DEPARTMENT OF ENGINEERING MANAGEMENT

FAT Flow: a Data Science Ethics Framework

David Martens

UNIVERSITY OF ANTWERP
Faculty of Business and Economics
City Campus
Prinsstraat 13, B.226
B-2000 Antwerp
Tel. +32 (0)3 265 40 32
www.uantwerpen.be

FACULTY OF BUSINESS AND ECONOMICS

DEPARTMENT OF ENGINEERING MANAGEMENT

FAT Flow: a Data Science Ethics Framework

David Martens

RESEARCH PAPER 2020-004
JULY 2020

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.226
phone: (32) 3 265 40 32
e-mail: joeri.nys@uantwerpen.be

D/2020/1169/004

FAT Flow: a Data Science Ethics Framework

David Martens

Abstract

The impact of data science in our society is undeniable, both in generating cost-efficiencies and in offering better services and products. As data science often involves making decisions for humans, from deciding on whether to give credit or not, to a self-driving car deciding how to drive, the manner in which companies conduct data science will have large implications for humans (including their customers) too. The interest in the ethical aspects of data science is growing, and have become an increased focus point in both research and practice. Data science ethics looks at what is right and what is wrong when conducting data science. This goes beyond what is legal, and considers aspects as privacy, discrimination against sensitive groups, the ability to explain predictions, and accountability. This paper provides a framework in which concepts, techniques and cautionary tales related to data science ethics can be placed. Companies can use the framework to think about the ethical aspects of their own data science projects, be it at the start of a project or to review current data science practices. It provides guidance on what the important concepts are, how techniques can be used to improve on their data science, and what cautionary tales exist in domains that might be similar to their own. The FAT Flow framework looks at three dimensions: (1) the role of the humans involved in the project, being data subject, data scientist, manager and model applicant; (2) the stage of the data science project: from data gathering to model deployment; and (3) the FAT evaluation criteria: Fair, Accountable and Transparent.

1 Introduction

Data science can be defined as *"a set of fundamental principles that support and guide the principled extraction of information and knowledge from data"* [16], and has become a mature technology that is being used in many businesses. The increased digitization leads to an ever growing mass of data that is being generated: companies now store each contact with their customer, be it through helpdesk calls, emails, payments or website visits. Tremendous value in terms of knowledge on customers and prospects lies within these massive datasets, which can be uncovered through data science practices. Data science, and predictive modeling in particular, has therefore taken up a crucial role in many organisations, such as banks, insurance companies, telco players, media companies, and tax administrations [15]. For many companies, data science is even the core of a company's added value. For example, it is hard to imagine the products and services of large tech companies Facebook, Google, Amazon, Netflix or Spotify without data science.

Additionally, recent advances in Artificial intelligence (AI) have led to "super-human" results in domains as speech and image recognition, where the models reach better results than humans [10]. These improvements are spurred mainly by deep learning (artificial neural networks) and the availability of massive image, textual and behavioural data. Though deep learning has led to positive economical and societal implications, they also suffer from leading to very complex models. The lack of understanding of how decisions are being made has important ethical challenges, from not discriminating against sensitive groups to simply being transparent.

You might wonder, why is this important for businesses, and why should I care about data science ethics? Although being ethical has been put forward as a life goal in itself, there are just as important societal and business reasons. First of all, there are huge reputational risks related to data science ethics. Not only large companies risk their reputation, also startups and smaller companies should care: they often rely even more on new data science products and services. Not getting the ethical aspects right can stop their growth (or even business) altogether or could lead them into trouble during due diligence or investment negotiations.

A second reason to care about data science ethics is the actual value it can bring. Ethical thinking can lead to improvements in your data quality and data science models, with potentially more accurate predictions, better user acceptance of the data science models, or a better brand image. This thereby could improve the business value, through more revenue, higher profits or lower costs.

Thirdly, we've reached an age where society expects business leaders and data scientists alike to be ethically responsible. The power of data science has become clear to all: from the data subjects to business leaders. The cases that regularly appear in the media, from privacy-related discussions to real data science cases showing to have unfairly treated certain sensitive groups, have educated the public. Members of generation Z, born between 1995 and 2010, specifically care much about social justice and ethics [8].

This resulted in an increased attention and acknowledgement of the importance of ethical science with much attention from main stream press on the cautionary tales, and ample new research being proposed, from new approaches on how to explain prediction models [11, 17], to studies discussing on whether to include ethical preferences in self-driving cars [3]. However, a general framework to discuss these concepts, techniques and cautionary tales is largely missing.

2 FAT Flow

We put forward an ethical data science framework for business (see Figure 1), using three dimensions:

1. Role of the human
2. Stage in the data science process
3. Evaluation criterion

The first dimension of the framework considers the four roles that exist when discussing data science ethics: data subject, data scientist, manager and model applicant. The second dimension follows the five stages of a data science project relevant to ethics: from data gathering to model deployment. The third dimension entails three properties that have emerged in the community: fair, accountable and transparent. This framework is general enough to include all ethical aspects of data science, and provides the flexibility to include novel techniques and cases. At the same time it provides a scientific instrument to follow with guidelines on concepts, techniques and cautionary tales.

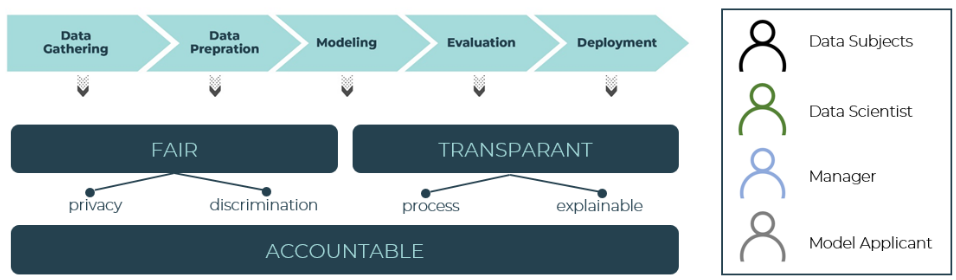

Figure 1: FAT Flow Data Science framework, using three dimensions: (1) role, (2) modeling stage, and (3) evaluation criterion.

2.1 FAT: Fair, Accountable and Transparent

Data science ethics can be evaluated using three criteria: fairness, transparency and accountability. The first two criteria evaluate ethical concepts, such as privacy, discrimination and explainability. Accountability is about the effective and verifiable implementation of these concepts. These criteria have become key, and the focus of conferences [7] and research groups on ethics in AI, for example at Microsoft [12].

Fair The Oxford dictionary defines fair as *"Treating people equally without favouritism or discrimination."*[1] In the context of data science the common meaning is:

Fair (a): "Not discriminating against sensitive groups."

[1] https://en.oxforddictionaries.com/definition/fair

This already shows the subjectiveness of ethics: what are sensitive groups? Typically discriminating against ethnicity, gender or religion is considered unfair, making for three important variables to test when considering fairness. But this is not always true and depends on the application: in medical diagnostics for example, race and gender can be important, scientifically motivated variables.

Next to the application-dependency, what is considered sensitive also varies over time and regions. Not discriminating against women for example is a relatively recent accepted standard. In the United States, the legal right for women to vote only was established nationally in 1920. Only in 1976 did West Point, a U.S. military academy, admit its first female cadets [4]. Slavery in the U.S., primarily of black people, was only prohibited nationwide in 1865 by the 13th Amendment, while the right for black people to vote was included in the 19th Amendment in 1920. Although now, most of us consider these to be self-evident, we too will be considered victims of our time in the future.

Also regional aspects play a role. The difference in U.S. versus European (GDPR) laws already demonstrates this. A 2018 study at MIT investigated the human preferences when dealing with the trolley problem [3]: if a driver could not break and had to choose between hitting and killing two adults versus one baby, what would be preferred? By putting such dilemmas forward to thousands of persons, considering different types of subjects (babies, children, cats, dogs, elderly, executives, homeless people, etc.) a ranking of ethical preference can be established. The study found that, in terms of sparing, children were preferred over adults, adults over elder people, and interestingly dogs over criminals. Important regional differences were also found in this study: the preference to spare young over older characters was much less pronounced in Eastern countries, such as Japan and Taiwan, as compared to Western countries. Once more, given the subjectiveness of ethics, deciding on what data science ethics practices to implement in your business is an endeavor that every business has to undertake.

Next to the discrimination aspect of fairness, there is also the privacy aspect. The fair use of personal data entails that the privacy of the data subject is respected. The second definition in Oxford Dictionary is: *"Without cheating or trying to achieve unjust advantage."* In the context of data science, we take over this definition, but specifically look at privacy: **Fair (b) - adverb: "Without cheating or trying to achieve unjust advantage w.r.t. privacy."**

Privacy is recognised as a human right. The United Nations' 1984 Universal Declaration of Human Rights states: *"No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation."*[2] Much has been written about privacy, from Orwell's book 1984 [14] to the European General Data Protection Regulation (GDPR[3]). In other words, fairness also relates to not cheating w.r.t. the respect for privacy, not observing or disturbing people when they don't want to be.

This too might be self-evident, but once more this is not a boolean but continuous criterion. Some applications allow for more latitude: for detecting fraud, solving crimes or medical diagnosis generally more private data can be used (of course still with limits) as compared to targeted advertising or recommending music.

Transparent The transparency criterion is arguably the most important one, as it has implications for both accountability and fairness. Although one often limits transparency to explaining the decisions made by the data science model, transparency goes much further, covering all stages of a data science project. The level of transparency required is also different depending on the different entities to which transparency is provided: the manager of the organisation will want full transparency on the process, which the data subject might not get (as not to reveal company secrets). On the other hand, the data scientist will want full transparency on all algorithmic steps taken in previous modeling exercises, the manager is likely not that interested in knowing what hyperparameter grid was used in cross-validating the

[2] `https://www.un.org/en/universal-declaration-human-rights/`

[3] `https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1532348683434&uri=CELEX:02016R0679-20160504`

regularisation parameter of a regularised logistic model (the very good and very bad managers might want to know).

Transparency also takes an important role in accountability and fairness. In order to know whether a model is fair, transparency is needed in the data used (privacy) or the evaluation of the model for different sensitive groups (discrimination). Making the data science process and predictions transparent, the data science model might even improve: unnecessary data sources might be detected, data biases which led to decreased performance can be removed or explanations of misclassifications can reveal insights as to how to improve the data quality or model.

Transparency does not imply you need to tell every detail to end-users. Doing so could infringe on company secretes or even privacy of data subjects. What exact data science technique you are using could be part of your "secret sauce", just as you don't need to provide the model applicant the exact prediction score of your model. This illustrates that transparency is not a boolean criterion, but more a continuous one.

Accountable Accountability might just be the least well-defined FAT evaluation criterion, mainly because it is a very broad concept, often used in different contexts. The GDPR directive and supporting documents are very useful to help us to understand why accountability is needed (see Opinion 3/2010 on the principle of accountability 2010 [2]). Accountability is all about moving from theory to practice: policies in itself are not enough. A recitation of a company's policy with regards to data science ethics can be very lofty and impressive, but there needs to be obligations to ensure that these are put into practice. Accountability is intended as to strengthen the responsibility of a company and its people, leading us to another definition (illustrated in Figure 2):

Accountability: "Obligation to (1) implement appropriate and effective measures to ensure that principles are complied with, (2) demonstrate compliance of the measures upon request, and (3) recognize potential negative consequences."

The first term is about taking appropriate measures to ensure that data

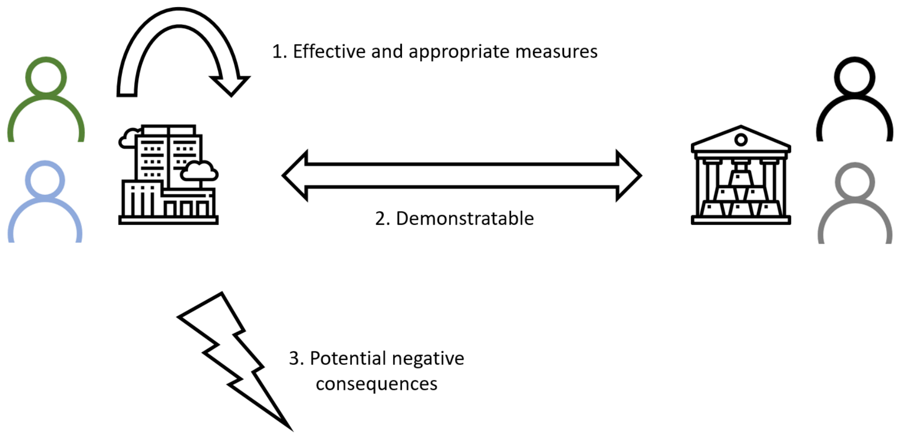

Figure 2: Three facets of accountability.

science practices are conducted according to the set policies and relevant regulations. The second term of accountable is being able to demonstrate that these measures are implemented, that these do indeed lead to compliance with the set forth policies and regulations. The ability to demonstrate what measures have been taken is important in this account, and involves a description of tools and training that are put in place, as well as a system for oversight and verification. The third aspect of accountability is having to face possible consequences for not complying with the obligations. When something does go wrong, accountability will be about this aspect mostly, and implies a liability. Financial negative consequences can be in the form of government fines or compensations awarded by judges. As motivated by Bovens [5], this possibility of sanctions (and not necessarily the actual imposition of sanctions) makes the difference between non-committal provision of information and actually being held to account.

2.2 The roles of people in data science

To discuss the different roles that humans can take in a data science project, let's consider the issue of being able to explain the decisions made by some data-driven credit scoring model (an issue that received renewed attention in November 2019 with the Apple Card [18]). This model will predict if a loan applicant (hence the model applicant or decision subject) is able to repay his or her loan, and hence whether to grant credit or not. First of all, a rejected loan applicant will want to know why the application is rejected (even more: it often is a legal requirement). Is it because the income is too low? Because the ratio of loan to value of the mortgage is too low? Is there some issue in the credit history? Is it a combination of factors? Simply stating "Computer says so..." is not enough [19]. Secondly, before a credit scoring model is being deployed, the manager of the bank will want some insight into how the model works. Simply deploying some black-box, incomprehensible model will not be allowed, even if it is accurate on an out-of-time test set. For the manager it is less important to know why a single customer is accepted or rejected, the manager will want the general idea of how it works. Thirdly, we have the data scientist. He or she will be eager to know why the model is making certain wrong decisions. Is it because not enough data was available on that specific group? Is it because of a data quality issue? Knowing this can help to improve the data science model. Finally, note that in this case the data subjects are different from the loan applicants: the data subjects are all customers who previously took out a loan. On these persons the bank knows the actual target variable: did this customer repay the loan or not. The loan applicant is the one applying for a loan, being scored by a model.

3 Stages in the Data Science Process

3.1 Ethical Data Gathering

Fair

The data gathering process is a first important stage, as illustrated in Figure 3. This needs to be fair to the data subjects and model applicants, in terms of privacy and discrimination against sensitive groups. Note that the roles of data subject and model applicant can be different in this case, as different data can be used on each in the data gathering process. Let's

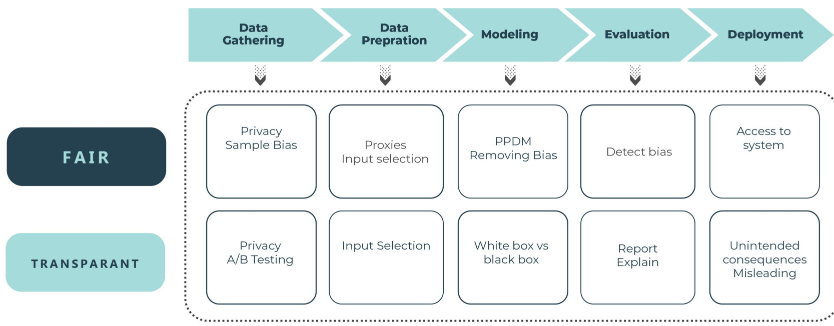

Figure 3: Concepts of Fairness and Transparancy within the FAT Flow framework.

revisit the credit scoring case, where data of the customers who ever have been granted a loan is kept, as to build credit scoring models with as target variable: did this customer repay the loan or not. These persons become a customer at the start of the loan (if they weren't customer before) and sign a contract which discusses what data is used, for what purposes, how long etc. Loan applicants on the other hand might well not be customers. They just provide their details (such as income, age, profession, etc.) to be scored and obtain an approval and/or interest rate. But is this personal data only used to score the applicant, or is this data kept for other data science modeling? If the latter, it is important to also obtain an informed consent. The fairness should not only cover privacy, also fairness towards sensitive groups should already be considered. Sample bias, where certain sensitive groups are over- or underrepresented, can lead to negative discrimination of the model built on this data.

Transparent

The transparency of the data gathering process also needs to consider the privacy of the data subjects and the model applicants. This includes informed consent: is the data subject and model applicant informed about the data gathering and is consent provided? There should also be transparency in what data is gathered, for what purpose and for how long. A specific setting in which data is gathered across all model applicants is A/B testing. The fact that one is part of an experiment can be sensitive and requires transparency. A/B testing is widespread and it is important to make model applicants aware of the fact that they are part of this A/B test in case the experiment has a potential impact on the physical or mental well-being of the participants. Also the data scientist and manager require transparency, as to understand how the data is gathered. The data scientist needs to know to ensure data quality and perform suitable data preparation and modeling, while the manager is the one signing off on the process, so he or she surely wants to know how this all occurs.

