Ethical Artificial Intelligence

Author, Sray Agarwal

A Data Science Foundation White Paper

March 2020

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: admin@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

Discrimination and fairness in AI and machine learning: How to detect and mitigate bias in the financial services industry.

By Sray Agarwal

Are artificial intelligence (AI) algorithms biased? The answer is "yes," according to recent studies that suggest, for example, that the wide use of AI to assess standardized testing in the United States could yield unfavorable results for certain demographic groups [1]. Another study shows that AI plays a deciding role in hiring decisions, resulting in 72% of resumes in the United States never being viewed by a human [2]. Perhaps most famously (and disturbing), Google's photo recognition AI led to African-Americans being misidentified as primates [3].

The fact is, numerous incidents of rampant discrimination (intentionally or unintentionally) have surfaced – be it Google ads for showing higher-paying jobs more often to men than women [4]; Bank of America paying a $335 million fine for charging higher rates based on race [5]; a court intervening and ordering a company to stop using proxies for race to make hiring decisions [6]; or companies using name and place of birth to identify race or nationality, which led to significantly reduced customer satisfaction and ultimately damaged brand reputation [7]. AI is being used to help decide whether you get that job interview, how much you pay for insurance, and even what rating you get in your annual performance review – with historical bias and discrimination embedded in them. In yet another instance, an algorithm commonly used to identify the need for extra care for sick people, reduced the number of black patients identified for extra care by more than half [8].

Not to be ignored, natural language processing reinforces gender stereotypes by embedding the word "man" much closer to the word "programmer" than it does the word "woman." The latest example is the Apple Card, where it was found that the service gave lower limits to females based on gender discrimination [9].

All in all, from image recognition to hiring decisions to medical assistance to language translation to criminal justice to natural language processing, AI bias is found everywhere.

Of course, if a business differentiates against a person solely due to the color of their skin, it would be considered unethical and illegal. However, some machine-learning models do precisely that. Machine learning (ML), as the name implies, learns whatever it is taught. It's a ramification of what it is fed. It's a fallacy that ML doesn't have perspective; it has the same perspective of the data that was used to make it learn. In simple words, algorithms can echo prejudices that data explicitly or implicitly have.

How to Ensure Fairness in AI?

Any AI algorithm can have bias creep into it. The best way to avoid this is to proactively look for and identify bias in your AI, eradicate it and then alter your approach to ensure future algorithms are fairer. Posing the following questions can help check for systematic bias in your data:

- Are any particular groups suffering from systematic data error or ignorance?
- Have you intentionally or unintentionally ignored any group?
- Are all groups represented proportionally, e.g., when it comes to the protected feature of race, are all races being identified or merely one or two?
- Are there enough features to explain minority groups?

- Are you sure you aren't using or creating features that are tainted?
- Have you considered stereotyping features?
- Are your models apt for the underlined use case?
- Is your model accuracy similar for all groups?
- Are you sure that your predictions are not skewed toward certain groups?
- Are you optimizing all required metrics and not just those that suit the business?

The first step forward would be to understand the distribution of sensitive features (age, gender, color, race, nationality) to the outcome features (default, reject, approve, high rate, etc). This would mean defining metrics that can quantify interpretability and fairness for a model. These metrics, along with traditional metrics, eventually can be used to measure the suitability of an ethical yet performing ML model.

In order to ensure fairness of models, some key metrics need to be defined. While there are many possible fairness metrics, the most important are statistical parity, mean difference and disparate impact, which can be used to quantify and measure bias or discrimination. For instance, metrics such as statistical parity reveals if the data in question is discriminatory against an unprivileged class for a favorable outcome.

Taking an example of Loan data (bondora data set) where we wanted to predict the probability of someone defaulting on a loan, we were able to shortlist features that were discriminatory in nature. Using this risk data (risk calculated for more than 61,000 customers over 200 features), we see that six to eight sensitive features (marriage or otherwise, single or otherwise, house owner, mortgage, age group, ethnicity, primary language – English or otherwise, etc.) scream discrimination (Figure 1). For illustration, we chose a discriminatory feature that indicates a candidate's marital status, and by using a bias removal technique, such as re-weighing on this feature, we were able to reduce bias from 0.13 points to zero.

