknown correlations. The analytics findings usually lead to new revenue opportunities, improved operational efficiency, more efficient marketing and other business benefits. Companies often rely on big data analytics to assist them in making strategic business decisions.

Big data analytics enable data scientists, predictive modellers and other professionals in the analytics field to analyze large volumes of transaction data. They can also use big data analytics to analyze data which might not have been discovered by conventional business programs.

**Challenges To Handle Big Data**

The programmers have to take decisions due to large availability of raw and complex data. An organization can collect, store, and analyze these large datasets in a number of ways. The Business can even use robust big data tools to store, access, and manage the structured and unstructured data collected from various sources in a faster and more efficient way. There are few challenges to address when handling big chunks of data. Some challenges listed below:

**Handling a Large Amount of Data**

The large availability of data makes the difficulty is making decisions. Data that enterprises can access has been increased exponentially from last several years. They have data for everything, right from what a consumer likes, to how they react, to a particular scent, to the amazing restaurant that opened up in Italy last weekend. This data exceeds the amount of data that can be stored and computed, as well as retrieved. The challenge is not so much the availability, but the management of this data . Along with rise in unstructured data, the availability of data is in multiple formats such as video, audio, social media, smart device data etc. Some of the newest ways developed to manage this data are a hybrid of relational databases combined with NoSQL databases. An example of this is MongoDB, which is an inherent part of the MEAN stack. There are also distributed computing systems like Hadoop to help manage Big Data volumes.

**Data Security**

In increasing of data, the major issue is to secure the data. Many organizations claim that they face trouble with Data Security. This happens to be a bigger challenge for them than many other data-related problems. The data that comes into enterprises is made available from a wide range of sources, some of which cannot be trusted to be secure and compliant within organizational standards. They need to use a variety of data collection strategies to keep up with data needs. This in turn leads to inconsistencies in the data, and then the outcomes of the analysis . This data is made available from numerous sources, and therefore has potential security problems. We may never know which channel of data is compromised, thus compromising the security of the data available in the organization, and giving hackers a chance to move in. Now it is essential to introduce Data Security best practices for secure data collection, storage and retrieval.

**Data Complexity**

Organizations must be aware of how to handle the massive amounts of data that are updated every second. Real-time data from customers' recent transactions, for instance, can be useful for a retail corporation looking to monitor client behaviour. Veracity and Velocity are two data analysis tools that can be used for the same. They include frameworks, compute engines, visualisation engines, ETL engines, and other required inputs. Along with the data that is static and constantly available, it is crucial for organisations to stay current with this data. Better insights will be developed as a result, and decision-making skills will be improved.

**Shortage of Skilled Resources**

Right now, there is a dearth of qualified Big Data specialists. Many businesses have started to bring up this in their efforts to improve their use of Big Data and create more efficient Data Analysis platforms. The current shortage of qualified Data Scientists and Data Analysts makes "number crunching" challenging and the development of insights cumbersome. Again, for a business dealing with new technologies, training employees at the entry level can be costly. Instead, many are focusing on automation strategies that use AI and ML to produce insights, but this too requires highly qualified personnel or the outsourcing of talented engineers.

**CONCLUSION**

Big data analytics is a sophisticated social phenomena that has a built-in dualism. It undoubtedly presents chances for human society to grow, but it also presents ethical difficulties for the parties involved. The sector with the fastest growth is information technology. Many individuals work in information technology (IT). Everywhere we go, we utilise IT to make life easier, but occasionally, it also causes problems. There are many instances of IT professionals misusing technology. Stopping young practitioners from acting unethically and reducing negative reputation are difficult tasks. Information about individual users is collected by the government, social media, and even organisations. Of course, how they use this information is a matter. There are occasions when IT infrastructures are built without considering the needs of the entire society. Big Data is used unethically by the government or political parties to manipulate or control the populace. Practitioners of big data analytics and data mining analyse data to improve society and daily life. This industry is crucial for reducing crime and antisocial behaviour, and it may also help fight disease.

**REFERENCES**

[1] Sorell, Tom, Nasir Rajpoot, and Clare Verrill. "Ethical issues in computational pathology." Journal of Medical Ethics 48.4 (2022): 278-284.

[2] Stahl, Bernd Carsten. "Responsible innovation ecosystems: Ethical implications of the application of the ecosystem concept to artificial intelligence." International Journal of Information Management 62 (2022): 102441.

[3] Ida Someh, Michael Davern, Christoph F. Breidbach, Graeme Shanks (2019), Ethical Issues in Big Data Analytics: A Stakeholder Perspective, Communications of the Association for Information Systems, volume 44,article 34 ISSN: 1529-3181.

[4] Youssra Riahi, Sara Riahi, “Big Data and Big Data Analytics: Concepts, Types and Technologies” in International Journal of Research and Engineering ISSN: 2348-7860 (O) | 2348-7852 (P) | Vol. 5 No. 9 | September-October 2018 | PP. 524-528.

[5] Martin, K. E. (2015). Ethical issues in the big data industry. MIS Quarterly Executive, 14(2), 67-85.

[6] Markus, M. L., & Topi, H. (2015). Big data, big decisions for science, society, and business. Bentley University.

[7] Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decisionmaking: A call for action on the long-term societal effects of “datification”. Journal of Strategic Information Systems, 24(1), 3-14.

[8] Chowdhury, Amin (2014). Dissertation Proposal on Big Data for Development: Assesing the fitness in M&E system. Module Assignment, Sheffield, Sheffield Hallam University.

[9] Zwitter, A. (2014). Big data ethics. Big Data & Society, 1(2).

[10] Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. Academy of Management Review, 22(4), 853-886.