3.2 Ethical Data Preparation

Fair

Not including a sensitive variable such as race does not necessarily imply that you are not discriminating against racially sensitive groups. Any variable that is strongly correlated with race, so called proxies, should also be removed. So if a bank is not allowed to use gender in its credit scoring process, is it allowed to use shoe size? A banker will likely not ask you for your shoe size (beware of bankers that do) but he or she will ask you for your address, and this can be a proxy for race.

Next to this fairness issue at the feature dimension (on discrimination), fairness also relates to the instance dimension (on privacy). When you claim to have anonymized your data, are you sure that the data can not be de-identified? Anonymizing your data can be very useful, as you can continue to work on it for aggregated analytics or to use the data without any personal information. The latter use case is also very useful to foster academic research by making datasets available, as often is done through systems such as Kaggle.

Transparent

The transparency aspects of the data preparation stage concerns the open communication of the aforementioned issues on data preparation. If the data is pseudonymized instead of anonymized, do the data subjects and model applicants know their data is still being kept? Is there sufficient effort being made to ensure the data is truly anonymous? The input selection procedure should similarly be well-documented, and be clear to (future) data scientists and managers, as to ensure that the potential ethical motivations for removing (or keeping) variables are well understood. Finally, the definition of the target variable should be made transparent to all, ensuring that everyone agrees on the definition and practical scenarios that could have impacted the measurement thereof.

3.3 Ethical Modeling

Fair

There are ways to include privacy within the data science modeling. Consider the case where you are the data scientist asked to build a variety of predictive models on datasets gathered from several data providers. First of all, personal data on the data subjects is not required for you as these won't be a part of your input set. The name of a person, exact date of birth, or social security number are not (or should not be) relevant in your exercise[4]. In other words, personally identifiable information should be protected and made invisible. Further use of Privacy Preserving Data Mining (PPDM) techniques when working with personal data is recommended in this stage. Secondly, we need to avoid that sensitive variables can be predicted from the datasets. Perhaps political preference is not explicitly included in the dataset, but could be easily predicted. In this case, we want to protect

[4]In some specific cases this might be relevant, for example in fraud detection to find patterns in the stated date of birth.

against patterns that should not be learned, for example by cloaking certain data [6].

Transparent

The arsenal of modeling techniques available to a data scientist is very broad. From the 'workhorse' logistic regression to advanced deep learning algorithms. One dimension on which these algorithms differ is how comprehensible the generated models are. We typically speak of white-box versus black-box models. As so often, the comprehensibility is a continuum, where rule or tree-based models are often regarded very comprehensible, while non-linear techniques are considered much less comprehensible. The choice of number of inputs (or regularization) and type of algorithm therefore have an important impact on the transparency of the model. The actual generation of explanations for a decision is closely related to this, commonly known as part of 'explainable AI' [1].

3.4 Ethical Model Evaluation

Fair

In the evaluation stage, the data science model is being evaluated on the aforementioned fairness criteria related to privacy and discrimination against sensitive groups. When we want to determine whether the model discriminates or not against sensitive groups, a paradox is revealed: the need for the sensitive attribute. To evaluate whether a model does not discriminate on race, the race variable would need to be available to assess its impact on the predictions. If this is available, several techniques have been put forward to evaluate the fairness, see e.g. [9]. But when we don't have the sensitive variable, the situation becomes much more difficult.

Transparent

When evaluating the model, transparency has several important roles. The first is on doing proper model evaluation. Yet, it is so important that it even has an ethical implication. Different performance metrics can lead to the different decisions on the model. Consider the example of predicting stock prices. A modeler stating 90% accuracy should lead to questioning remarks. One easy explanation could be that the evaluation was done on 10 days only, in which period the stock market increased on all but one day. Simply always predicting an increase leads to a 90% accuracy. A good model though? No. A good data scientist would not make such a mistake. But the intuitiveness of accuracy can sometimes lead to the decision to also report the "easy" accuracy metric. Only do so when also reporting the accuracy of good baseline models.

Explaining the decisions and predictions that are made is another important area related to transparency in the evaluation stage. Can you explain in non-technical terms to the model applicant why his or her credit was denied? Can you explain why some website was classified as containing hate speech and hence not allowing ads to run there? Similarly a data scientist would like to know why certain misclassification were made, and the manager will want to know how decisions are being made. Such explanations of predictions have become an important area in data science [1], because of the increased complexity of the data science technique and data, but also because of the increased use of data science and awareness in the business.

3.5 Ethical Model Deployment

Fair

In deployment of data science systems, decisions are sometimes made on who gets access to the data science model. Who are you considering to score. In credit scoring, banks are often limited to those who have previous credit history or at least a checking account. It's important to be aware of who exactly has access to the system and who doesn't. Censorship is a specific, deliberate case of providing limited access to your data science system. A second fairness concept to consider at deployment is the ability to 'correct' automated decisions. However, one should try to keep this limited and each overruling decision needs to be properly motivated. Also for the model applicant this is important, as to ensure a coherent and fair policy, not favoring persons because of unethical reasons (e.g. 'friends with' or 'quid pro quo').

Transparent

The way a system behaves in production versus during development can be quite different. Unintended consequences are by definition unintended, yet can already be thought through during development. One should be transparent on what the unintended consequences might be and how these could be mitigated.

Misleading people through data science models is an ethically very questionable practice. Deep Fake is such a technology [13], based on deep learning, to create real-looking yet fake videos. Transparency in disclosing clearly when fake footage is being distributed is required.

4 Conclusion

Data science ethics is all about what is right and what is wrong when using data science. As data science practices are entering the life of many organisations, with an impact on an increasing number of persons, getting the ethics right is becoming a key aspect of data science. Whether you are a data scientist, manager or simply a data subject, you should think about the potential impact and misuse of such technology.

One should note that the proposed FAT framework is not intended to be complete, as any framework or checklist is bound to become outdated quickly (and hence needs to be updated regularly): new data sources, techniques, applications and ethical considerations are continuously being proposed. Rather, data scientists and businesses can use the FAT Flow framework at the start of any data science project, and for reviewing existing data science projects.

Finally, the framework shows that ethics is not an aftermath to a data science project, it is intertwined with it. So try to consider ethics from the start of your data science project. Furthermore, don't shy away from including an ethical section in your reports or even your emails when discussing data science projects.

References

[1] A. B. Arrieta, N. Diaz-Rodriguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019.

[2] Article 29 Data Protection Working Party. Opinion 3/2010 on the principle of accountability. Technical Report WP173 00062/10/EN, European Commission, 2010.

[3] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. The moral machine experiment. *Nature*, 563, 11 2018.

[4] C. Barkalow. *In the Men's House*. Poseidon Press, 1990.

[5] M. Bovens. Analysing and assessing public accountability. a conceptual framework. *CONNEX and EUROGOV networks, European Governance Papers (EUROGOV)*, 01 2006.

[6] D. Chen, S. P. Fraiberger, R. Moakler, and F. J. Provost. Enhancing transparency and control when drawing data-driven inferences about individuals. *Big Data*, 5(3):197–212, 2017.

[7] FAT. Conference on fairness, accountability, and transparency. `https://facctconference.org/`, 2020.

[8] T. Francis and F. Hoefel. 'true gen': Generation z and its implications for companies. `https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/true-gen-generation-z-and-its-implications-for-companies`, November 12, 2018. Online, accessed July 13, 2020.

[9] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 16, page 21252126, New York, NY, USA, 2016. Association for Computing Machinery.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] D. Martens and F. Provost. Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–100, Mar. 2014.

[12] Microsoft. Microsoft. `https://www.microsoft.com/en-us/research/theme/fate/`. Online, accessed July 13, 2020.

[13] Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. Technical report, 04 2020.

[14] G. Orwell. *1984*. Tandem Library, centennial. edition, 1950.

[15] S. K. Prashant Gandhi and S. Ramaswamy. Which Industries Are the Most Digital (and Why)? *Harvard Business Review*, 2016.

[16] F. Provost and T. Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1), 2013.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[18] C. Thorbecke. New york probing apple card for alleged gender discrimination after viral tweet. `https://abcnews.go.com/US/york-probing-apple-card-alleged-gender-discrimination-viral/story?id=66910300`, November 11, 2019. Online, accessed July 13, 2020.

[19] United States. Equal credit opportunity act. https://www.law.cornell.edu/uscode/text/15/1691, 1974.