April 2021

Investigating the Impact of Unfairness Mitigation Techniques on Privacy Preserving Machine Learning

Raphaëlle Tseng

260767650 - COMP 400

An increasing reliance on data and machine learning systems has resulted in more and more important decisions being made by models. We are faced with the challenge of learning from data whilst ensuring that sensitive aspects of said data, such as sex or education, remain private, and that the decisions made by the automated systems and inference models are fair. We investigate the relationship between privacy and fairness by first ensuring our data remains differentially private, and then applying bias mitigation techniques. We then evaluate models trained with both differential privacy and bias mitigation against a variety of fairness metrics.

1. Introduction

As our reliance on data and machine learning decision systems continues to increase, we have a responsibility to ensure that the models we train guarantee individual privacy and do not exacerbate existing social disparities and unfair judgements. Privacy and fairness have been discussed more and more as the topic of ethics in artificial intelligence (AI) has gained prominence, however they are oven covered as distinct topics. Both privacy and fairness strive to protect the rights of users and subjects of software systems, and more often than not, work in tandem with each other. Privacy may be defined as 'The condition of not having undocumented personal knowledge about one possessed by others" Dwork and Roth (2014). Fairness has always been a more complicated notion to define, especially in the realm of computer science. Fairness, with regards to the law, is 'seeking to treat people justly on an individual basis with regards to the use of information regarding them'. Algorithmic fairness stipulates that 'A person's experience with an information system should not irrelevantly depend on their personal characteristics' Ekstrand et al. (2018). We measure an AI system's fairness based on its impact on people and the harm it may cause. Allocation harms occur when a system makes decisions, for example, to assign loans. A quality of service harm occurs when a system works differently for one group of people vs another.

When the fairness of a model is not considered or perhaps even taken for granted, there is an important risk that people may be negatively impacted and harmed. Cases of allocation harms and quality of service harms have been recently documented in the case of software systems. In 2018, commercially available facial recognition programs by major technology companies like IBM and Microsoft were recorded as demonstrating skin-type and gender-type biases, with far greater error rates when trying to determine the gender of darker-skinned women than light-skinned men. The effect of unfair systems can be even more sinister when applied to the context of the judiciary and criminal system. In 2016, a study showed that a software program used to predict a defendant's likelihood of re-offending by giving them a score from 1 (low risk) to 10 (high risk) was more likely to allocate a low number to a white defendant. Black defendants were almost twice as likely as white defendants to be labelled as higher risk but not actually re-offend. If we are trusting automated decisions to make decisions that impact people's lives, we have a responsibility to ensure that they treat all people fairly.

Whilst the importance of fairness and fair automated systems has only recently become more apparent, a large portion of the general public has always been concerned about the privacy risks associated with data leakage and data usage. Especially in cases where automated systems are trusted with sensitive data, such as health records, it is important that we prevent personal information from being misused, from being used for tasks the user has not expected or explicitly consented to, and from falling into the wrong hands, perhaps due to hacks.This last point is particularly pertinent in the case of machine learning where we want to ensure training data privacy, and guarantee that a malicious actor will not be able to reverse engineer our model to access our training data, or infer any information about our training data from the predictions made by our models.

Both privacy and fairness search to protect people from exploitation, and we should be looking to ensure that future machine learning systems do not disenfranchise any group for any reason. However, existing work by Cummings et al. (2019) has asked if it is currently possible to guarantee the privacy of training data, fairness of predictions made by a model, and reasonable overall accuracy, concluded that it is possible to satisfy only two of these three criteria. Their findings force us to ask the question, if none of the current, existing algorithms and models used today to make decisions and allocate resources are fair, private and accurate, which conditions are being prioritised and which are being ignored. Furthermore, who ends up suffering from these shortcomings? More often than not, work by Farrand et al. (2020) reveals that minority subgroups of data suffer more utility loss compared to others. When it comes to accuracy, the less represented groups which already achieve lower accuracy, end up losing more utility: the poor become poorer. In addition, as stricter privacy guarantees are imposed, this gap gets wider. This gap can have hugely significant societal and economic implications for people in all areas. In this work, we will set out to examine how bias mitigation techniques impact privacy preserving machine learning models.

2. Related Works

Currently, there has not been much research specifically regarding the trade-offs between privacy and fairness. There are a few works that try to achieve private and fair learning, but the works are limited to a specific privacy preserving approach, and don't provide a comprehensive comparison between different methods.

'Privacy for All: Ensuring Fair and Equitable Privacy Protections' Ekstrand et al. (2018) provides an important, in-depth literature review of the actual state of research at the intersection of privacy and fairness, exploring the main concepts in both areas, key definitions, important questions worth asking, and how we might go about answering them.They argue that privacy-related literature fails to address the potential discriminatory risks which may have unfair impacts on members of vulnerable groups. More focus needs to be placed on ensuring that recent research ensures sociotechnical systems are fair and non-discriminatory to the privacy projections those systems may provide, as privacy literature rarely considers whether proposed privacy schemes protect all people uniformly. In fact, more often than not, privacy regimes may disproportionately fail to protect vulnerable members of their target population, leading to disparate impact. Technology and policies intended to protect users of information systems should strive to provide such protection in an equitable fashion. Their work investigates how fairness and privacy interact, complement, and compete with each other, by reviewing the existing definitions and methods that are currently used. They define the goal of fair privacy protection as the probability of failure and expected risk being statistically independent of the subject's membership in a protected class. This position paper offers a direction for research, stressing the importance of discussion on the topic to promote more fair and nondiscriminatory sociotechnical systems that ensure privacy protection, but it does not touch on what could be done in industry or with regards to policy at the moment, if we hope to live in a society with equitable, just systems.

An investigation by Farrand et al. (2020) looked at how different levels of imbalance in the data affect the accuracy and fairness of decisions made by a model given different levels of privacy. They concluded that even small imbalances and loose privacy guarantees can cause disparate impacts. Their work demonstrates that as datasets become more imbalanced and stricter privacy guarantees are used, the fairness of the model is reduced. They used differential privacy to mitigate the challenges created by deep neural networks memorizing information from training sets and being exploited to extract sensitive information. Their experiment measured the fairness outcomes with the metrics demographic parity and difference in equality of opportunity. Their three main findings were: 1) The disparate impact of differential privacy on model accuracy is not limited to highly imbalanced data and can occur where the classes are only slightly imbalanced. 2) The disparate impacts are not limited to high privacy levels. Even for loose guarantees, differential privacy has disparate impacts on model accuracy. 3) By increasing privacy levels, we don't always see an increase in disparate impacts, since tighter privacy guarantees degrade the utility so much that the model becomes more random and therefore more fair.

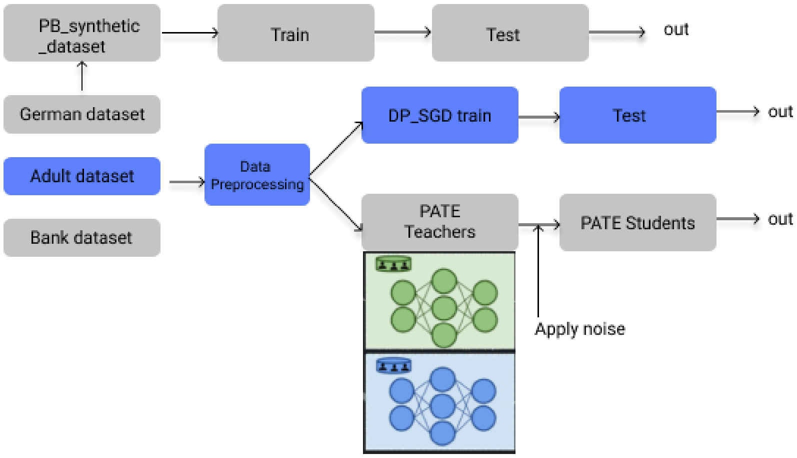
Further research titled 'On the Compatibility of Privacy and Fairness' Cummings et al. (2019) asks whether or not privacy and fairness can be simultaneously achieved by a single classifier in several different models, considering the trade-offs between differential privacy and fairness with respect to equal opportunity. They investigate the tensions existing between differential privacy and statistical notions of fairness, such as equality of false positives and false negatives. They look to find an efficient algorithm for classification whilst maintaining utility and satisfying both privacy and approximate fairness with a high probability. They are able to find an algorithm that is differentially private, approximately fair, and accurate, but not necessarily efficient; it does not have a polynomial time implementation in general. They claim there is no classifier that achieves a good enough differential privacy that satisfies equal opportunity whilst preserving accuracy.

The paper 'Fair Decision Making Using Privacy-Protected Data' Pujol et al. (2020)also demonstrates that noise added to privatize data, as is the case with differential privacy, may disproportionately impact some groups over others, leading to disparity in accuracy in decision-making tasks. Stronger privacy leads to much greater accuracy disparity. Agarwal (2020) extends the work by Cummings et al. (2019) and proves the impossibility theorem - a social-choice paradox illustrating the flaws of ranked voting systems. His work shows how impossible it is to design an accurate (even approximately) learning algorithm that is both differentially private and fair. Abadi et al. (2016) introduced differentially private stochastic gradient descent (DP-SGD), a modification of the stochastic gradient descent algorithm, which is the basis for many optimizers popular in machine learning. Models trained with DP-SGD have provable privacy guarantees expressed in terms of DP, by requiring that the probability of learning any particular set of parameters stays roughly the same if we change a single training example in the training set. As a result, if a single training point does not affect the outcome of learning, the information contained in that training point cannot be memorized and the privacy of the individual who contributed this data point to our dataset is respected. Cummings et al. (2019), Agarwal (2020) and Pujol et al. (2020) define fairness as equal opportunities, different groups should have relevantly equal true positives based on prediction accuracy.

3. Method

This project builds upon the existing work done by Andrea Jang, exploring the extent to which privacy algorithms impact group and individual fairness in machine learning decision making systems. Andrea's project investigates the role of privacy preserving machine learning techniques in all the parts of the pipeline. Three differentially private learning methods are used: 1) Using a differentially private synthetic dataset as input (pre-processing stage) 2) Optimizing privacy using differentially private stochastic gradient descent (DP-SGD) (in-processing stage) 3) A machine learning differentially private framework called the PATE method (which can be thought of as a post processing method as noise is added after the first teacher models are trained). PATE is an example of an ensemble model, where multiple models are learning at the same thing at the same time. Andrea then evaluates these different methods against a set of fairness metrics: Equalized Odds, Error Parity, and Demographic Parity. Three different datasets are also utilised.

Figure 1. Andrea's existing work. The stream this project follows is in blue.



This project focuses on a particular stream of Andrea's work. Firstly, we only focus on one dataset: the UCI Adult Data set, which can be used to predict whether income exceeds $50K/yr based on census data. The dataset has 48842 instances and 14 different categorical and continuous attributes, such as age, education, martial-status and sex. Secondly, we consider only the in-processing method of differential privacy: DP-SGD. We then apply bias mitigation techniques and evaluate the model's accuracy and its performance against the fairness metrics of error and disparity. The goal is to compare the model's performance with different combinations of bias mitigation and DP-SGD.

Differential privacy inadvertently usually results in the compromising of a model's accuracy as noise is added. In the case of DP-SGD, a privacy optimiser is attached to the existing SGD optimiser. As SGD iterates to optimize the objective function, random noise is added by the privacy optimizer to essentially randomise the process. As a result, adversary models cannot directly locate each update that occurs or access any of the hidden parameters.

The importance of fairness in machine learning systems has only become apparent in recent years, as ML models have risen in prominence and more and more industries have become reliant on such systems to make key decisions. As a result, there do not exist many widely available, or even thoroughly documented, tools to measure and mitigate fairness, bias, and discrimination in machine learning models throughout the AI application life cycle. In addition, most of the tools that do exist have been developed by major technology companies like IBM and Microsoft. 'The Grey Hoodie Project' Abdalla and Abdalla (2021) draws parallels between Big Tech's funding of large AI and AI fairness conferences to Big Tobacco funding medical research in order to sway and influence academic and public discourse. The authors examine how the funding of academic research may be used as a tool by Big Tech to put forward a socially responsible public image and distort the academic landscape to suit its needs.

The two most prominent unfairness mitigation tools currently are IBM's Aif360 and Microsoft's Fairlearn. There exist other tools that range in rigour and usefulness. PWC's Responsible AI Toolkit and the University of Chicago Center for Data Science and Public Policy's Aequitas act more as checklists and risk-assessments to try and ensure informed and equitable decisions regarding predictive tools are made. Facebook has its own Fairness-Flow but it is not open source. Google has also developed two tools, What-if, and the Model Card Toolkit. These allow us to visualise and analyse model behaviour for different fairness metrics, but not to mitigate unfairness.

Aif360 by IBM supports four key fairness metrics: 1) Statistical Parity Difference: The difference of the rate of favourable outcomes received by the unprivileged group to the privileged group. 2) Equal Opportunity Difference: The difference of true positive rates between the unprivileged and the privileged groups. 3) Average Odds Difference: The average difference of false positive rate (false positives/negatives) and true positive rate between unprivileged and privileged groups. 4) Disparate Impact: The ratio of rate of favourable outcome for the unprivileged group to that of the privileged group. Whilst Aif360 does contain a wider range of different metrics and bias mitigation algorithms, the less developed documentation and the fact that Andrea's work already utilised Fairlearn metrics encouraged us to first explore Fairlearn's algorithms. Both these tools are constrained by their need for sklearn classifiers. In the future, this project could be expanded to utilise Aif360's tools too.

Fairlearn presents a set of tools to assess fairness and mitigate unfairness of predictors for classification and regression, based on the works of Agarwal et al. (2018), Agarwal et al. (2019). Fairlearn uses the definition of fairness known as group fairness, which asks which groups of individuals are at risk for experiencing harms? The relevant groups are defined using a sensitive feature, which is passed to the Fairlearn estimator. As such, the system designer should be sensitive to these features when assessing group fairness. These attributes may or may not have privacy implications. In this work, our sensitive feature is sex.

Group fairness is determined by applying constraints on the behaviour of the predictor. These constraints are known as parity constraints. Fairlearn considers Demographic (or Statistical) Parity and Equalized Odds. In binary classification, a classifier h satisfies demographic parity under a distribution over (X,A,Y) if its prediction h(X) is statistically independent of the sensitive feature A. Demographic Parity seeks to mitigate allocation harms. It requires that individuals are offered the opportunity (making >$50k/year) independent of membership to the sensitive class A. Simply put, men and women should make >$50k/year at the same rate. Disparity metrics are also used to evaluate how far a given predictor departs from satisfying a parity constraint. This work primarily utilises Demographic Parity, a stronger version of the US Equal Employment Opportunity Commission's 'four fifths rule' which requires that the 'selection rate for any race, sex, or ethnic group must be at least 80% of the rate for the group with the highest rate.

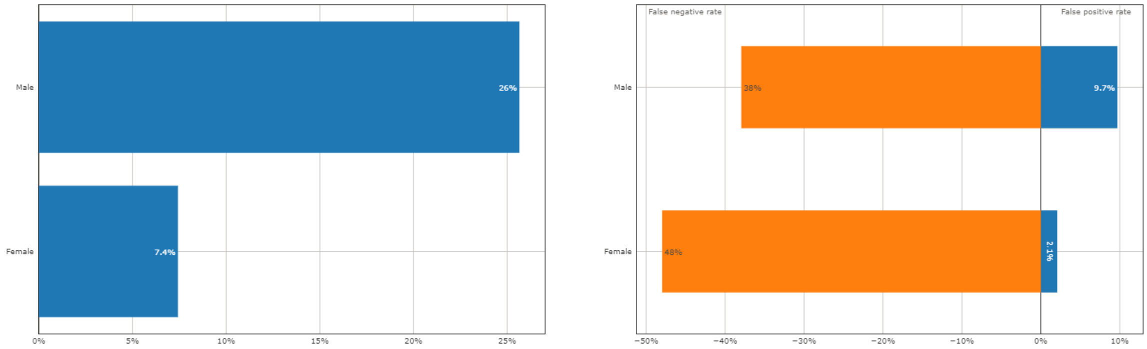
Fairlearn provides three possible bias mitigation algorithms. The Threshold Optimizer, based on 'Equality in Supervised Learning' Hardt et al. (2016), is a post-processing technique that takes an existing classifier and sensitive feature and derives a monotone transformation of the classifier's prediction to enforce the specified constraints. Exponentiated Gradient, effective with a categorical sensitive feature, and Gridsearch, effective with a binary sensitive feature, are both wrapped (reduction) approaches based on 'A Reductions Approach to Fair Classification' Agarwal et al. (2018). In this project, the first mitigation technique we use is the Grid Search. The Grid Search works well with Demographic Parity, as it minimises the number of constraints. Whilst it does often lower accuracy, it is noted that selecting a deterministic classifier, even if that means lower accuracy or a modest violation of the fairness constraints, is sometimes preferable. The method takes a standard black-box machine learning estimator and generates a set of retrained models using a sequence of re-weighted training datasets; a sequence of re-labellings and re-weightings is generated, and a predictor trained for each.

Fairlearn is not designed to work with pyTorch or any framework other than skLearn. As such, we had to first wrap our pyTorch model with Skorch to ensure we could access the necessary *fit* and *predict* functions for the bias mitigation to work. We also wrote our own dataset class to ensure we could crop the sensitive attribute A, in this case 'sex', but also access it throughout the training process. We used batch sizes of 32. The privacy optimiser necessary to do the DP-SGD was also not compatible with Fairlearn's expected input and as such, we had to rewrite a different version of Grid Search that would work with our new optimiser. The Grid Search was completed with Demographic parity as the constraint, and we logged the Error Rate, the Disparity (from the Demographic Parity), and Accuracy. Due to complications with time and computer power, the model was only tested on half the dataset.

4. Results

Initial outputs without DP enabled, gave extremely high fluctuations in the results. We experimented with changing the learning rates but this sometimes caused the accuracy to drop below 0.5. Using an unmitigated SkLearn logistic regression model on the adult census data yields results with vast disparities in accuracy and opportunity. Figure 2a shows the selection rate in each group, meaning the fraction of points classified as 1, where 1 signifies making >$50K/year. Figure 2b shows the false negative and false positive rates in each group. The orange false negative rate represents a prediction of 0 when the true value is 1, and the blue false positive rate demonstrates a prediction of 1 when the true value is 0. Through the Fairlearn Dashboard, it is clear that men are about three times more likely that women to make more than $50k/yr. Despite removing the sensitive attribute 'sex' from the training data, the predictor still discriminates based on sex. Simply removing or ignoring the sensitive feature to try and eliminate unfairness is therefore clearly insufficient a method. Table 1 reflects the scores of the unmitigated sklearn logistic regression model. (It is reflected separately from our own results as these results are taken directly from Fairlearn's own examples and displayed using Fairlearn's own Fairness Dashboard, where as we used weights and balances.ai). Error represents the misclassification error rate.

Figure 2. Unmitigated sklearn model results.



(a) Selection Rate

(b) False Positives and False Negatives

Table 1: Unmitigated sklearn.

	Accuracy	Disparity	Error
Baseline SkLearn Model	0.8540	0.289	0.382

In total, we ran five different tests to compare. Firstly, we ran baseline tests without DP-SGD and without any bias mitigation on both the sklearn model and our skorch wrapped pyTorch multilayer perceptron model. We then ran both these models with bias mitigation, before finally attempting a run with the privacy optimiser attached to the pyTorch model. The accuracy of the pyTorch model began above 75% and the sklearn one around 85%. We know that adding noise with differential privacy compromise the accuracy of the model. A drop in accuracy is also to be expected from the bias mitigation. The results between the unmitigated and mitigated sklearn models show that we should expect a larger reduction in disparity for a small loss in accuracy. In a perfect world, we would want an accuracy of 1.0, a disparity of 0.0, and an error rate of 0.0 too.

Our current results contain large amounts of fluctuation and noise, even without the privacy optimiser, giving a wide range of values. The mitigated run with pyTorch gave an accuracy that peaked at 0.7561 and dropped to 0.6829. The disparity for this run oscillated between 0.002714 and 0.1275. Similarly, the error rate was between 0.2456 and 0.2972. The results listed in Table 2 represent the model's values after having completed all the steps.

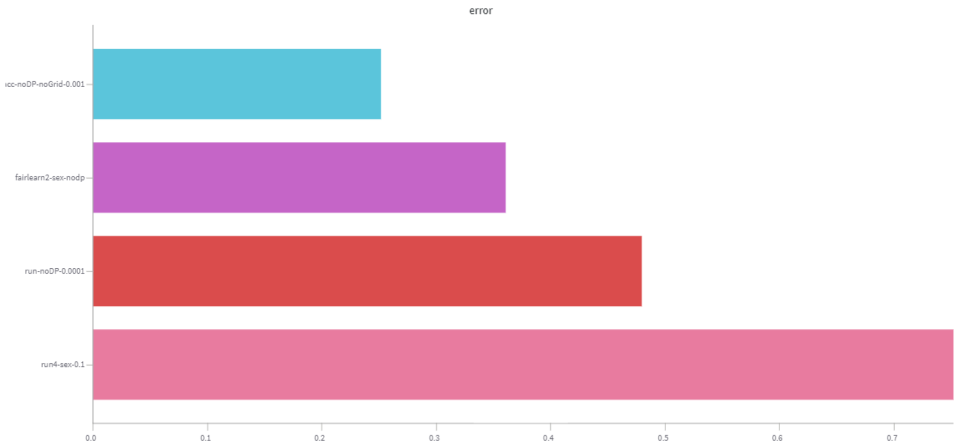
The disparity of the privacy optimised Skorch model is listed as NA because it outputted 0.0. This extremely low result may be due to the fact that we had to rewrite the Fairlearn Grid Search in order for it to be compatible with DP-SGD, and it is not working entirely as anticipated just yet. The accuracy of the private model drops massively, whilst the error rate jumps. The noise level for the privacy optimiser is kept at a low 0.1, forcing us to consider how the model results might be further negatively exacerbated by a higher noise level.

The mitigated pyTorch model performed with results comparable to the mitigated sklearn model. Although it has a slightly higher error rate and disparity score, it also has a higher accuracy score. So whilst it does not necessarily have the best fairness metric value, it achieves a solid performance overall. Having Fairlearn's algorithms be compatible with pyTorch allows us to potentially plan more elaborate experiments, and gather more information.

Table 2: Comparing methods.

	Accuracy	Disparity	Error
Baseline PyTorch Model	0.7505	0.0082	0.2517
SkLearn Model + Grid Search	0.6322	0.4679	0.3607
PyTorch Model + Grid Search	0.7561	0.4783	0.4796
PyTorch Model + Grid Search + DP (0.1)	0.2466	NA	0.7517

Figure 3. Comparison of Error Rates



5. Discussion and Conclusion

In order for machine learning systems to positively and successfully contribute to and advance in a just and equitable society, it is necessary that they consider both the privacy of their users' data and the fairness of their outputs. They ought to strive towards providing privacy protection mechanisms equitably to all people, and ensuring fairness and non-discrimination in the in all sociotechnical settings. Here, we have only just begun to explore the links weave both privacy and fairness together, as well as the trade-off between achieving fair results and ensuring privacy of data. Our results show that currently, attempting to build a model that successfully utilises Fairlearn's bias mitigation Grid Search procedure whilst applying DP-SGD, results in a detrimental effect on the error rate, disparity and accuracy of the model. However the pressing need and the social importance of both privacy and fairness mean that despite the clear challenge achieving both of these presents, we must continue to work at finding a solution.

Going forward, there is plenty of work that can be done. The imbalance of the dataset may be the cause for the vastly fluctuating results. Adding a sampler to try and even out the different classes in the pre-processing stage may help achieve better numbers. Furthermore, it may be worth exploring alternative routes to getting the privacy optimiser to be compatible with the Grid Search, instead of wrapping our model in Skorch and rewriting the Grid Search class. Utilising the whole dataset and continuing to test different learning rates might also lend better outcomes.

In the long term, the work of this project should be expanded to test more bias mitigation tools and compare them to more privacy protecting mechanisms like PATE and using synthetic datasets, in order to evaluate the trade-offs between each different method. Hopefully, gathering more information will allow us to come closer to finding a solution where both privacy and fairness can be guaranteed.

In the broader field, achieving fairness when training time access to protected attributes is unavailable and navigating trade-offs between accuracy and multiple fairness definitions remain open problems that also need solving. Tackling these problems are the first step of many that can be taken in order to achieve a just, equitable society, where decision making machine learning systems can be trusted to protect the interests of its users, mitigating discrimination and unfairness whilst also respecting the autonomy of members in the use and disclosure of their personal information.

Acknowledgments

I would like to thank Dr. Jin Guo and Andrea Jang of the Knowledge and Software Technology Lab for their patience, support, and guidance throughout this project. They provided plenty of valuable discussions and interesting insights.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, B. H., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security.*

Abdalla, M. and Abdalla, M. (2021). The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity.

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*.

Agarwal, A., Dudik, M., and Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning*.

Agarwal, S. (2020). Trade-offs between fairness, interpretability, and privacy in machine learning.

Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. (2019). On the compatibility of privacy and fairness.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy.

Ekstrand, M. D., Joshaghani, R., and Mehrpouyan, H. (2018). Privacy for all: Ensuring fair and equitable privacy protections. In *Proceedings of Machine Learning Research*. Conference on Fairness, Accountability and Transparency.

Farrand, T., Mireshghallah, F., Singh, S., and Trask, A. (2020). Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of 2020 Workshop on Privacy-Preserving Machine Learning in Practice*. ACM Conference on Fairness, Accountability and Transparency.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)*.

Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., and Miklau, G. (2020). Fair decision making using privacy-protected data. ACM Conference on Fairness, Accountability and Transparency.