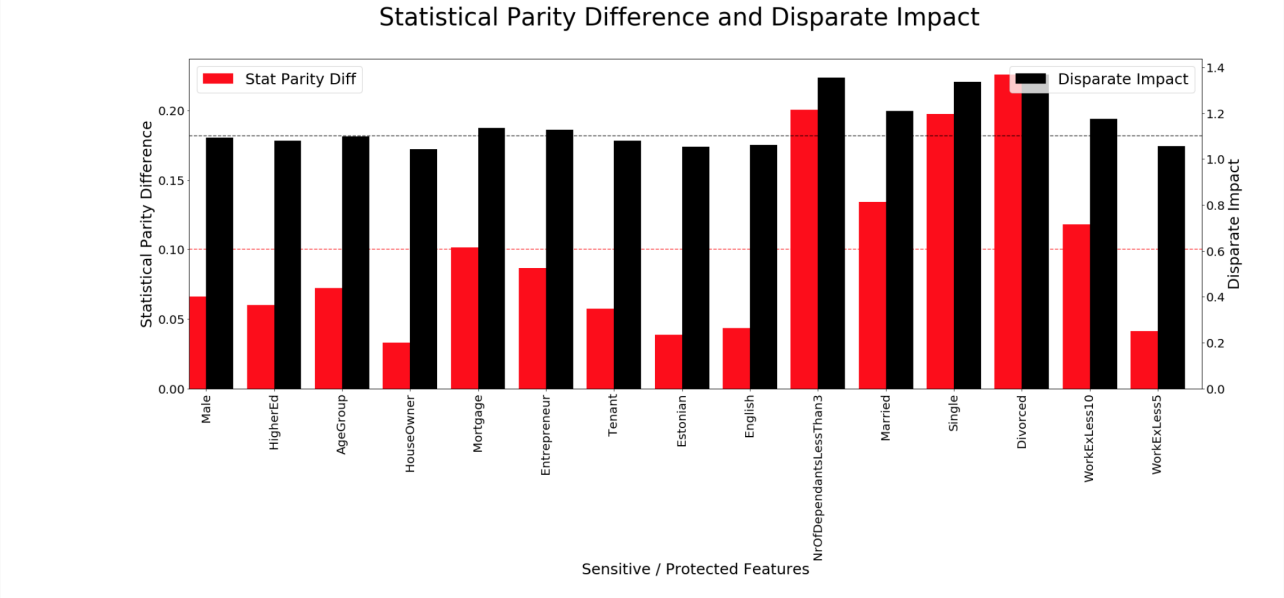

Figure 1: It's clearly visible that 5 sensitive features are having bias as per Statistical Parity Difference and Disparate Impact

The weights thus can be used as sample weights for most of the commonly used ML algorithms. These weights would penalize the cost function for any error with weights specified. In simple words, the algorithm won't give equal weights for all errors it makes, but it would have weighted errors, thus penalizing the maximum for unprivileged/disadvantageous and favorable combination.

Using this generated fairness-induced sample weights as a parameter to a logistic regression, we saw the overall accuracy difference between married and otherwise candidates reduced by 0.60%, while the overall accuracy increased by 0.13. The point to note here is that by using sample weights (generated using the reweighing technique), the model accuracy difference (or discrimination) between two groups of a sensitive class significantly decreased, thus ensuring similar performance for both the groups.

What is "Fairness"?

It is imperative to define "fairness" in the first place. The above example illustrates how one can detect bias and, to some extent, remove bias. However, in the above example, bias is defined as discriminatory of the overall accuracy of a model. But in the practical implementation of a use case, accuracy or business acceptability of a model is defined in many other ways. For example, in credit risk data, a bank would like to optimize and maximize its true positive rates and lower its false positive rates, thus producing huge cost/loss savings that may occur due to false positives.

In another scenario, a pharmaceutical company would like to increase its true positive rate and reduce its false negative rate to ensure maximum coverage of its treatment. For the uninitiated, a true positive is a situation when a model predicts an outcome as positive when its actual value in data was true; a false positive is a situation when a model predicts an outcome as positive when its actual value in data was false; and false negative is a situation when a model predicts an outcome as negative when its actual value in data was true.

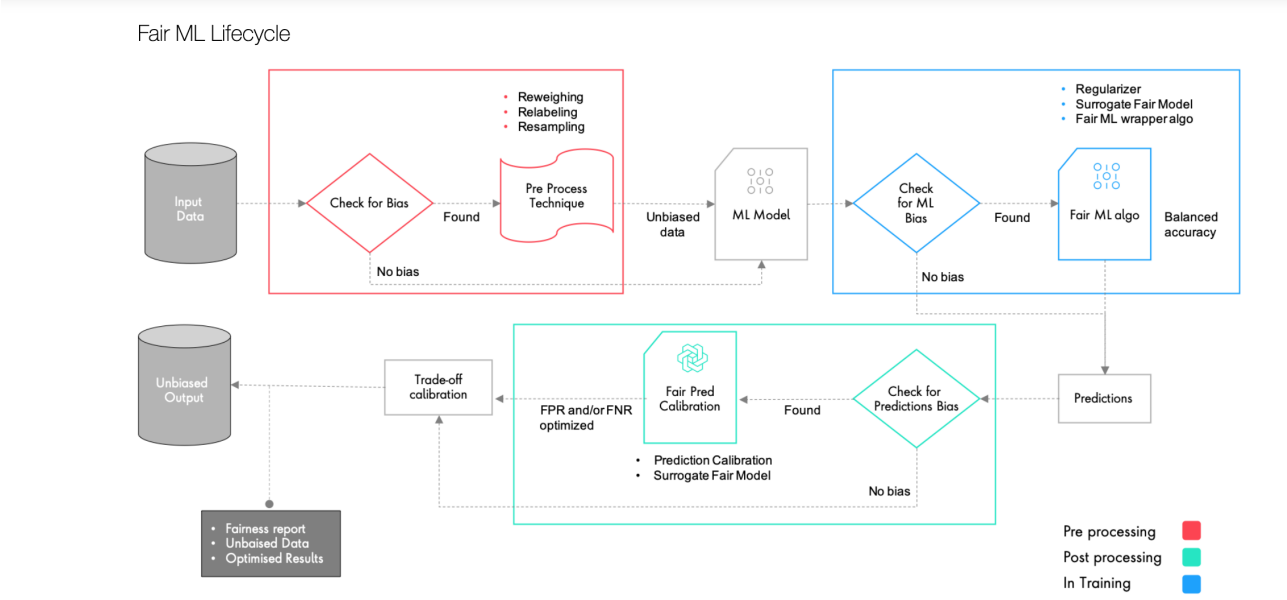

Figure 2: The practical implementation of a use case, accuracy or business acceptability of a

model is defined in many ways.

In order to define fairness metrics [10], it's important to understand what is the business asking for. In many cases (given the business objective and modeling approach), it makes sense to choose fairness metrics that put a lot of stress on true positive and false positive only. For instance, in the example we are discussing, it would make sense to optimize "average odds ratio" defined as an average of the difference in false positive rate (FPR) and true positive rate (TPR) for unprivileged and privileged groups and may also look at predictive equality (for FPR where both protected and unprotected groups have equal FPR) (Figure 3). In a few other cases, a secondary metric(s) like "equal opportunity" or "equalized odds" or "predictive parity" can be added to the list in case the primary fairness metrics fail to paint the complete picture. The choice of fairness metrics needs to be defined when the model performance (accuracy) metrics are being defined.

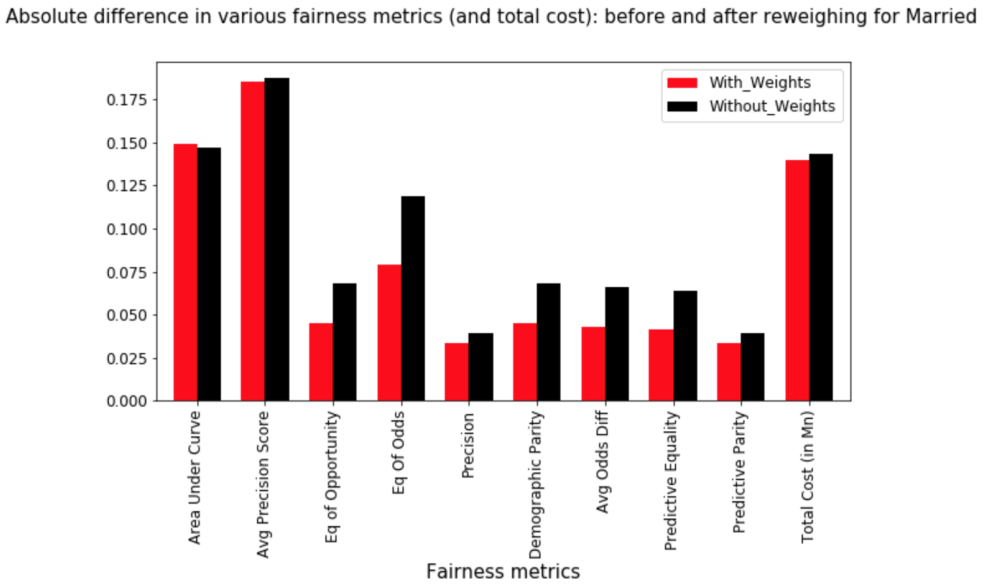

Figure 3: Absolute difference in various metrics (and total cost) before and after reweighting for "married."

As noted, bias detection is a half-baked solution without a remedy. Once you've detected bias in any and all stages of the data science lifecycle, it's worth looking at how to change your approach to algorithms to help ensure fairness in your overall modeling approach. You can do this in five ways:

1. **Use statistical calibration:** Leverage various statistical techniques to resample or reweigh data to reduce bias.
2. **Use a regularizer:** Add a fairness regularizer (a mathematical constraint to ensure fairness in the model) to existing ML algorithms.
3. **Use surrogate models:** Wrap a fair algorithm around baseline ML algorithms already in use.

4. **Use fair machine learning models:** Wrap a fair algorithm around baseline ML algorithms already in use.
5. **Calibrate the threshold:** Leverage various statistical techniques to resample or reweigh data to reduce bias.Calibrate the prediction probability threshold to maintain fair outcomes for all groups with protected and sensitive features.

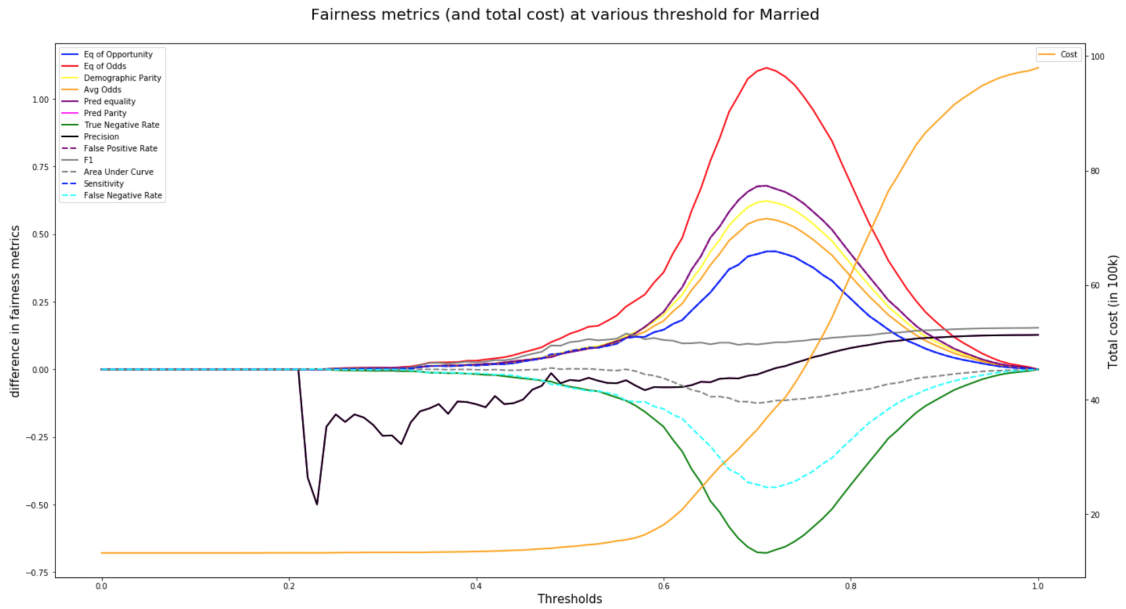

Figure 4: Fairness metrics (and total cost) of various thresholds for "married."

Sometimes it's not mandatory to reach an ideal value as defined for fairness metrics. For example, a model can be useful despite not reaching 100% accuracy; similarly, a model can be bias-free without reaching a zero value of discrimination. In many situations, it makes sense to reduce the bias by manifolds until the trade-off on accuracy (and monetary cost) is unobjectionable. This is much easier in all modeling that deals with thresholds (typical classification algorithms, Figure 4).

