SMU Data Science Review

Volume 1 | Number 2 Article 2

2018

On Identifying Factors Affecting Ethical Practices in Data Science Domains

yanqin wang
Southern Methodist University, yanqinw@mail.smu.edu

Earl Shaw
Southern Methodist University, edshaw@smu.edu

Brian Kruse
Southern Methodist University, bkruse@smu.edu

Mehdi Ghods
m1.ghods@comcast.net

On Identifying Factors Affecting Ethical Practices in Data Science Domains

Yanqin Wang[1], Earl Shaw[1], Mehdi Ghods, Brian Kruse[1]

[1] Southern Methodist University, Dallas, Texas 75205
{yanqinw, edshaw, bkruse}@mail.smu.edu, m1.ghods@comcast.net

Abstract. In data science domains, ethics and ethical approaches are important to minimize adverse effects that may arise in data collection, analysis, and storage. What factors are influential for ethical practices in data science? In this research study, we designed a survey to capture an assessment of ethical concerns and practices from those currently active in the field by soliciting the attitudes/feelings of data science students and practitioners via the questionnaire. We analyzed the extent of their attitudes and identified factors contributing to the difference.

1 Introduction

We are living in a world in which humans, and even machines, are producing a vast amount of data. Storage of this data, along with analysis on it using data science tools such as machine learning aids in decision-making for many business entities, nonprofit organizations, governments, and even individuals. While this can provide a host of benefits, we should be increasingly alarmed as more personal data is collected on individuals and the checks and balances currently in place on the storage and use of this data is minimal. Questions we should continually raise are: What data should not be collected? How should data be stored and protected after collection? And to what extent should the data be used for analysis?

To address these questions, one approach could be to pass laws regulating data usage. For example, the European Union's General Data Protection Regulation (GDPR) was approved in 2016, while in numerous other countries similar laws were discussed or were in the process of approval [1]. Another approach to prevent the misuse and abuse of an individual's information could be to establish a core set of ethical practices in the data science domains.

Data science uses computing technology that raises ethical and social issues that differ fundamentally from other technologies [2][3]. One typical example is cloud computing, which uses the most advanced technologies in data science [4]. When users store personal data in the cloud, they lose full control of who has access to it. This represents an obvious ethical issue, as it creates a situation in which an individual's data could potentially be viewed and/or used by a party with no right to do so. Moreover, with

more organizations expanding their use of cloud computing and cloud storage, if not managed properly, this could exacerbate the access issue. So it becomes increasingly necessary to establish a set of ethical guidelines for cloud users.

Another field generating numerous ethical issues is social networking. Social media platforms such as Facebook and Twitter collect a large amount of personal information from their users who may not be well informed about how their data is used. This information may be provided to the user prior to using the social media platform, but considering the length of such documents and the sophistication with which they are written, it is safe to say the average user may have trouble understanding much of it. So, a key question to consider is: "How do we define unethical actions in this domain and prevent inappropriate use of users' personal information?" Currently, we rely largely on society to explicitly perceive, define, and agree that an act is wrong and therefore unethical [5]. These norms and values cannot match the rapid emergence of new technologies used in various data science domains. Therefore, analyzing ethical issues and performing ethical practices has become a daunting task.

While there are some general guidelines for data science ethics, each organization may apply them differently [6][7][8]. For example, different organizations may have different standards for ethics in data protection due to various business needs. Additionally, at the core of data science, practitioners may judge ethics differently due to their age, gender, working experience or educational background. These issues decrease the efficiency of guidelines currently used in data science fields. Therefore, we must consider the question: "How can we build a set of guidelines effective to most ethical concerns in data science?" To address this a critical step is to identify primary factors affecting ethical practices in the data science field.

In this study, we aim to explore data science students, practitioners, and individuals in data science management positions about their perceptions and their organizations' attitudes towards ethics in data science. To this end, we designed a survey to capture an assessment of ethical concerns and practices from these data science students and practitioners. We analyzed the extent of their attitudes/feelings solicited via the questionnaire and identified factors contributing to the difference in ethical practices.

2 Research Methods and Designs

The attitudes-based nature of this research requires the construction of a survey questionnaire for gathering perception data from the intended population. A literature search was conducted to extract usable items and questions from surveys related to ethics and ethical practices to help in constructing the survey. The intended survey instrument was based on research developed by Likert [9]. All variables were represented by attitude and feeling related questions based on a theoretical rating scale [9][10]. The design included generic demographic related questions, along with a range of perception-based items, and a section for overall thoughts, comments and/or suggestions.

Next, the survey questionnaire was validated before being sent out to collect responses. Face validity (or logical validity) is one approach to validate the survey questionnaire. It is designed to determine whether the survey questions actually measure the attitudes of data science learners and practitioners [11]. Employing a research method premiered by Ghods [10], a minimum of four experts with wide ranging experiences in survey methodology and research were asked to review and rate each item of the questionnaire and the survey as a whole. Analysis of the ratings provided helpful insights toward revisions of the survey. The questionnaire initially contained 28 questions. Three of the initial questions (item 11, 12 and 13) were removed, as they did not show enough face validity, leaving 25 questions with a face validity score of 0.79, which is between the established guidelines of 0.75 and 1, providing evidence of very high face validity (Table 1).

Following the face validity for the survey, the questionnaire was posted online for convenient access: https://survey.zohopublic.com/zs/ViB0nw. The web link along with the e-consent form was distributed to eligible participants of the sample population based on SMU's Institutional Review Board (IRB) protocol. The intended population for the purpose of this research included data science and analytics practitioners at organizations and companies in our professional network, in addition to students in the data science program at SMU. Selection of this population allowed for understanding how practitioners perceive and approach ethics and ethical practices prior to and during their professional work.

Table 1. Face validity for all items in the questionnaire

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	F-Score
Item 1	4	4	3	4	4	3	0.917
Item 2	3	3	3	4	4	4	0.875
Item 3	4	4	3	3	4	4	0.917
Item 4	4	4	3	3	4	4	0.917
Item 5	4	4	3	3	4	4	0.917
Item 6	3	3	3	4	3	3	0.792
Item 7	4	4	3	4	3	2	0.833
Item 8	2	3	3	4	3	4	0.792
Item 9	3	4	3	3	4	4	0.875
Item 10	3	3	3	4	4	4	0.875
Item 11	2	1	1	1	4	4	0.542
Item 12	2	1	1	2	4	4	0.583
Item 13	2	1	1	1	4	4	0.542
Item 14	3	4	3	4	2	3	0.792
Item 15	4	4	3	4	3	4	0.917
Item 16	4	4	3	3	4	4	0.917
Item 17	4	4	3	3	4	4	0.917
Item 18	1	4	3	4	4	4	0.833
Item 19	3	4	3	3	4	3	0.833
Item 20	3	4	3	4	2	4	0.833

Item 21	3	4	3	4	4	3	0.875
Item 22	3	4	3	3	4	4	0.875
Item 23	3	4	3	4	4	3	0.875
Item 24	4	3	2	3	4	4	0.833
Item 25	3	4	2	4	4	3	0.833
Item 26	3	4	3	3	4	4	0.875
Item 27	3	4	3	4	4	1	0.792
Item 28	4	4	3	4	4	1	0.833

Prior studies of this type of survey support a sample size of 80-100 responses for the analysis methods required for the data and its intended results [12][13][14]. The final data set was prepared for analysis from the testing of the hypotheses to the following range of statistical analysis methods appropriate for this survey-type data.

The primary analysis methods and procedures are based on Structural Equation Modeling (SEM), which is a hybrid between Analysis of Variance (ANOVA) or regression, and factor analysis, allowing exploration into relationships among factors. Initially, to form and establish the set of constructs for the study, the collected data is exposed to exploratory and confirmatory factor analyses (EFA and CFA) [15]. The resulting set of variables/constructs form the basis for hypothesis testing, various regression models and procedures to answer the questions initially formed. Further statistical methods and analyses are performed to provide for a comprehensive results and conclusions, and for implications and recommendations drawn from the survey data. In particular, for testing the hypotheses, a stepwise regression algorithm was applied to fit the nature of specific variables within this study. Attempts were made to determine the second order constructs, if any, for a deeper analysis into understanding how a collection of variables influence the choices. For all analyses, a number of software applications/tools were used, such as the advanced statistical packages SAS and SPSS, in addition to AMOS and/or LISREL for SEM.

3 Results

3.1 Exploratory Analysis to Examine the Profile of Respondents

A total of 90 responses were collected from the survey over a 2-month period. All items in the questionnaire received more than 64% response.

We first explored the profile of respondents based on information from age, gender, highest level of education/degrees, and area of education/degrees (Fig. 1). A majority of respondents were male and older than 25 years of age. Additionally, close to 57% of them have a master's degree or higher and their degree varies from data science to business.

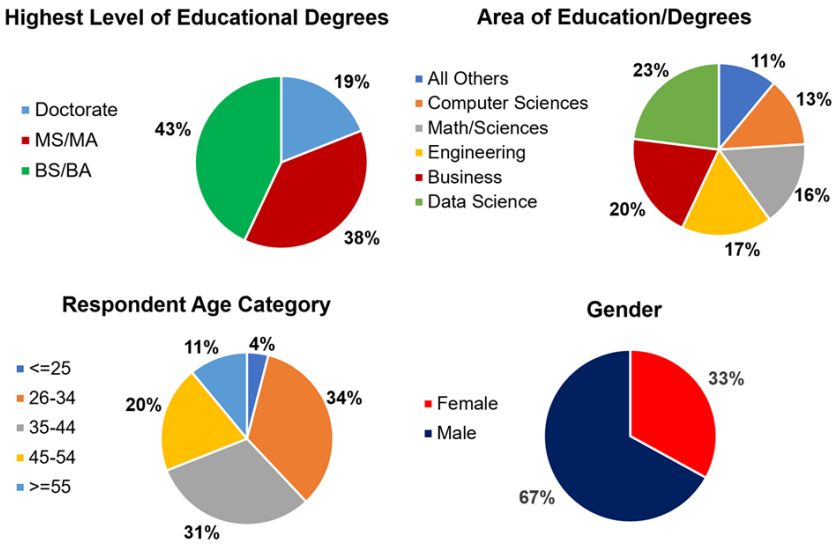

Fig. 1. Profile of respondents.

Next, we examined how respondents think in regards to the importance of ethical practices in data science. We found that practicing data science ethically is important to over 80% of respondents based on responses to the question "practicing ethics in my data science work/functions is one of the most important things in my life" (Fig. 2). Moreover, a majority of respondents think it is essential to incorporate established ethical standards in their data science practices (Fig. 3 top); but many are unsure as to whether or not their organizations will support and emphasize those practices (Fig. 3 bottom).

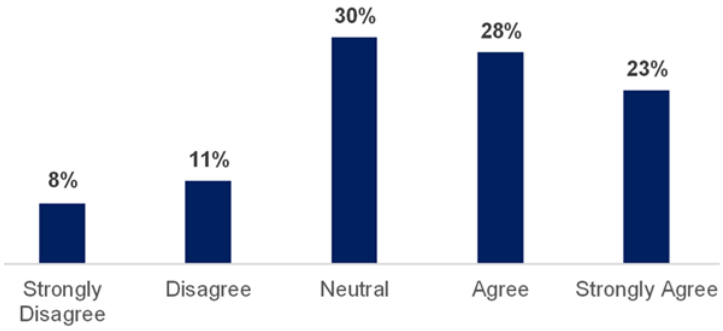

Fig. 2. Attitudes towards the importance of practicing data science ethically.

The respondents represented diverse educational backgrounds, a variety of age groups, and possessed different skill levels, in addition to their gender differences. Some of the questions we considered in response to the data are: (a) Are there correlations between variables such as respondents' age, gender, educational level/background or skills with the extent of their satisfaction towards ethical practices in data science? (b) If so, how does the variable contribute to the difference?

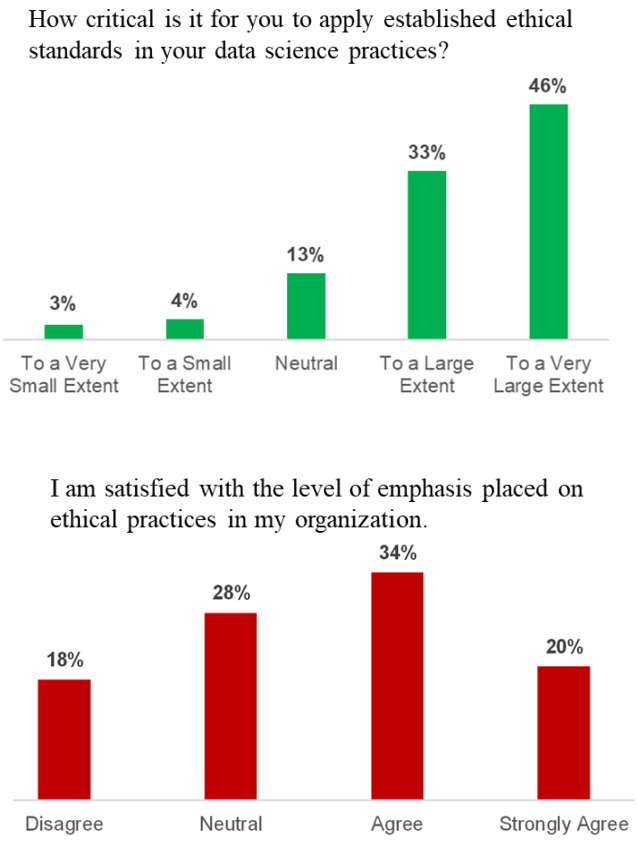

Fig. 3. Attitudes towards support of ethical practices from organization.

3.2 Primary Factors Affecting the Extent of Data Science Ethical Practices in Organizations

We built a model to measure the relationships between the predictors 1 – N (in our case, independent variables such as age, education level, experience…) and the dependent variable (e.g., satisfaction). A stepwise regression method of the analysis of variance was applied to determine the relationship of predictor 1 with the dependent variable (positive or negative). If the relationship was significant at the .05 level of significance

(95% level of confidence), the variable was retained in the model, and then the second predictor (variable) was tested. This process continued until testing for the remaining predictors (3 – N) was completed. In this model, R^2 shows the percentage the selected predictors contributed to the variance of the model. The value for "F" is from the F-test, while "p" shows the level of significance. Finally, regression coefficients are denoted by "b".

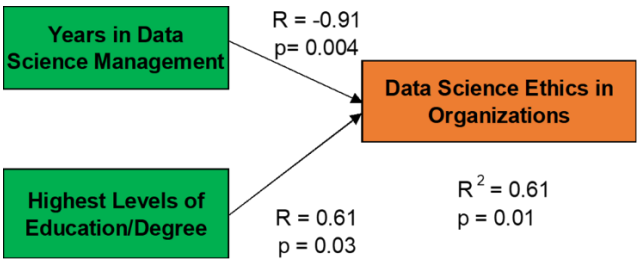

Fig. 4. Primary factors affecting the extent of ethical practices in data science.

As mentioned above, we tested a list of variables derived from respondents against the extent of ethical practices. We discovered that age, gender, and educational background have no significant impact on the extent of ethical practices in data science. On the other hand, we found that 61% of the total variance related to the extent of ethical practices in data science was explained by the years of practice in management and by advanced degrees (Fig. 4). The more advanced the degree, the greater the positive impact on attitudes towards ethical practices in data science. Additionally, the more years in data science management, the more extreme negative feelings with respect to organizational ethical practices.

3.3 Primary Factors Affecting the Emphasis Placed on Data Science Ethics in Organization

Although 80% of the respondents feel ethical practices are important in data science and are willing to apply the established ethical standards in practice, they are not certain their employers share their values. Thus, in order to have successful ethical practices in data science, the attitudes of both data science practitioners and their organization must be shared to some degree. After this discovery, our attention turned to explore what factors would affect the ethical emphasis in organizations.

In our questionnaire, items 1 to 25 were designed to obtain the attitudes for the following areas in data science: extent of ethical practice in the organization, ethical barriers within the organization, ethical emphasis in the organization, direction of ethical practices, personal commitments to ethical practice, and ethical practice impacts on the organization. We applied the step-wise model using the scores from the 6 areas and found that personal commitment to ethical practice and the impact of ethical practice on the organization determine the ethical emphasis in the organization (Fig. 5). The model suggested that for an organization to support ethics in data science, the

ethical practices need to influence an organization positively and significantly, in addition to there being a positive personal commitment to the practice.

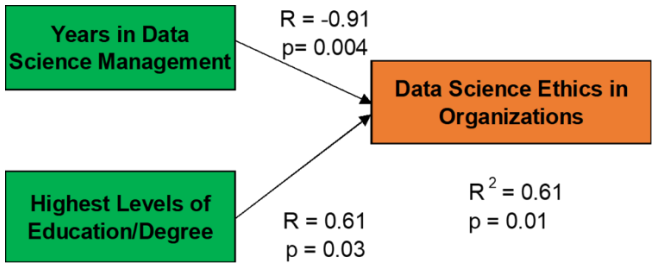

Fig. 5. Primary factors affecting the emphasis placed on ethics in organization.

4 Conclusion

Overall, ethics plays an important role in regulating data misusage in data collection, analysis, and storage. As such, developing effective guidelines for ethics in data science is vital. In this study, a survey was designed to assess ethics and ethical practices in data science and the related work environment. Scores reflecting attitudes/feelings of data science students and practitioners were recorded and subjected to further analysis. Additionally, we identified primary factors affecting ethical practices from two different perspectives: one was related to individual data practitioners and the other related to the organization in which practitioners work. Considering this, we discovered that practitioners and organizations should strive to communicate more regarding ethical practices in data science to ensure their values are in agreement. We also found that two primary factors: years of practice in management and advanced degrees, affect the extent of ethical practices in data science. That is, the more advanced the degree(s), the greater the positive impact on attitudes towards ethical practices in data science. However, with more years in data science management, feelings on organizational ethical practices veer in a more extreme negative direction. Furthermore, we found that personal commitment to ethical practices and the impact of ethical practices on an organization determine the ethical emphasis within the organization. Therefore, if the goal is to garner more support for ethics in data science across an organization, the ethical practices need to influence the organization positively and significantly, and a consistent personal commitment to ethical practice by employees must be present. To do this a culture among practitioners of personally committing to ethical practices should be fostered.

Going forward, considering the technologies used in data science evolves rapidly, the ethical issues raised by these technologies will evolve rapidly also. To align with the changes in ethical issues, the guidelines should be updated frequently. In future survey studies, the field of occupation or intended occupation of the respondents should be considered since occupations might provide a different understanding for data misuse.

Moreover, more details are required on the definition of established ethical behaviors so that all respondents will have an agreed upon basis.

References

1. Osterman Research, Inc. (2017). A Practical Guide for GDPR Compliance. Osterman Research.

2. Wiener, N. (1950/1954). The Human Use of Human Beings: Cybernetics and Society. Houghton Mifflin, 1950. (Second Edition Revised, Doubleday Anchor, 1954.)

3. Stahl, G., Tung, R. (2016). Stahl & Tung 2016.

4. Durao F., Carvalho J.F., Fonseka A., Garcia V.C. (2014). A systematic review on cloud computing. The Journal of Supercomputing, Volume 68, Number 3, Page 1321

5. Stahl, G. et al (2012). Six principles of effective global talent management. Sloan Management Review, 53, 25-42. MIT Sloan Management Review. 53. 25-32.

6. Gass S. I. (2009). Ethical guidelines and codes in operations research. Omega 37, 1044-1050.

7. American Statistical Association, Ethical Guidelines for Statistical Practice, August 7, 1999.

8. U.S. federal regulations regarding human subjects protection are contained in Title 45 of the Code of Federal Regulations, Chapter 46 (45 CFR 46).

9. Likert, R. (1932). A Technique for the Measurement of Attitudes. Archives of Psychology 140, 1–55.

10. Ghods, M. (1980). Michigan extension agents attitudes towards computers and computerized extension forward planning and consulting programs: The telplan system. East Lansing: Michigan State University.

11. Gravetter, F. J.; Forzano, L. B. (2012). Research Methods for the Behavioral Sciences (4th ed.). Belmont, Calif.: Wadsworth. p. 78. ISBN 978-1-111-34225-8.

12. Fowler, F. J. (1995). Improving survey questions: Design and evaluation. Thousand Oaks, CA: Sage Publishing, Inc.

13. Ho, F.W. (2005). Survey as a source of statistics and factors affecting the quality of survey statistics. International Statistics Review, 73(2), 245-248

14. Hole, K. (2015). An Investigation of the quality of student-developed surveys and rating scales and psychometric reporting practices in doctoral dissertations. University of Kansas. ProQuest Dissertation Publishing. 3713528.

15. Kline, R. B. (2010). Principles and practice of structural equation modeling (3rd ed.). New York, New York: Guilford Press.

Appendix: Ethics and Ethical Practices in Data Science Discipline – A Survey Measurement

This questionnaire is designed to assess ethics and ethical practices in data science (DS)/analytics functions and related work environment. More specifically, the intent is to understand the application and implications of ethics in data science as it relates to the university degree programs and in the actual practice settings. This study is also being undertaken to supplement the requirements for the Capstone Project/research of the Masters' Program in Data Science (MSDS) at Southern Methodist University (SMU). The data gathered will be instrumental in the understanding of how ethics and ethical practices are perceived by MSDS students, as well by professionals in practice.

It is important that you answer the questions in the survey carefully and honestly. This measurement will be valuable only if you record what you really think and feel.

We have taken measures to ensure the strict confidentiality of your responses:

- The data will be collected and analyzed by three MSDS students at SMU (below) with the supervision of faculty and an industry advisor, at large. These individuals will be the only ones who will have access to the coded data and the database. The completed questionnaires will be destroyed leaving the coded data as the only record of your responses

- Your responses are anonymous. No reference to the identity of the respondent or the organization is included in the questionnaire.

- All analysis of data and reporting of results will be done at a summary level, aggregating responses from a number of individuals. No analysis will be performed nor will any data be reported at the individual level.

Instructions

There are two sections in this questionnaire: Section 1 contains demographics data and quantitative questions related to your organization/program. Items and questions related to your perceptions, thoughts and feelings are included in Sections 2.

Throughout these sections, the term "organization" refers to the entity such as a company, firm or your place of employment. The term "program" is intended to include a university degree program. In this way, all questions and items of the survey can be answered by both students and data science professional/practitioners. The last item of the survey is meant to gather thoughts, comments and suggestions, if any, with respect to ethics and data science practice. The comments in text format will be subjected to text mining collectively for inclusion in the overall analysis of data.

Thank you for your time. If you have any questions please contact the undersigned.
Brian Kruse, bkruse@smu.edu

Earl Shaw, edshaw@smu.edu
Yanqin Wang, yankinw@smu.edu

Section 1: Organization, Functions and demographics data

(Note: You may not have the exact number for some of the questions in this section. Regardless, please provide your best estimate)

- Are you a current Data Science (DS) student or a DS/Analytics practitioner?
  __Student, __ Practitioner

- If, a practitioner, how long have you been in practice? __Years.

- Total number of years of experience in DS (irrespective of years of practice and/or being a current student)?
  __ Years

- Have you ever been in a management position in DS? __Yes, __No

  If Yes, how many years in DS management? __Years

- Your highest level of education/degree?
  __No Degree, __ BS/BA, __ MS/MA, __ Doctorate

- Please specify areas of education/degrees other than DS: _____

- What is your age category?
  __ (25 & under), __ (26-34), (35-44), __ (45-54), (55 & above)

- What was/is the field of study for your highest degree? _____

- Gender? ____ M, ____F

Section 2: Feelings/perceptions with respect to ethics and ethical practices in data science

	To A Very Small Extent		**To Same Extent**		**To A Very Large Extent**
To what extent are ethical standards and procedures set out to guide the data science practices in your organization/program	**1 __**	**2 __**	**3 __**	**4_**	**5 __**

To what extent has your organization /program developed strategies and procedures to deal with demands for ethical practices in the field of data science **1 __ 2 __ 3 __ 4 __ 5 __**

How critical is it for your organization /program to establish ethical policy and practices in performing analytics **1 __ 2 __ 3 __ 4 _ 5**

How critical is it for you to apply established ethical standards in your data science practices **1 __ 2 __ 3 __ 4 __ 5 __**

	Strongly Disagree		**Neutral**		**Strongly Agree**
I feel I am not prepared to deal with unethical practices in my data science work/functions	**1 __**	**2 __**	**3 __**	**4 __**	**5 __**
I feel I am expected to ignore established ethical practices in my analytics work/functions	**1 __**	**2 __**	**3 __**	**4 __**	**5 __**
Practicing ethics in my data science work/ functions is one the most important things in my life	**1 __**	**2 __**	**3 __**	**4 __**	**5__**
My organization inspires the very best in me when it comes to combine established ethical standards with data science work/practice	**1 __**	**2 __**	**3 __**	**4**	**5 __**
I am satisfied with the level of emphasis placed on ethical practices in my organization/program	**1 __**	**2 __**	**3 __**	**4**	**5 __**
I am willing to put greater time and effort than normally expected in order to fully comply with data science ethical standards	**1 __**	**2 __**	**3 __**	**4 __**	**5 __**
My clients/customers with whom I work/deal generally expect ethically accepted outcomes from my analytics work	**1 __**	**2 __**	**3 __**	**4 __**	**5 __**
Management/administrators in my organization /program go out of their way to emphasize the importance of applying ethics to data science work/practice	**1 __**	**2 __**	**3 __**	**4 __**	**5 __**
Management/administrators in my organization	**1 __**	**2 __**	**3 __**	**4 __**	**5 __**

/program rarely consult with me about practical applications of ethics in my data science work/practice

In recent years, our organization/program has been developing programs to educate employees in ethics of data science **1 __ 2 __ 3 __ 4 __ 5 __**

As a result of our ethics related programs, our clients/customers have increased their understanding of the role ethical practices in data science work/functions **1 __ 2 __ 3 __ 4 _ 5__**

In general, the initiatives to incorporate ethical practices in Data Science work/functions have…

	To A Very Small Extent		**To Same Extent**		**To A Very Large Extent**
• …created a large demand on my workload affecting my main responsibilities	**1 __**	**2 __**	**3 __**	**4 _**	**5 __**
• … become a source of dissatisfaction my co-workers	**1 __**	**2 __**	**3 __**	**4 _**	**5 __**
• … diverted from my time/effort needed for technical/project work	**1 __**	**2 __**	**3 __**	**4 _**	**5 __**
• … been a source of conflict between us and our clients/customers	**1 __**	**2 __**	**3 __**	**4 __**	**5 __**
• **… not** been supported by our clients /customers	**1 __**	**2 __**	**3 __**	**4 _**	**5 __**

	Dramatically Decreased		**Remained Stable**		**Dramatically Increased**
In recent years, the emphasis placed in practicing data science/analytics ethically has	**1 __**	**2 __**	**3 __**	**4 __**	**5 __**
In recent years, the involvement of data science professionals in ethics related decisions has	**1 __**	**2 __**	**3 __**	**4 _**	**5 __**

In general and within the field of Data Science, initiatives with respect to ethical practices have ….

a)	… been one of the least important factors for our organization/program success		… had moderate importance for our organization/program success		… been among the most critical factors for our organization/program success
	1	2	3	4	5
b)	… had inconsequential organization/program impact		… had moderate organization/program impact		… had tremendous organization/program impact
	1	2	3	4	5
c)	… been largely ignored by organization/program		… often needed responses		… always needed responses
	1	2	3	4	5

Below, please provide your overall thoughts, comments, suggestions and points of interest with respect to ethics and ethical practices in Data Science discipline and practice: