Analysis of Ethical Issues in Data Science

Jasmeet Kaur

[1]*Student, Computer Science, Sri Guru Gobind Singh College of Commerce, Delhi University, Delhi, India.*

Corresponding Author: jasmeet10kaur@gmail.com

Abstract: - The twin drivers of data and information are driving technology and innovation forward in every aspect of human initiatives. Data science intensely influences how business is done in diverse fields as the life sciences, smart cities, and transportation. The dangers of data science without ethical consideration are apparent- whether be it the protection of personally identifiable data, the bias in automated decision making or lack of trust in virtual world. Also, data scientists need to be part of the community that establishes data science as proper profession in the view of Airaksien [1], a philosopher whose work focus on professional ethics.

Key Words: — ***Data Ethics, Data Science, Ethics of Data, Code of Ethics.***

I. INTRODUCTION

In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques [2]. Data Science provides huge opportunities to improve private and public life, as well as our environments. Unfortunately, such opportunities are also coupled to significant ethical challenges. The extensive use of increasingly Big Data—often personal if not sensitive (Big Data)—the growing reliance on algorithms to analyze them in order to shape choices and to make decisions (including machine learning, AI, and robotics), as well as the gradual reduction of human involvement or even oversight over many automatic processes pose pressing issues of fairness, responsibility, and respect of human rights, among others.

A systematic approach to identify and describe the ethical issues in data science was undertaken by CNIL (Commission nationale de l'informatique et des liberties') in 2017 [3][4]. In the present paper we would try to understand some of the ethical issues that need to be addressed through some real life case studies. We will also discuss the professional ethics code needed for this relatively new emerging field of data science.

Manuscript revised March 28, 2021; accepted March 29, 2021. Date of publication March 30, 2021.
This paper available online at www.ijprse.com
ISSN (Online): 2582-7898

II. WHAT IS DATA SCIENCE

There is no agreed definition of data science. Donoho [5] defines data science as follows:

Data Science is the science of learning from data; it studies the methods involved in the analysis and processing of data and proposes technology to improve methods in evidence based manner. The scope and impact of this science will expand enormously in coming decades as scientific data and data about science become ubiquitously available.

Donoho also provides a classification of the related activities into six divisions:

- Data gathering, preparation, and exploration
- Data representation and transformation
- Computing with data
- Data modelling
- Data visualization and presentation
- Science about data science

Item 6 is differentiated from other parts as what he calls *greater data science*.

III. ETHICAL ISSUES

Margo Boenig-Liptsin points out that our ever-increasing reliance on information technology has fundamentally transformed traditional concepts of "privacy", "fairness" and "representation", not to mention "free choice", "truth" and "trust"[6]. These mutations underline the increasing footprint and responsibilities of data science — beyond the bytes and

bits, data science shakes the perceptual foundations of value, community, and equity.

In this section we will look at the concerns that arises with the increasing volume of big data on the internet and hence lead to the need for professional code of ethics.

A. Privacy

Privacy is considered as a basic human need. Data Privacy or Information privacy is a part of the data protection area that deals with the proper handling of data with the focus on compliance with data protection regulations. Data Privacy is centred on how data should be collected, stored, managed, and shared with any third parties, as well as compliance with the applicable privacy laws like General Data Protection Regulation (GDPR [7]).

Since there is vast amount of data available in the virtual world, the chances of breach of privacy cannot be ruled out. There are drivers of privacy violation. The first driver is surveillance which can be done by government or a company. The second one is the metadata or big data stored at servers which can provide information like employee information, health records and so on. The third one is advertising. To show focused and relevant ad, personal data can be collected without the consent of individual.

Case study:

Aadhaar is a 12 digit number that uniquely identifies the citizens of India and Aadhaar database (established by Unique Identification Authority of India (UIDAI)) is one of the largest government databases on the planet. This database contains both the demographic as well as biometric data of the citizens. With the sheer amount of private and confidential data amassed in one singular database, concerns about the security of the private data of individuals are inevitable.

In 2018, the entire controversy around Aadhaar and privacy concerns captured centre stage after a french security researcher pointed the flaws in the mAadhaar app that was available on Google Play Store. The app based flaws had resulted in the leak of private data. For instance, an IIT graduate was arrested for illegally accessing the Aadhaar database without authorisation back in August 2017. He created an app called 'Aadhaar eKYC' by hacking into the servers related to an 'e-Hospital system'. These data leaks also lead to misuse of personal details (biometric and demographic) by others. For instance, the investigation by the Tribune in 2018, uncovered that anonymous individuals were ready to sell the Aadhaar card details of any individual with an Aadhaar number against the payment of a sum of ₹500 [8][9][10].

B. Anonymity

Data anonymization is the process of protecting private or sensitive information by erasing or encrypting identifiers that connect an individual to stored data [11].

One way to ensure anonymity is the de-identification. De-identification is the removal of identifiable information data. For example, if attributes like name, phone, and address are taken out of the data set, there are no personally identifying attribute values in the data set. But it also has its flaws.

Case Study:

In the Massachusetts Re-identification case, state General Insurance Commission released de-identified health records with the aim to serve the public good. Though the records were de-identified to ensure privacy protection of data, the researcher Latanya Sweeney used these to locate the health records of the governor William Weld. And in particular she was able to look up his diagnosis and prescriptions and she sent him a letter with all his intimate health history to make his point [12].

C. Validation of Data

Data validation means checking the accuracy and quality of source data before using, importing or otherwise processing data. Different types of validation can be performed depending on destination constraints or objectives [13]. There are multiple reasons of having errors in data analysis. There may be defect with the choice of representative sample since there is limitation of what data is present for analysis. Another reason could be errors in data processing or errors in data that might result from misinterpretation of available data. For example, sarcasm makes interpretation of data difficult when the task is to extract sentiment from the text. There could be errors in the model design itself.

Case study:

Another source of error with the data is change itself. Managing change is critical to maintaining validity of the results. Most changes may not impact analysis but some do. There is famous case of Google Flu predictor which worked well for a while and then crashed. First introduced in 2008, the idea of this product was that Google would use search terms that would enter into their search engines to estimate prevalence of flu. It's generally believed that this was because it dependent heavily on Google search algorithms which aren't static [14][15].

D. Bias and discrimination

Algorithms and artificial intelligence can create biases, discrimination and even exclusion towards individuals and groups of people [3]. The biases can be there because of different reasons. The first being that training data set isn't representative of the population. Secondly the past population isn't representative of future. For example, one issue with fairness of algorithm is where we have correct results but they are misleading or unfair in some way.

Case Study:

The probably most famous example is COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) a software used in the US judicial system to classify the probability of defendants' recidivism [16].It was shown in a detailed analysis [17] that the privately owned algorithm used in the juridical system gave far better prognoses for white than for black people, thus it discriminated implicitly based on colour. The machine generated prognosis was intended just to help the judges, but in interviews it could be seen, that it played a crucial rule in the judgements. Especially decisions by the judges whether defendants could get out on parole or had to go to jail were strongly influenced by the algorithm's output and discriminated against black people.

IV. CODE OF ETHICS

Professional codes are present in many professions. For example Hippocrates was a Greek doctor who introduced the notion of Hippocratic Oath [18]. The demand for inclusion of ethics in the different field and professions of digital technologies has been growing. Websites like fatml.org (fairness, accountability and transparency in machine learning) [19] provides overview on the ethical values in information and technology field. Data science is an evolving field that have professionals from different fields of computer science, statistics, and mathematics. But there need a consensus on the fact that ethical guidelines are a shared interest. Since universities where the course of data science is taught are the places where people from different disciples meet, there should be inclusion of professional ethics should be part of the curriculum.

To set general guidelines, the existing guidelines can be reached out. The ethical guidelines of ASA (American Statistical Association) [20] are detailed for the issues of data scientists in the sense of Donoho [5]. The code of ethics and professional conduct by Association of Computing Machinery (ACM) [21] outlines the societal impact of implementing algorithms related to data science and concerns about privacy of human data.

Regulation and compliance are the twin essentials for the development of any profession. But for the profession of data science, they may be necessary but not sufficient. There are many other factors involved that need to be taken care of. But since the rules can be complex, vary from organization to organization or list can be long enough to be memorized, there are two basic points every data scientist should follow. First, own the outcomes. The results available after analysis of data, optimization of data and the impact of the outcomes after analysis or optimization on the individual and society should be owned by data scientist. When they consider the consequences of their action they would be more ethical in their approach. Second, respect the data owner. Since, vast amount of data is available in the digital era; data science professional should keep in mind the rights of owner of data and make a careful and conscious decision of which data can be used.

V. CONCLUSION

This research paper is a theoretical approach to understand the ethical issues and challenges involved with data science and the need for ethical professional code for the community of data scientists. There is a need for understanding that the morality of the data science community is evolving and that it is a shared task to develop it, which in turn needs open discussions.

REFERENCES

[1]. https://www.researchgate.net/publication/264555968_Big_Data_Analytics_A_Literature_Review_Paper
[2]. https://www.researchgate.net/deref/https%3A%2F%2Fwww.cnil.fr%2Fen%2Fhow-can-humans-keep-upper-hand report-ethical-matters-raised-algorithms-and-artificial-intelligence
[3]. https://www.researchgate.net/deref/https%3A%2F%2Fwww.cnil.fr%2Fen%2Falgorithms-and-artificial-intelligence-cnils-report-ethical-issues
[4]. David Donoho (2017), 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766.
[5]. https://data.berkeley.edu/news/it%E2%80%99s-time-data-ethics-conversations-your-dinner-table
[6]. https://www.firstpost.com/india/an-iit-graduate-has-been-arrested-for-illegally-accessing-the-aadhaar-database-report-3892045.html
[7]. https://www.firstpost.com/tech/business/twitter-user-highlights-potential-security-flaws-in-uidais-maadhaar-app-

for-android-devices-user-data-could-be-compromised-4298719.html
[8]. https://m.tribuneindia.com/news/archive/nation/rs-500-10-minutes-and-you-have-access-to-billion-aadhaar-details-523361
[9]. https://www.informatica.com/in/services-andtraining/glossary-of-terms/data-validation-definition.html
[10]. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/
[11]. http://governance40.com/wp-content/uploads/2019/03/Weapons-of-Math-Destruction-Cathy-ONeil.pdf
[12]. https://www.researchgate.net/deref/https%3A%2F%2Fwww.propublica.org%2Farticle%2Fhow-we-analyzed-the-compas-recidivism-algorithm
[13]. https://www.medicinenet.com/hippocratic_oath/definition.htm
[14]. https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx
[15]. https://www.acm.org/code-of-ethics