Things get trickier when it comes to accessing discrimination and inducing fairness in ML models where the outcome is not merely yes or no but a continuous number. In use cases where it is required to predict, for instance, the amount of the loan to be approved, the interest rate to be charged, credit limit to be sanctioned or probable income of a person, the approach can be a bit different. Here the ML model needs to compare the distribution of the outcome between a privileged and unprivileged group. The bias/discrimination would be defined as a stark difference between the two distributions in terms of mean, skewness and kurtosis.

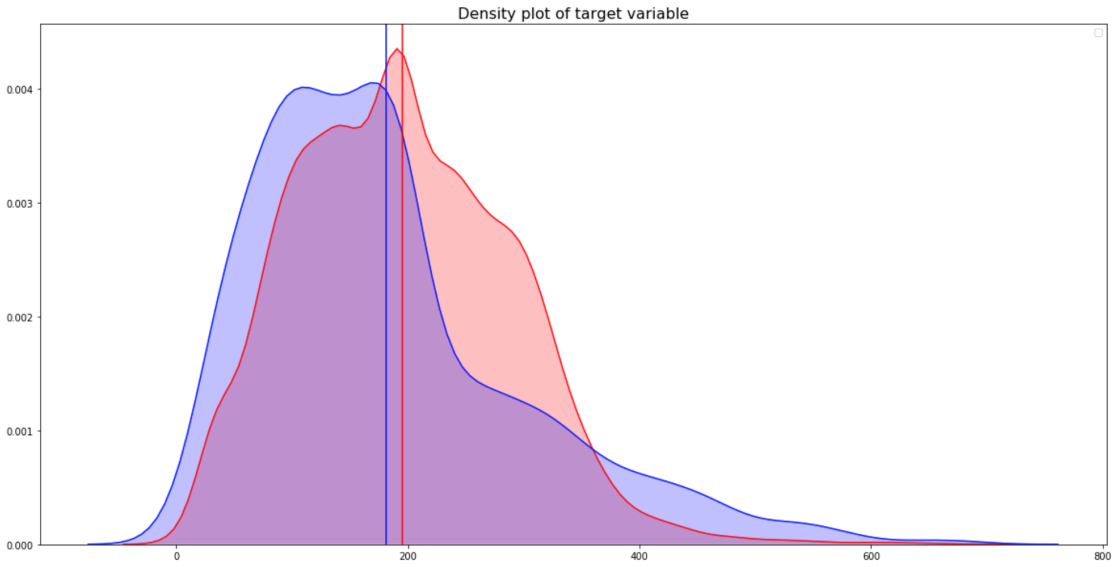

Figure 5A: The distribution of target value (probability of default) between two groups is quite varied in the original data (evident from the large difference in mean, skewness and kurtosis)

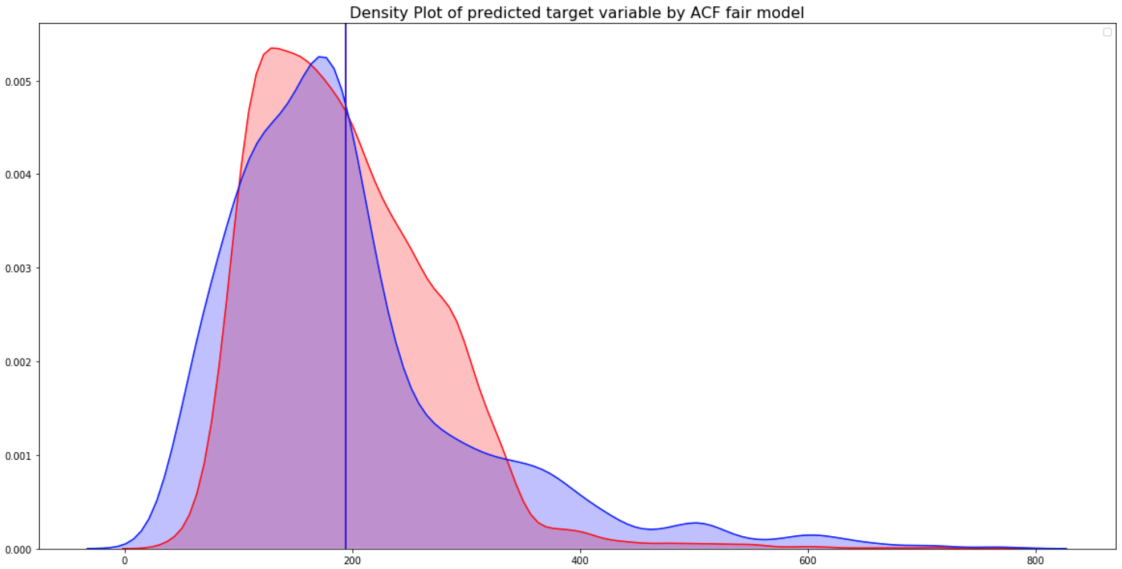

Figure 5B: ACF model show that the distribution of target value among the two groups is very similar and the difference in mean, skewness and kurtosis has reduced dramatically.

In such a case, it's important to ensure that the distribution of outcomes between two groups is as close

as possible. Furthermore, the concept of counterfactual [11] [12] methods may also be considered. For instance, noting the model error for original data and then comparing the error of the model for counterfactual (inverting protected features, e.g., changing all males to females, all married to unmarried and so forth) can give a very good idea about the impact of protected attributes on the model. If the difference between both cases is quite significant, approaches such as additive counterfactual fair models can be adopted. This approach negates the effects of protected features on the model without much trade-off on accuracy (and cost).

It is clear that there is no single definition of fairness. Even from our simple example, in a trivial logistic regression, different fairness metrics give seemingly different outcomes. Some would suggest our model is fair, some would suggest it is biased, and some would even suggest (arguably) that the model is biased against a particular group. Therefore, it is critical that the correct metric is chosen. The choice has to be sensible and well-reasoned, based on the expected impact that the metric has for the users of the model. Furthermore, the metrics selected for each use case must align with the stated principles of the organization and the model.

Selecting an appropriate metric allows the organization to justify to its stakeholders its choice of model and the outcomes that it produces. Not only does this build trust between the model and its users as it demonstrates that their personal interests were considered during model development, but it also allows feedback from the users – possibly raising misalignments between the stated principles of the company and the outcome of its models. This feedback should quickly alert a company to issues that users have with its models before they are alienated.

Sray Agarwal is a London-based data scientist for Publicis Sapient. He has more than 10 years of experience in predictive modeling, forecasting and advanced machine learning with a deep understanding of algorithms and advanced statistics. His current areas of interest are fair and explainable machine learning.

References:

1. Amorim, E., Can¸cado, M. and Veloso, A., 2018, “Automated essay scoring in the presence of biased ratings,” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (long papers), pp. 229-237, June, New Orleans: Association for Computational Linguistics.
2. O’Neil, G. M., 2017, “Hiring algorithms are not neutral,” Harvard Business Review, https://hbr.org/2016/12/hiring-algorithms-are-not-neutral.
3. Simonite, T., 2018, “When it comes to gorillas, Google photos remains blind,” Conde Nast (June), https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.
4. Post, W., 2015, “Google shows far more ads for high-paying jobs to men than women. Is the algorithm sexist – or is it us?,” https://nationalpost.com/life/is-the-google-algorithm-sexist.
5. BBC, 2011, “Bank of America fined $335 million for minority discrimination,” December, https://www.bbc.co.uk/news/business-16296146.
6. Ajunwa, I., 2015, “The other big U.S. Supreme Court decision we should be celebrating is one no one’s talking about,” Quartz, https://qz.com/438704/the-other-big-supreme-court-decision-we-should-be-celebrating-is-one-no-ones-talking -about/.
7. Schwartz, O., 2019, “Untold history of AI: Algorithmic bias was born in the 1980s,”

https://spectrum.ieee.org/ tech-talk/tech-history/dawn-of-electronics/untold-history-of-ai-the-birth-of-machine-bias.

8. Tom Simonite, 2019, “A Health Care Algorithm Offered Less Care to Black Patients”, https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/
9. Diane Harris, 2019, “Apple Card Gender Bias? Don't Assume its Discrimination, Experts Warn”, https://www.newsweek.com/apple-card-gender-bias-credit-limit-goldman-sachs-1471146
10. Verma, S. and Rubin, J., 2018, “Fairness definitions explained,” Proceedings of the International Workshop on Software Fairness (FairWare 18), doi: 10.1145/3194770.3194776.
11. Kusner, J. M., Loftus, R. J., Russell, C., et al., 2018, “Counterfactual fairness,” https://arxiv.org/abs/1703.06856.
12. Niki, Ball, J., P., Kusner, J., M., Weller, et al., 2019, “The sensitivity of counterfactual fairness to unmeasured confounding,” https://arxiv.org/abs/1907.01040.

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation