Association for Information Systems

AIS Electronic Library (AISeL)

ICIS 2019 Proceedings

Cyber-security, Privacy and Ethics of IS

Empowering Users to Detect Data Analytics Discriminatory Recommendations

Sepideh Ebrahimi
York University, sepidehe@yorku.ca

Khaled Hassanein
McMaster University, hassank@mcmaster.ca

Ebrahimi, Sepideh and Hassanein, Khaled, "Empowering Users to Detect Data Analytics Discriminatory Recommendations" (2019). *ICIS 2019 Proceedings*. 39.
https://aisel.aisnet.org/icis2019/cyber_security_privacy_ethics_IS/cyber_security_privacy/39

Empowering Users to Detect Data Analytics Discriminatory Recommendations

Short Paper

Sepideh Ebrahimi
School of Administrative Studies
York University
Toronto, ON, Canada
sepidehe@yorku.ca

Khaled Hassanein
DeGroote School of Business
McMaster University
Hamilton, ON, Canada
hassank@mcmaster.ca

Abstract

Notwithstanding the various benefits ascribed to using Data Analytics (DA) tools in support of decision-making, they have been blamed for their potential to generate discriminatory outputs. Although several purely technical methods have been proposed to help with this issue, they have proven to be inadequate. In this research-in-progress paper, we aim to address this gap by helping users detect discrimination, if any, in DA recommendations. By drawing upon the moral intensity literature and the literature on explaining black box models, we propose two decisional guidance mechanisms for DA users: (i) aggregated demographic information about the data subjects (ii) information on the variables that drive the DA output and the extent of their contribution along with information about demographics of the data set being analyzed. We suggest that these mechanisms can help decrease users' readily acceptance of discriminatory DA recommendations. Moreover, we outline an experimental methodology to test our hypotheses.

Keywords: Data Analytics, discrimination, decisional guidance, demographic transparency, ethical decision-making, moral intensity, features importance

Introduction

Data Analytics (DA) tools are being increasingly used in organizations to better analyze the tremendous amounts of available data ("Big Data"), make data-driven decisions, and eventually enhance financial and strategic performance (Côrte-Real et al. 2017; Davenport and Bean 2018; Ghasemaghaei et al. 2018). DA tools are typically a combination of several processes and tools, including SQL queries, statistical analysis, data mining, fact clustering, and data visualization. Despite their observed benefits, the use of DA tools to support managers' decision-making has raised concerns including issues associated with privacy, control, and discrimination (Newell and Marabelli 2015; O'Neil 2016). In 2015 Gartner predicted that "by 2018 half of business ethics violations will occur through the improper use of Big Data Analytics" (Gartner 2015). Our focus in this study is on developing and testing methods to reduce the incidence of managers readily adopting possible discriminatory DA recommendations.

Whereas it is suggested that technology has been employed to mitigate the problem of discriminatory decisions made based on personal prejudice (Gates et al. 2002; Tene and Polonetsky 2013), the issue has not been resolved by using DA tools. On the contrary, it has been argued that using such tools and techniques can contribute to discrimination in societies (Johnson 2014; Lyon 2003; Newell and Marabelli 2015). For instance, it was shown that COMPAS, a predictive tool designed to generate scores for the risk of recidivism (i.e., the chance of a person committing another crime within two years if released) has a strong racial bias (Courtland 2018).

It is worth noting that such discriminatory outputs of DA tools are not produced because of the user or the DA tools basing the recommendation on sensitive variables (e.g., sex and race). In addition, excluding those sensitive variables from the analysis and/or only using them for training DA models do not resolve the problem of DA tools generating discriminatory recommendations (Lipton et al. 2018). In fact, biased or non-representative data and/or inadvertent modeling procedures in the DA tools are the main culprits to blame (Žliobaitė and Custers 2016). Existing research identifies three main reasons for the generation of discriminatory recommendations by unbiased DA algorithms: (i) relations between non-sensitive and sensitive attributes in data that lead to non-sensitive attributes acting as proxies for sensitive variables; (ii) erroneous/biased data labeling; as historical data is used to build and train DA models and therefore, discriminatory data can bring about discriminatory models (Custers 2013); and (iii) flaws in the data collection process that lead to some groups of individuals being over- or under-represented in the data sets used to train or test DA models (Barocas and Selbst 2016; O'Neil 2016; Schermer 2011).

As such, even fair and well-intentioned individuals can make a discriminatory decision drawing on a discriminatory recommendation generated by a DA tool. It should be noted that recommendations that treat a demographic class (e.g., females) less favorably than other class(es) are considered potentially discriminatory. However, according to civil rights legislations, such recommendations are actually discriminatory only when there is no legitimate business reason to explain the discrepancy (Yinger 1998). For instance, when making a recommendation about hiring salespeople, if a DA tool mainly puts forth names of male applicants, the recommendation is considered as potentially discriminatory. If it is found out, however, that there are legitimate business necessities behind such a situation (e.g., when driving a truck is a requirement for the job and fewer females have a license to drive trucks), then it can be concluded that the recommendation is not discriminatory against females. Similarly, throughout this paper, the notion of discriminatory recommendation refers to a recommendation that is potentially discriminatory and needs to be investigated further.

To date some technical methods have been suggested to discover and remove discrimination in DA recommendations (e.g., Dwork et al. 2012; Pedreschi et al. 2008). However, research is yet to find methods that eliminate such discrimination altogether (Žliobaitė and Custers 2016). Therefore, recent scholarly and practitioners' (e.g., Holstein et al. 2018; McDonald 2016) and governments' (Executive office of the [US] president 2014; Federal Trade Commission 2016) publications have raised concerns and awareness about the potential of making discriminatory decisions when using DA tools. Furthermore, in a statement on algorithmic transparency and accountability, the Association for Computing Machinery recently asserted "institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results" (Dopplick 2017). However, due to the complexity and opacity of DA tools and techniques, mostly in organizations understanding the causes and consequences of particular patterns found by DA tools are neglected and finding significant connections is considered as sufficient (Newell and Marabelli 2015). Thus, a DA recommendation, even if unethical or discriminatory, is often readily approved by users (Ebrahimi and Hassanein 2019).

Therefore, the main objective of this study is to devise decisional guidance (Silver 1991) tools to support DA users in being able to discern traces of discrimination, if any, in DA recommendations. To achieve this goal, we will pursue the following research objectives: (i) to employ decisional guidance tools that provide DA users with aggregated demographic information regarding the human subjects of the DA recommendations; (ii) to design and develop decisional guidance tools that provides DA users with information about the contribution of each feature (i.e., variable) used in the generation of a DA recommendation as well as the distributions of these variables across various demographic classes; (iii) to develop and empirically validate a theoretical research model explaining the impact of using the decisional guidance tools above on users' perceptions, their ability to identify potentially discriminatory DA recommendations, and finally their behavior in terms of accepting/rejecting these recommendations.

The remainder of this research-in-progress paper will unfold as follows. First, the theoretical background, the research model and proposed hypotheses are described. The methodology in support of the experimental design, data collection, and pertinent analyses is presented in the third section. Finally, concluding remarks, and potential contributions to theory and practice close the paper.

Theory and hypotheses development

The ethical decision-making process and Moral Intensity

Before proceeding, we wish to note that, in line with much prior ethical decision-making research, we use the terms moral and ethical interchangeably. So, for example, we consider the terms "moral recognition' and "ethical recognition" to be synonyms. To study the un/ethical behavior of approving/rejecting a discriminatory recommendation generated by DA tools, we draw on theories in the business ethics literature, where ethical behavior has been described as a systematic framework that involves making principled assessment in questionable situations (Rest 1986). Individuals engage in ethical behavior after they realize that the situation at hand has an ethical aspect (Craft 2013). This realization is called recognition of the moral issue or moral recognition. A person who fails to recognize the moral aspects of a particular issue will not activate her/his moral decision-making schemata and will make the decision based on other schemata such as economic factors, etc. (Douglas et al. 2001; Jones 1991). Recognition of the moral issue, then, prompts the decision maker to make a judgment of what potential action is most moral, establish a moral intent, and finally engage in a moral behavior (Rest 1986).

However, individuals do not automatically know that they are facing an ethical dilemma and that they should choose to act ethically (Trevino and Brown 2004). Jones (1991) proposed that the moral intensity of an issue influences moral issue recognition. Moral intensity of an issue is comprised of six factors: magnitude of consequences (i.e., sum of harms done to the victims), social consensus (i.e., social agreement that a proposed behavior is evil), probability of negative effect (i.e., likelihoods that the issue occurs and causes the anticipated harm), temporal immediacy (i.e., shorter interval between when the decision is made and when the consequences occur), proximity (i.e., the degree to which the decision-maker feels close to the victims), and concentration of effect (i.e., when the consequences affect fewer individuals as opposed to the same consequences being more broadly distributed) (Jones 1991).

