Trade-Offs between Fairness and Privacy in Machine Learning

Sushant Agarwal
University of Waterloo, Canada
sushant.agarwal@uwaterloo.ca

Abstract

The concerns of fairness, and privacy, in machine learning based systems have received a lot of attention in the research community recently, but have primarily been studied in isolation. In this work, we look at cases where we want to satisfy both these properties simultaneously, and find that it may be necessary to make trade-offs between them. We prove a theoretical result to demonstrate this, which considers the issue of compatibility between fairness and differential privacy of learning algorithms. In particular, we prove an impossibility theorem which shows that even in simple binary classification settings, one cannot design an accurate learning algorithm that is both ϵ-differentially private and fair (even approximately).

1 Introduction

Technology has entered most aspects of our lives, with automated systems being deployed to make consequential decisions, such as predicting recidivism rates in released prisoners, and estimating the probability of an applicant returning a loan. Now, because these systems are making decisions that are potentially life-altering for many people, there have been many ethical questions raised about how they function. We will look at two ethical considerations in this work: fairness, and privacy. We would like the system to be *fair*, and not discriminate against an applicant just because of their membership in a minority/protected group (which could be a particular race, gender, etc.). The second concern is *privacy*. Now, because these decision making systems are typically machine learning models, and are trained on potentially sensitive data, we would not like to inadvertently leak information about people in the training data, and would like to protect their privacy.

These concerns have received a lot of attention in the research community in the last few years [Chaudhuri *et al.*, 2011; Corbett-Davies and Goel, 2018; Dwork *et al.*, 2006, 2014, 2012; Kleinberg *et al.*, 2017; Feldman *et al.*, 2015]. However, they have primarily been studied in isolation, that is, people have primarily looked at scenarios in which we would want to satisfy one of these properties at a time. In this work, we look at a case where we want to satisfy these properties simultaneously, and analyse how they interact. We find that that these properties are at odds with each other, and it is necessary to make trade-offs between them. We show a theoretical result to demonstrate this, which talks about the clash between the requirements of differential privacy, accuracy, and fairness in learning algorithms. It is an impossibility theorem which states that even in a very simple binary classification setting, no learning algorithm that is ϵ-differentially private (for any $\epsilon < \infty$), and approximately fair (i.e., the algorithm is guaranteed to output an *approximately fair classifier*[1]), can have non-trivial accuracy.

2 Related Work

The work that is most relevant to ours is that of Cummings *et al.* [2019]. Cummings *et al.* [2019] consider the trade-offs when considering learning algorithms that satisfy differential privacy and one particular notion of fairness (equal opportunity), and one of their results is a weaker version of ours. In particular, they claim that there is no learning algorithm that achieves ϵ-differential privacy, satisfies equal opportunity, and has accuracy better than a constant classifier. However, to the best of our understanding, we believe that there is a gap in their argument (see Section 4 where we describe what it is), and so their proof idea does not go through. So, our contribution here can be summarized as correcting their proof and also generalizing their result, by showing that such an impossibility holds with respect to many different notions of (even approximate) fairness.

Apart from the work of Cummings *et al.* [2019], another important paper that was a motivation for pursuing this line of work was that of Pujol *et al.* [2020]. Pujol *et al.* [2020] empirically show how there might be privacy-fairness trade-offs involved in certain settings. In particular, they consider three resource allocation settings and use census data to which noise has been added to demonstrate how adding noise to achieve differential privacy could disproportionately affect some groups over others in the settings that they consider. Besides the paper mentioned above, there is also work by Dwork

[1] An approximately fair classifier refers to classifiers that satisfy even relaxed or approximate versions of common notions of fairness (such as error rates, or false positive/negative rates, being approximately equal across different groups). Proving an impossibility for such relaxations makes our result stronger.

and Mulligan [2013], and Ekstrand *et al.* [2018], where they consider the issues of privacy and fairness together. They intuitively discuss about why there may be trade-offs involved in deploying algorithms in real life scenarios, and advocate for the issues of privacy and fairness to be studied together. Bagdasaryan *et al.* [2019] empirically demonstrate that in neural networks trained using differentially private stochastic gradient descent (DP-SGD), accuracy of DP models drops much more for the underrepresented classes. Our work can be considered as essentially lending some theoretical support to the observations in the above works, as we prove how even in very simple settings it may be impossible to achieve fairness and privacy together, while maintaining accuracy. There is also work by Jagielski *et al.* [2019], where they shows two algorithms that satisfy (ϵ, δ)-differential privacy and a particular notion of fairness (equalized odds). However, they consider the relaxed notion of approximate differential privacy ((ϵ, δ)-DP, $\delta > 0$), while we consider the stricter notion of pure differential privacy ($(\epsilon, 0)$-DP) throughout.

3 Trade-Offs between Fairness and Privacy

The result essentially shows that even in a simple binary classification setting, there is no learning algorithm that is fair (even approximately), and differentially private, while maintaining good accuracy. Hence, we see that, the properties of fairness, differential privacy, and accuracy, can be at odds with each other and it may not possible to satisfy the three of them simultaneously.

3.1 Setup

Throughout, we use $\mathcal{X}$ to denote the domain set. There is a probability distribution $\mathcal{D}$ over $\mathcal{X}$. The domain set consists of elements of the form $z = (x, a, y)$, where x refers to the element's features (e.g., this could be income, age, etc.), a is a protected (binary) attribute (we have a protected/minority and a majority group, and use $a = 0$ to denote the minority class that we wish to protect from discrimination). y is a binary label, that is what we want to predict. Additionally, throughout, we assume that $y = 0$ denotes the *bad* label—meaning, for instance, in the context of, say, giving loans, this means that the person will not return the loan.

3.2 Privacy

The notion of privacy we consider is called differential privacy. Differential privacy aims to protect the privacy of each individual in a database. In the case of learning algorithms, the database is the training set.

Differential Privacy

Differential privacy protects the privacy of an individual by ensuring that an algorithm will generate similar outputs on neighboring databases. It roughly protects the privacy of an individual in the database in the following way; changing an individual's entry, or deleting or adding it, will lead to what we call a neighboring database, and because the algorithm will generate similar outputs on neighboring databases, an observer seeing its output essentially cannot tell if a particular individual's information was used in the computation, or what that information is.

Definition ((ϵ, δ)-differential privacy [Dwork *et al.*, 2006])**.** For any $\epsilon, \delta \geq 0$, a randomized algorithm $\mathcal{A}$ is said to be (ϵ, δ)-differentially private if for all pairs of neighboring databases $\mathcal{D}$, $\mathcal{D}'$ and for all sets $S \in \text{Range}(\mathcal{A})$ of outputs,

$$\Pr[\mathcal{A}(\mathcal{D}) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{A}(\mathcal{D}') \in S] + \delta.$$

Remark. Although we have defined differential privacy in its full generality, note that we will be talking about $(\epsilon, 0)$-differential privacy throughout.

Database

We mention two different notions of a database.

1. The first one, is a finite sample, with entries drawn i.i.d. from the distribution $\mathcal{D}$ over the domain $\mathcal{X}$.
2. The second notion is to consider the whole distribution $\mathcal{D}$ as a database.

The first notion is standard in the privacy literature, where databases are viewed as a finite collection of data points from n individuals. The second notion is standard for statistical notions of fairness, where the goal is to ensure fairness over a large population. Notion 2 can simply be considered a generalization of Notion 1. As in Cummings *et al.* [2019], we will be using the second notion, but the same result and proof idea also work for the first notion.

Neighbouring Databases

Given our definition of a database, it now remains to be defined what we mean by *neighboring databases*. Here we use the notion of σ-closeness, which is also used by Cummings *et al.* [2019].

Definition (σ-closeness [McGregor *et al.*, 2010])**.** Distributions (i.e., databases) $\mathcal{D}$ and $\mathcal{D}'$ are said be σ-close if

$$\frac{1}{2} \sum_{z \in \mathcal{X}} |\mathcal{D}(z) - \mathcal{D}'(z)| \leq \sigma.$$

As in Cummings *et al.* [2019], we calculate the distance between two distributions (databases) by the above expression (this is also known as total variation distance), and if the distance is lesser than σ, for some pre-specified value of σ, then the distributions are said to be neighboring.

3.3 Fairness

What notion of fairness do we use? Our results hold for pretty much all the common notions proposed in popular literature (for example: Demographic Parity, Equal Opportunity (used in Cummings *et al.* [2019]), Equalised Odds, etc., see below for definitions) [Dwork *et al.*, 2012; Hardt *et al.*, 2016; Verma and Rubin, 2018]. Essentially, any reasonable notion of fairness, that does not allow one group to be treated much worse than the other. More importantly, our results hold for even relaxed or approximate versions of these notions (This means that, for example, instead of demanding equality in false positive/negative rates for both groups, we require that there should not be a high difference in these rates between the two groups). Proving an impossibility for such relaxations makes our result stronger.

Fairness definitions

Definition (Demographic parity)**.** A binary classifier h satisfies demographic parity if with respect to random variables A and Y

$$\Pr_{z\sim D}[h(z)=1|A=1] = \Pr_{z\sim D}[h(z)=1|A=0].$$

Definition (Equal opportunity [Hardt *et al.*, 2016])**.** A binary classifier h satisfies equal opportunity if with respect to random variables A and Y

$$\Pr_{z\sim D}[h(z)=1|Y=1,A=1] = \Pr_{z\sim D}[h(z)=1|Y=1,A=0].$$

In words, h satisfies equal opportunity if it produces equal true positive rates across the two groups.

Definition (Equalized odds [Hardt *et al.*, 2016])**.** A binary classifier h satisfies equalized odds if

- h has equal false positive rates across the two groups, i.e., with respect to random variables A and Y

$$\Pr_{z\sim D}[h(z)=1|Y=0,A=1] = \Pr_{z\sim D}[h(z)=1|Y=0,A=0]$$

- h satisfies equal opportunity.

3.4 Result

Our main result is an incompatibility theorem showing how differential privacy and fairness can be at odds with each other when we consider a learning algorithm with non-trivial accuracy. In particular, we consider a simple binary classification setting where the learning algorithm is given full access to the underlying distribution, and show that even under this severe restriction[2], any learning algorithm that is $(\epsilon, 0)$-differentially private, and even approximately fair, cannot achieve accuracy better than that of a constant classifier (that outputs the same label for every input).

Theorem 1. *If a learning algorithm $\mathcal{A}$ is $(\epsilon, 0)$-differentially private and is guaranteed to output an approximately fair classifier, then $\mathcal{A}$ is constrained to output a constant classifier, i.e., $\mathcal{A} : \tilde{D} \to \Delta(\mathcal{H})$, where $\tilde{D}$ denotes the set of all distributions, and*

$$\mathcal{H} = \{h : \mathcal{X} \to \{0,1\} \mid h \textit{ is a constant function}\}.$$

We start with an informal overview of the proof. The main idea in the proof is to first observe that, due to differential privacy constraints, if there is a classifier that is output with positive probability by $\mathcal{A}$ on a distribution $\mathcal{D}_1 \in \tilde{D}$, then $\mathcal{A}$ has to output this classifier with positive probability on any other distribution $\mathcal{D}'_1 \in \tilde{D}$. Now, what the claim above implies is that, if algorithm $\mathcal{A}$ has to be (approximately) fair as well, and it outputs classifier h on some input distribution, then h is always (approximately) fair, irrespective of the underlying distribution. Now, once we have the observation above, then it just remains to show that such classifiers—i.e., ones that are (approximately) fair with respect to any underlying distribution—belong to a very restricted set, namely $\mathcal{H}$ as defined in the theorem. Below, we present a formal argument by first proving the following claim.

Claim 2. *Let $\mathcal{A}$ be a learning algorithm that is $(\epsilon, 0)$-differentially private. Then, $\forall \mathcal{D}_1, \mathcal{D}'_1 \in \tilde{D}$, and for all classifiers h,*

$$\Pr[\mathcal{A}(\mathcal{D}_1) = h] > 0 \implies \Pr[\mathcal{A}(\mathcal{D}'_1) = h] > 0.$$

Proof. Consider an arbitrary distribution $\mathcal{D}_1 \in \tilde{D}$ and a classifier h such that $\Pr[A(\mathcal{D}_1) = h] > 0$. Next, consider any arbitrary distribution $\mathcal{D}'_1 \in \tilde{D}$. We need to show that $\Pr[\mathcal{A}(\mathcal{D}'_1) = h] > 0$.

To see this, first let us consider, for any $i \in [n]$ and $\eta > 0$, two η-close distributions $\mathcal{D}_i$ and $\mathcal{D}_{i+1}$ (i.e., they are neighboring databases). Since $\mathcal{A}$ is ϵ-differentially private, if $\Pr[\mathcal{A}(\mathcal{D}_i) = h] > 0$, then we have that $\Pr[\mathcal{A}(\mathcal{D}_{i+1}) = h] > 0$, for if otherwise, then we have,

$$0 < \Pr[A(\mathcal{D}_i) = h] \leq \exp(\epsilon) \Pr[A(\mathcal{D}_{i+1}) = h] = 0,$$

which is a contradiction.

Now, given the observation above, observe that, for any $\eta > 0$, one can construct a (finite) series of distributions $\mathcal{D}_2, \cdots, \mathcal{D}_n$ such that $\forall i \in [n]$, $\mathcal{D}_i$ and $\mathcal{D}_{i+1}$ are η-close (i.e., they are neighboring databases) and where $\mathcal{D}_{n+1} = \mathcal{D}'_1$. This in turn implies that we have,

$$\begin{aligned}
\Pr[A(\mathcal{D}_1) = h] > 0 &\implies \Pr[A(\mathcal{D}_2) = h] > 0 \\
&\implies \Pr[A(\mathcal{D}_3) = h] > 0 \\
&\vdots \\
&\implies \Pr[A(\mathcal{D}_{n+1}) = h] > 0,
\end{aligned}$$

where all the implications above are obtained by using the argument made above that for two neighboring databases $\mathcal{D}_i$ and $\mathcal{D}_{i+1}$, $\Pr[\mathcal{A}(\mathcal{D}_i) = h] > 0 \implies \Pr[\mathcal{A}(\mathcal{D}_{i+1}) = h] > 0$. This in turn proves our claim.

□

Proof of Theorem 1. From Claim 2 we know that if a learning algorithm $\mathcal{A}$ is $(\epsilon, 0)$-differentially private and is guaranteed to output a fair classifier, then for all fair classifiers h and $\forall \mathcal{D}_1, \mathcal{D}'_1 \in \tilde{D}$, $\Pr[\mathcal{A}(\mathcal{D}_1) = h] > 0 \implies \Pr[\mathcal{A}(\mathcal{D}'_1) = h] > 0$. In other words, what this implies is that, for a fair learning algorithm $\mathcal{A}$, any fair classifier h that is output by $\mathcal{A}$ is fair with respect to any distribution in $\tilde{D}$. Below, we show how any h satisfying the property mentioned above should belong to $\mathcal{H}$, where $\mathcal{H}$ is as defined in the statement of the theorem.

To do this, consider for the sake of contradiction any $h \notin \mathcal{H}$. This implies that, there exist points $p_1 = (x_1, 0, y_1)$ and $p_2 = (x_2, 1, y_2)$ classified differently by h (Because h is not constant, we can find two points $a, b \in \mathcal{X}$ such that $h(a) \neq h(b)$. If they are in different groups, we are done. If they are in the same group, then choose any $c \in \mathcal{X}$ in the other group. It will hold that either $h(c) \neq h(a)$, or $h(c) \neq h(b)$, and we are done). Then, either of the following two cases holds:

1. $h(p_1) = 0$ and $h(p_2) = 1$, or
2. $h(p_1) = 1$ and $h(p_2) = 0$.

[2]The result also holds for the case where the algorithm has access to a finite training set, and not the underlying distribution. Giving the algorithm access to the underlying distribution is an easier task (equivalent to providing infinite training samples). Proving an impossibility for this case makes our result stronger.

Now, if this is the case, then we will construct a distribution on which h is unfair. We construct a distribution for Case 1. To construct such a distribution, let us first consider the following points.

$$q_1 = (x_1, 0, 1) \qquad q_2 = (x_2, 1, 0)$$

Next, let us define the following distribution $\mathcal{D}'$.

$$\mathcal{D}'(q_1) = \frac{1}{2} \qquad \mathcal{D}'(q_2) = \frac{1}{2}$$

Note that $h(q_1) = 0$ and $h(q_2) = 1$. However, if this is the case, then note that by any reasonable notion of fairness, h is unfair to group 0 as compared to group 1, since group 0 always has true label 1 but is always labeled 0, whereas group 1 always has true label 0 but is always labeled 1. For example, note that demographic parity is clearly not satisfied, even approximately.

Note that in the above distribution, some common fairness notions such as equal opportunity are not defined (because false positive rate for group 0 is undefined). To construct a distribution on which a broader class of fairness notions are defined, let us first consider the following points.

$$q_1 = (x_1, 0, 1) \qquad q_2 = (x_1, 0, 0)$$
$$q_3 = (x_2, 1, 0) \qquad q_4 = (x_2, 1, 1)$$

Next, for some small $\epsilon > 0$, let us define the following distribution D'.

$$D'(q_1) = \frac{1}{2} - \epsilon \qquad D'(q_2) = \epsilon$$
$$D'(q_3) = \frac{1}{2} - \epsilon \qquad D'(q_4) = \epsilon$$

Since h depends only on the observable attributes, note that $h(q_1) = h(q_2) = 0$ and $h(q_3) = h(q_4) = 1$. However, if this is the case, then note that by any reasonable notion of fairness, h is unfair to group 0 as compared to group 1, since most of their points actually have true label 1 but they are all labeled 0, whereas most of the points of group 1 have true label 0 but they are all labeled 1. For example, note that equal opportunity and equalised odds are clearly not satisfied, even approximately.

We omit the construction for Case 2. Essentially the same idea as Case 1 can be used for Case 2. □

4 Gap in Proof in Previous Work

As mentioned previously, our result here is a stronger version to the one in Cummings *et al.* [2019], but we believe that their proof has a gap. Below we describe what this gap is.

Error in Theorem 1. On a high level, what their proof tries to do is, given a distribution $\mathcal{D}$ and a classifier h that satisfies equal opportunity (and is output with non-zero probability) for this distribution, to construct a neighboring distribution $\mathcal{D}'$ on which h does not satisfy equal opportunity. Now, h is output with non-zero probability on $\mathcal{D}$, and because of differential privacy constraints, it is output with non-zero probability on $\mathcal{D}'$ as well. This would imply the algorithm is not fair, because on input distribution $\mathcal{D}'$, it outputs an unfair classifier (h) with non-zero probability. However, there is error in this construction, and h does indeed satisfy fairness (equal opportunity) on the distribution $\mathcal{D}'$, contrary to what is claimed. The equal opportunity notion of fairness requires that a classifier h satisfies (with respect to group A and label Y on distribution $\mathcal{D}'$)

$$\Pr_{z \sim D'}[h(z) = 1 | Y = 1, A = 1] = \Pr_{z \sim D'}[h(z) = 1 | Y = 1, A = 0].$$

It is therefore crucial for their proof to show inequality of group-conditional true positive classification rates for classifier h on the distribution $\mathcal{D}'$, denoted by

$$\gamma_{ya}(h) = \Pr_{D'}[h = 1 | Y = y, A = a].$$

I.e., they require that

$$\gamma_{10}^{D'} \neq \gamma_{11}^{D'}.$$

which does not hold. The claim is that

$$\gamma_{10}^{D'} = \frac{1}{4} - \tau \neq \frac{1}{4} + \tau = \gamma_{11}^{D'}.$$

However, it is easy to see that

$$\gamma_{10}^{D'} = 1 = \gamma_{11}^{D'}.$$

and therefore h does indeed satisfy equal opportunity on the distribution $\mathcal{D}'$, contrary to what is claimed. The error seems to stem from an incorrect usage of conditional probability arguments, and unfortunately this error does not seem fixable within the same proof idea. In any case, we do think that the statement is correct, and we prove a stronger claim. □

5 Conclusion and Future Work

Through this work, we see that in machine learning based decision systems, the desiderata of fairness, and privacy may be at odds with each other and it is often necessary to make trade-offs between them, if we want to maintain accuracy. We prove a theoretical result to demonstrate this, an incompatibility theorem showing a setting where pure differential privacy and (even relaxed notions of) fairness are at odds with each other when we want a learning algorithm with non-trivial accuracy. In particular, we consider the task of learning a classifier for a simple binary classification setting and show that any learning algorithm that is $(\epsilon, 0)$-differentially private, and even approximately fair, cannot achieve accuracy better than that of a constant classifier.

The current statement allows the learning algorithm to be faced with any underlying distribution (without any restrictions). But in reality, it's probably more likely that the set of distributions the learning algorithm will encounter follow some niceness properties. So, if we restrict the distributions by these niceness properties, can we prove something similar? Additionally, in the result, we require each output classifier to be fair. An algorithm that generates a fair classifier with high probability could also be considered as fair, and such relaxations could definitely be looked at.

Acknowledgments

We would like to thank Shai Ben-David, Rachel Cummings, Gautam Kamath, Vijay Menon, Yaoliang Yu, and anonymous reviewers for useful comments.

References

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, pages 15453–15462, 2019.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Fairness in User Modeling, Adaptation and Personalization (FairUMAP 2019)*, 2019.

Cynthia Dwork and Deirdre K Mulligan. It's not privacy, and it's not fair. *Stan. L. Rev. Online*, 66:35, 2013.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Michael D Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*, pages 35–47, 2018.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008, 2019.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 81–90. IEEE, 2010.

David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 189–199, New York, NY, USA, 2020. Association for Computing Machinery.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.