Theoretical Model and Hypotheses

Figure 1 presents the proposed research model. As the figure depicts, two decisional guidance mechanisms (i.e., demographic transparency and Feature importance + demographic distributions) are suggested to increase moral recognition and moral behavior (i.e., decrease the acceptance of a DA discriminatory recommendations). Two moral intensity factors (i.e., probability of effect and magnitude of consequences) are suggested to play a mediating role in this process. Next, we elaborate on the logic of our proposed research model.

Demographic Transparency

The majority of discrimination measures rely on the outcome of a decision as opposed to the process of decision-making (Zliobaite 2015). One such measure used in several studies is called extended lift (elift), which measures how the probability of granting a benefit to an individual changes as a result of their belonging to a protected group (Pedreschi et al. 2008). More specifically, extended lift is defined as the ratio between the proportion of individuals from a protected class obtaining a benefit over the overall proportion of them in the data set.

Similarly, to detect traces of discrimination in a recommendation put forth by a DA tool, we draw on the notion of elift and suggest providing DA users with demographic transparency (DT), a decisional guidance mechanism that depicts statistics about the proportion of members of each demographic class (e.g., females and males) in both the original pool of all subjects and the DA recommended sample (to receive a specific benefit). For instance, imagine the case of internal applications for a position in an organization. If a DA tool is used to analyze the existing data and make recommendations, DT will provide the user with the proportion of female and male applicants in both the original pool of all applicants and the recommended sample of applicants to be further interviewed (see Figure 2 for an example of DT). We suggest that if the recommendation includes discrimination against a demographic group (e.g., females), receiving DT along with the recommendation will likely increase the user's perception of the probability of the negative effects. As defined by the moral intensity literature, the probability of negative effect assesses the likelihood that the decision outcome will actually take place and whether it will cause the anticipated harm (Haines et al. 2008). In the context of this study, the probability of a negative effect refers to the probability of the data analytics recommendation being discriminatory against a demographic class (e.g.,

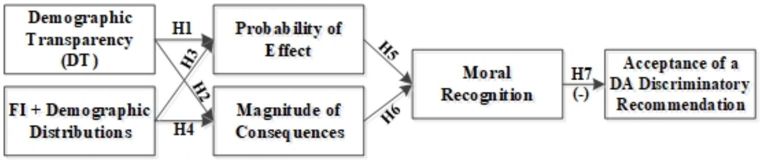
Demographic Transparency (DT)
FI + Demographic Distributions
H1
H3
H2
H4
Probability of Effect
Magnitude of Consequences
H5
H6
Moral Recognition
H7
(-)
Acceptance of a DA Discriminatory Recommendation

Figure 1. Proposed Research Model

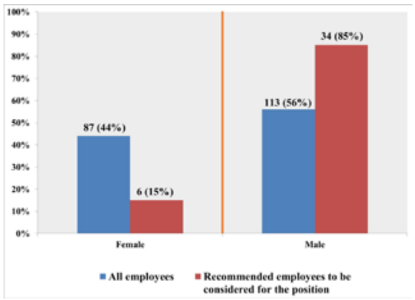

Figure 2. Sample of a DT Chart

females) and the negative effects of accepting/approving the recommendation on the members of that class. This is since if there is a substantial increase (decrease) in the proportion of males (females) in the sample recommended by the DA compared to the full data set of all applicants, then there is a possibility that females are being treated unfairly. Such a situation can take place due to several reasons such as when the data includes historical records of discrimination against women and the analysis includes variables (e.g., prior annual reviews, prior promotion decisions) in the analysis that bring in the prejudice of previous decision makers in the organization or variables that more adversely impact females than males (e.g., tenure as females are more likely to go on parental leave than males). Therefore,

H1: *In the context of a potentially discriminatory DA recommendation, providing users with demographic transparency charts will be positively associated with users' perception of the probability of negative effects.*

In addition to increasing perceptions of negative effects, we suggest that providing DA users with aggregated demographic information in the form of DT will bring about an increase in users' perceptions of the magnitude of the negative consequences. Magnitude of consequences of the moral issue is defined as the sum of the harms (or benefits) done to victims (or beneficiaries) of the moral act in question (Jones 1991). In the context of a discriminatory DA recommendation, demographic transparency presents DA users with the proportion and number of individuals from a protected class (e.g., females), who are denied a benefit. Therefore, DA users are more likely to have a sense of the sum of the harms approving such a recommendation can cause. Thus, we hypothesize:

H2: *In the context of a potentially discriminatory DA recommendation, providing users with demographic transparency charts will be positively associated with users' perception of the magnitude of the negative consequences.*

Features Importance and Demographic Distributions

DA users' readily acceptance of potentially discriminatory recommendations put forth by DA tools stems from the fact that in an organization only a few individuals actually understand the DA algorithms and what has been included in them (Lipton 2016; Newell and Marabelli 2015). Thus, to most users, a DA tool is a "black box". Methods have been suggested to explain a black box algorithm through an interpretable and transparent model that mimics the behavior of the black box and is easily understood by humans. Although several approaches (e.g., decision trees) have been proposed for explaining black box algorithms embedded in DA tools, they are not widely used (Johansson and Niklasson 2009).

Decision trees, decision rules, and features importance (FI) are the most frequently used approaches to generate interpretable models (Freitas 2014; Guidotti et al. 2018). Some researchers, however, have raised concerns about the interpretability of decision trees. This is due to the fact that in most real-life applications, the size of decision trees can grow very fast; and given the capacity of human cognition, they might become uninterpretable (Huysmans et al. 2011; Lipton 2016). Further, decision trees do not reveal the relative importance of features (i.e., variables) used in a model. These shortcomings may explain why decision trees are not widely used (Johansson and Niklasson 2009). Generally speaking decision rules have the same shortcomings as decision trees (Henelius et al. 2014). Thus, we argue that decision trees and decision rules are not well-suited for helping users to easily identify traces of discrimination in DA recommendations. On the other hand, FI is a simple yet effective way to provide explanations for black box models. It works by sharing with users the weights of the features used by a black box model (Lou et al. 2013; Vidovic et al. 2016). These weights are the coefficients of a linear model that is generated as an interpretable model approximating a black box model. Such weights convey to users the relative importance of the main features used by a DA algorithm to generate a specific recommendation. One advantage of FI is being an agnostic explanator. In other words, the frameworks used to generate them often work on almost all types of black box models (e.g., Naïve Bayes, Neural Networks) (Guidotti et al. 2018). Thus, in this study we employ FI as a concise and easily comprehensible method as the basis for our decision guidance tool to empower DA users to detect discriminatory DA recommendations.

We further suggest that if the information about FI is accompanied by data that delineates the distribution of each of the important features for each demographic class (e.g., females and males) in a demographic category (e.g., sex), then DA users are more likely to notice the possibility of potential discrimination in a DA recommendation. More specifically, if the distribution(s) of one/several important feature/s in the disadvantaged class (e.g., females) is significantly different from the corresponding distribution(s) in the other class(es), then DA users are more likely to have higher perceptions of the probability of the negative consequences of putting the recommendation in question into action. For instance, imagine the case of a DA making recommendations regarding which employees will receive a promotion in an organization. An example of FI + demographic distribution for this case is depicted in Figure 3. As can be seen in the figure, performance evaluation has 31% contribution to the recommendation. The distribution of this feature across the two demographic classes of males and females are shown in the diagram. Since the distribution of the performance evaluation feature is substantially different between the two classes, it can be concluded that the recommendation might include potential discrimination against females. It is noteworthy that such a distribution should be provided for other demographic categories (e.g., race and age) and for all the important features. In light of the above discussion, we hypothesize a relationship between providing FI + demographic distributions and DA users' perceptions of the probability of negative effects, if any. Therefore,

H3: *In the context of a potentially discriminatory DA recommendation, providing users with Feature importance and demographic distributions for each feature will be positively associated with users' perception of the probability of negative effects.*

Moreover, providing FI + demographic distributions, we suggest, can increase DA users' perceptions of the magnitude of negative consequences of a potentially discriminatory recommendation. This results from the fact that by reviewing this information, a DA user can estimate the proportion and number of individuals from a protected class who are treated unfairly by one/a few features included in the DA algorithm. For instance, by receiving FI + demographic distributions depicted in Figure 3, a DA user can see that about half of the female employees have received lower evaluation of performance than their male counterparts. Given the importance of this feature in generating the recommendation about employees to receive a promotion, the DA user will likely have a higher perception of the magnitude of negative consequences of approving the recommendation for female employees. Thus:

H4: *In the context of a potentially discriminatory DA recommendation, providing users with Feature importance and demographic distributions for each feature will be positively associated with users' perception of the magnitude of the negative consequences.*

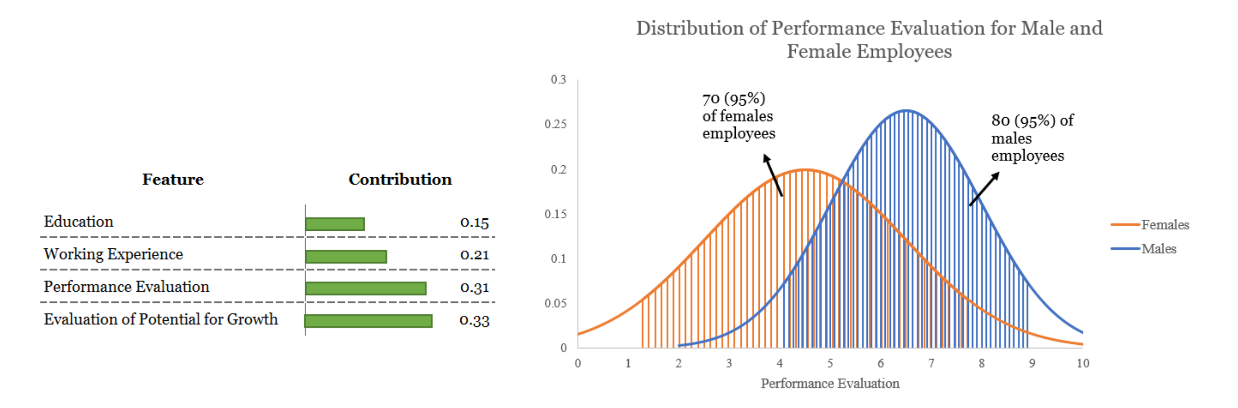

Figure 3. Sample of a FI + Demographic Distribution

Ethical Recognition and Approval of the Discriminatory Recommendations

Jones (1991) argued that the moral intensity of an issue influences decision makers' recognition of it as a moral problem and their subsequent behavior. Jones posited that issues that involve a high probably of causing harm (i.e., probability of effect) and the issues that can bring about large amounts of harm (i.e., magnitude of consequences) are more likely to be recognized most readily as moral issues. The concept of probability of (negative) effect is consistent with other theories in the business ethics literature. For example, Hunt and Vitell's (1986) notion of probabilities of consequences, which is a component of an individual's teleological evaluation of issues. Following Jones, several studies that empirically examined the relationship between probability of negative effects and moral recognition found significant effects (e.g., May and Pauli 2002; Singhapakdi et al. 1996). Similarly, we expect a DA user who realizes the possible negative effects of approving a discriminatory DA recommendation to be more likely to form a moral recognition about the situation at hand. Therefore,

H5: *In the context of a potentially discriminatory DA recommendation, the greater the perceived probability of negative effect, the more likely that a DA user will recognize the moral issue at hand.*

Magnitude of consequences, the other dimension of moral intensity that is of interest in this study, suggests that as consequences of an issue becomes more severe, a decision maker is more likely to recognize the ethicality of the issue (Jones 1991). Previous empirical research has confirmed that a relationship exists between magnitude of negative consequences and a decision-maker's moral recognition and behavior (e.g., Singhapakdi et al. 1996; Watley and May 2004). Likewise, when a DA user realizes that approving a DA recommendation can have negative consequences for a protected group (e.g., several females being denied a chance to receive a promotion), s/he is more likely to recognize the fact that s/he is facing an issue with moral facets. As such, we hypothesize that:

H6: *In the context of a potentially discriminatory DA recommendation, the greater the perceived magnitude of consequences, the more likely that a DA user will recognize the moral issue at hand.*

Ethical reasoning has been described as a systematic framework that involves making principled assessment in questionable situations (Ferrell et al. 1989; Rest 1986). Therefore, when individuals realize an ethical situation, they consider and evaluate courses of actions based on their morality. Such assessments subsequently affect their ethical intentions and actions (Loe and Weeks 2000; Rest 1986). Therefore, it is reasonable to expect that users who become morally aware of a potentially discriminatory recommendation of a DA system they are using are more likely not to accept that recommendation compared to users who are not aware. Hence,

H7: *In the context of a potentially discriminatory data analytics recommendation, users' recognition of the moral issue is negatively associated with their acceptance of the recommendation.*

Methodology

The proposed research model will be empirically validated through an online experimental study involving middle managers who use DA tools for decision-making at work. It is noteworthy that the real purpose of the study will not be revealed to participants until after they complete the experiment. Specifically, participants will be told that the purpose of the experiment is to better understand how data analytics users interact with DA tools. They will be asked to work with the tool and decide whether they approve its recommendation.

A full-factorial design will be employed using all four combinations of the two treatments: the presence or absence of DT charts and the presence or absence of FI along with demographic distributions of each of the features. A fictitious DA tool will be designed for the experiment. The tool, drawing on various objective and subjective factors, will make classification recommendations for promotions in an organization. However, a few of the subjective factors (e.g., annual reviews) will include traces of historical discrimination against a group of employees and therefore will bring in the discrimination to the generated recommendation (Barocas and Selbst 2016). As a result, the proportion of members of the disadvantaged group who are recommended for promotion will be significantly lower than that for the other groups.

Participants in the control group will not receive any decisional guidance while participants assigned to groups with DT and/or FI will receive the respective decisional guidance information along with the DA recommendation. Participants will then be asked to make a decision about whether they accept the recommendation. Subsequently, they will answer a survey about the reason/s for their decision, manipulation checks, and their perceptions of the other variables in the model. To ensure content validity, scales for all constructs will be selected from the extant literature with appropriate adaptation to context. Recognition of the moral issue will be measured using a 3-item scale from Reynolds (2006) and the two moral intensity constructs (i.e., probability of negative effect and magnitude of negative consequences) will each be measured using a 2-item scale adapted from Singhapakdi et al. (1996) and Frey (2000).

Power analysis indicates that 172 subjects (43 subjects for each group) would assure a sufficient statistical power of 0.80 to detect a medium effect size (f= 0.25) (Cohen 1988). To account for potential outliers and spoiled or incomplete responses, 200 participants will be recruited. Structural Equation Modeling (SEM) will be used to validate the proposed model. Further, ANOVA analysis will be performed to compare the differences between the four treatment groups in regard to the mediating and endogenous variables.

Conclusion

Potential Contributions

This study will be one of the first empirical studies to focus on examining the issue of discriminatory decision-making using DA tools and strives to tackle the problem by focusing on the human aspects of decision-making using such tools. Thus, this research will advance the DA literature by proposing decisional guidance tools to address the issue by focusing on the cognition and attitude of the human decision maker. The methods proposed in this research provide means for reducing the likelihood of discriminatory decision-making while using DA tools' recommendations without compromising individuals' privacy as individual sensitive data (e.g., sex, race) is not revealed to users but is only presented in an aggregated form. Moreover, this research focuses on inducing an ethical behaviour in DA users (i.e., rejecting a DA recommendation they find potentially discriminatory) by enhancing the moral intensity of the situation. This study thereby not only contributes to the DA literature but also to the social psychology literature by extending the theory of moral intensity to a new context. This study also promises strong implications for practitioners who are ultimately responsible for any discriminatory decisions they may make relying on DA recommendations. The outcomes of this research are especially important given the complete reliance of organizational decision makers on DA tools to deal with complexities and uncertainties in the era of big data. As such, the outcome of this study is beneficial for those well-intentioned practitioners who seek to lower the incidents of discriminatory decision-making in their corporations. Finally, this study also benefits DA developers who seek to gain a competitive advantage by making their tools more attractive to organizational users.

Limitations

Notwithstanding its significance and contributions, several limitations exist for this study. It is noteworthy that more than one definition for fairness exists in the literature. In this study we adopt the group fairness (aka statistical parity) definition. Future research is warranted to devise decisional guidance based on other notions of fairness (e.g., individual fairness). Moreover, the participants for this study will be selected from North American data analytics users. Thus further research should be conducted prior to generalizing the findings of this research to other cultures.

Future Plans for the Study

In the subsequent phases of this research, the proposed research model will be empirically validated. To that end, the experimental DA tool as discussed in the methodology section, will be developed. Moreover, ethics approval from the ethics research board at the first authors' university will be secured. Subsequently, prior to data collection, a pilot study with 40 middle managers will be conducted. We hope to have the preliminary results of this study ready for presentation at ICIS2019.

References

Barocas, S., and Selbst, A. D. 2016. "Big Data's Disparate Impact," *California Law Review* (104), pp. 671–732.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale, NJ: Lawrence

Earlbaum Associates.

Côrte-Real, N., Oliveira, T., and Ruivo, P. 2017. "Assessing Business Value of Big Data Analytics in European Firms," *Journal of Business Research* (70:1), pp. 379–390.

Courtland, R. 2018. "Bias Detectives: The Researchers Striving to Make Algorithms Fair.," *Nature* (558:7710), pp. 357–360.

Craft, J. L. 2013. "A Review of the Empirical Ethical Decision-Making Literature: 2004–2011," *Journal of Business Ethics* (117:2), pp. 221–259.

Custers, B. 2013. "Data Dilemmas in the Information Society: Introduction and Overview," in *Discrimination and Privacy in the Information Society*, B. Custers, T. Calders, B. Schermer, and T. Zarsky (eds.), Berlin Heidelberg: Springer-Verlag, pp. 3–26.

Davenport, R. B. T. H., and Bean, R. 2018. "Big Companies Are Embracing Analytics, But Most Still Don't Have a Data-Driven Culture," *Harvard Business Review*.

Dopplick, R. 2017. "New Statement on Algorithmic Transparency and Accountability by ACM U.S. Public Policy Council." (https://techpolicy.acm.org/2017/01/new-statement-on-algorithmic-transparency-and-accountability-by-acm-u-s-public-policy-council/).

Douglas, P. C., Davidson, R. A., and Schwartz, B. N. 2001. "The Effect of Organizational Culture and Ethical Orientation on Accountants' Ethical Judgments," *Journal of Business Ethics* (34:2), pp. 101–121.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. 2012. "Fairness through Awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, Cambridge, Massachusetts: ACM, pp. 214–226.

Ebrahimi, S., and Hassanein, K. 2019. "Can the Use of Data Analytics Tools Lead to Discriminatory Decisions?," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, Grand Wailea, Maui: IEEE, pp. 6739–6748.

Executive office of the [US] president. 2014. "Big Data: Seizing Opportunities, Preserving Values," White House, Executive Office of the President.

Federal Trade Commission. 2016. "Big Data: A Tool for Inclusion or Exclusion."

Ferrell, O. C., Gresham, L. G., and Fraedrich, J. 1989. "A Synthesis of Ethical Decision Models for Marketing," *Journal of Macromarketing* (9:2), pp. 55–64.

Freitas, A. A. 2014. "Comprehensible Classification Models: A Position Paper," *ACM SIGKDD Explorations Newsletter* (15:1), ACM, pp. 1–10.

Frey, B. F. 2000. "The Impact of Moral Intensity on Decision Making in a Business Context," *Journal of Business Ethics* (26:3), pp. 181–195.

Gartner. 2015. "Gartner Says, By 2018, Half of Business Ethics Violations Will Occur Through Improper Use of Big Data Analytics." (https://www.gartner.com/en/newsroom/press-releases/2015-10-07-gartner-says-by-2018-half-of-business-ethics-violations-will-occur-through-improper-use-of-big-data-analytics, accessed January 5, 2019).

Gates, S. W., Perry, V. G., and Zorn, P. M. 2002. "Automated Underwriting in Mortgage Lending: Good News for the Underserved?," *Housing Policy Debate* (13:2), pp. 369–391.

Ghasemaghaei, M., Ebrahimi, S., and Hassanein, K. 2018. "Data Analytics Competency for Improving Firm Decision Making Performance," *Journal of Strategic Information Systems* (27:1), pp. 101–113.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. 2018. "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys (CSUR)* (51:5), ACM, pp. 1–43.

Haines, R., Street, M. D., and Haines, D. 2008. "The Influence of Perceived Importance of an Ethical Issue on Moral Judgment, Moral Obligation, and Moral Intent," *Journal of Business Ethics* (81:2), Springer, pp. 387–399.

Henelius, A., Puolamäki, K., Boström, H., Asker, L., and Papapetrou, P. 2014. "A Peek into the Black Box: Exploring Classifiers by Randomization," *Data Mining and Knowledge Discovery* (28:5–6), Springer, pp. 1503–1529.

Holstein, K., Vaughan, J. W., Daumé III, H., Dudík, M., and Wallach, H. 2018. "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?," *ArXiv Preprint ArXiv:1812.05239*.

Hunt, S. D., and Vitell, S. 1986. "A General Theory of Marketing Ethics," *Journal of Macromarketing* (6:1), pp. 5–16.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. 2011. "An Empirical Evaluation of the Comprehensibility of Decision Table, Tree and Rule Based Predictive Models," *Decision Support*

Systems (51:1), pp. 141–154.

Johansson, U., and Niklasson, L. 2009. "Evolving Decision Trees Using Oracle Guides," in *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, IEEE.

Johnson, J. A. 2014. "The Ethics of Big Data in Higher Education," *International Review of Information Ethics* (21:1), pp. 3–10.

Jones, T. M. 1991. "Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model," *Academy of Management Review* (16:2), pp. 366–395.

Lipton, Z. C. 2016. "The Mythos of Model Interpretability," *ArXiv Preprint ArXiv:1606.03490* (558:7710), pp. 357–360.

Lipton, Z., McAuley, J., and Chouldechova, A. 2018. "Does Mitigating ML's Impact Disparity Require Treatment Disparity?," in *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, Montréal, Canada.

Loe, T. W., and Weeks, W. A. 2000. "An Experimental Investigation of Efforts to Improve Sales Students' Moral Reasoning," *Journal of Personal Selling & Sales Management* (20:4), pp. 243–251.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. 2013. "Accurate Intelligible Models with Pairwise Interactions," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 623–631.

Lyon, D. 2003. *Surveillance as Social Sorting: Privacy, Risk, and Digital Discrimination*, London: Routledge.

May, D. R., and Pauli, K. P. 2002. "The Role of Moral Intensity in Ethical Decision Making: A Review and Investigation of Moral Recognition, Evaluation, and Intention," *Business & Society* (41:1), Sage Publications Sage CA: Thousand Oaks, CA, pp. 84–117.

McDonald, C. 2016. "Is There a Sexist Data Crisis?," *BBC News*, United Kingdom. (http://www.bbc.com/news/magazine-36314061).

Newell, S., and Marabelli, M. 2015. "Strategic Opportunities (and Challenges) of Algorithmic Decision-Making: A Call for Action on the Long-Term Societal Effects of 'Datification,'" *Journal of Strategic Information Systems* (24:1), pp. 3–14.

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, NY: Crown Publishing Group.

Pedreschi, D., Ruggieri, S., and Turini, F. 2008. "Discrimination-Aware Data Mining," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 560–568.

Rest, J. R. 1986. *Moral Development: Advances in Research and Theory*, New York: Praeger.

Reynolds, S. J. 2006. "Moral Awareness and Ethical Predispositions: Investigating the Role of Individual Differences in the Recognition of Moral Issues," *Journal of Applied Psychology* (91:1), pp. 233–243.

Schermer, B. W. 2011. "The Limits of Privacy in Automated Profiling and Data Mining," *Computer Law & Security Review* (27:1), pp. 45–52.

Silver, M. S. 1991. "Decisional Guidance for Computer-Based Decision Support," *MIS Quarterly* (15:1), pp. 105–122.

Singhapakdi, A., Vitell, S. J., and Kraft, K. L. 1996. "Moral Intensity and Ethical Decision-Making of Marketing Professionals," *Journal of Business Research* (36:3), pp. 245–255.

Tene, O., and Polonetsky, J. 2013. "Judged by the Tin Man: Individual Rights in the Age of Big Data," *J. on Telecomm. & High Tech. L.* (11), HeinOnline, pp. 351–368.

Trevino, L. K., and Brown, M. E. 2004. "Managing to Be Ethical: Debunking Five Business Ethics Myths," *The Academy of Management Executive* (18:2), pp. 69–83.

Vidovic, M. M.-C., Görnitz, N., Müller, K.-R., and Kloft, M. 2016. "Feature Importance Measure for Non-Linear Learning Algorithms," *ArXiv Preprint ArXiv:1611.07567*.

Watley, L. D., and May, D. R. 2004. "Enhancing Moral Intensity: The Roles of Personal and Consequential Information in Ethical Decision-Making," *Journal of Business Ethics* (50:2), pp. 105–126.

Yinger, J. 1998. "Evidence on Discrimination in Consumer Markets," *Journal of Economic Perspectives* (12:2), pp. 23–40.

Zliobaite, I. 2015. "Measuring Discrimination in Algorithmic Decision Making," *Data Mining and Knowledge Discovery* (31:4), pp. 1060–1089.

Žliobaitė, I., and Custers, B. 2016. "Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models," *Artificial Intelligence and Law* (24:2), pp. 183–201.