Blind Justice: Fairness with Encrypted Sensitive Attributes

Niki Kilbertus [1 2] **Adrià Gascón** [3 4] **Matt Kusner** [3 4] **Michael Veale** [5] **Krishna P. Gummadi** [6] **Adrian Weller** [2 3]

Abstract

Recent work has explored how to train machine learning models which do not discriminate against any subgroup of the population as determined by sensitive attributes such as gender or race. To avoid disparate treatment, sensitive attributes should not be considered. On the other hand, in order to avoid disparate impact, sensitive attributes must be examined—e.g., in order to learn a fair model, or to check if a given model is fair. We introduce methods from secure multi-party computation which allow us to avoid both. By encrypting sensitive attributes, we show how an outcome-based fair model may be learned, checked, or have its outputs verified and held to account, *without users revealing their sensitive attributes*.

1. Introduction

The growing field of *fair learning* seeks to formalize relevant requirements, detection, and mitigation of potential illegal or unfair discrimination against certain subgroups of the population (Schreurs et al., 2008; Calders & Žliobaitė, 2012; Friedler et al., 2016). Most legally-problematic discrimination centers on *sensitive attributes* (Barocas & Selbst, 2016). Simply not inquiring about sensitive attributes to avoid *disparate treatment* (Grgić-Hlača et al., 2018), does not protect against *disparate impact* (Dwork et al., 2012), which occurs when the *outcomes* of decisions—perhaps unintentionally—disproportionately benefit or hurt particular sensitive groups (Handel et al., 2014). Much recent work has focused avoiding various notions of disparate impact (Feldman et al., 2015; Hardt et al., 2016; Zafar et al., 2017).

[1]Max Planck Institute for Intelligent Systems [2]University of Cambridge [3]Alan Turing Institute [4]University of Warwick [5]University College London [6]Max Planck Institute for Software Systems. Correspondence to: Niki Kilbertus <nk470@cam.ac.uk>.

A longer version of this paper has been accepted at ICML 2018: arxiv:1806.03281.

5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2018), Stockholm, Sweden, 2018.

In order to check and enforce such requirements, the modeler must have access to the sensitive attributes in the training data (Žliobaitė & Custers, 2016; Lipton et al., 2017)—however, this may be undesirable for several reasons. First, individuals are unlikely to want to entrust sensitive attributes to modelers. Even if a modeler was trusted, the wide provision of sensitive data creates heightened privacy risks in the event of a data breach.

Second, legal barriers such as EU's General Data Protection Regulation (GDPR), may limit collection and processing of sensitive personal data, which modelers cannot justify with their "legitimate interests" (Veale & Edwards, 2018). Veale & Binns (2017) addressed this problem by involving a highly trusted third party, requiring individuals to disclose their sensitive attributes (risking breaches or hacks) and the modeler to disclose their model (resulting in intellectual property concerns) to this third party.

Contribution. Based on recent methods from *secure multi-party computation* (MPC), we propose an approach to detect and mitigate disparate impact ensuring that both individuals' sensitive attributes and the modeler's model remain private to all other parties. This reflects the notion that decisions should be blind to an individual's status—depicted in courtrooms by a blindfolded Lady Justice holding balanced scales (Bennett Capers, 2012). In our framework, we assume the existence of a regulator with fairness aims (e.g., a data protection authority or anti-discrimination agency).

Desirable fairness and accountability applications we enable include: 1. **Fairness certification:** Given a model and a dataset of individuals, check that the model satisfies a given fairness constraint; if yes, generate a certificate. 2. **Fair model training:** Given a dataset of individuals, learn a model guaranteed and certified to be fair. 3. **Decision verification:** To prevent a malicious modeler from using a non-certified model in practice (Kroll et al., 2016), we provide for an individual to challenge an outcome and check that it matches the outcome from a certified model.

Our extension of recent theoretical developments in MPC to admit linear constraints may be of independent interest. We demonstrate the real-world efficacy of our methods, and make our code publicly available at `https://github.com/nikikilbertus/blind-justice`.

2. Fairness and Privacy Requirements

Assumptions and Incentives. We assume three categories of participants: a *modeler* M, a *regulator* REG, and *users* $\mathrm{U}_1 \dots \mathrm{U}_n$. Each user has a vector of binary sensitive attributes $\mathbf{z}_i \in \{0,1\}^p$ (e.g., ethnicity or gender), a vector of non-sensitive features $\mathbf{x}_i \in \mathbb{R}^n$ (discrete or real), and a non-sensitive recorded outcome $y_i \in \{0,1\}$—the *label*. We collect user data into matrices $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{Z} \in \{0,1\}^{n \times p}$ and a label vector $\mathbf{y} \in \{0,1\}^n$ and refer to non-sensitive data by $\mathbf{D} = (\mathbf{X}, \mathbf{y})$. The source of societal concern is that sensitive attributes $\mathbf{z}_i$ are potentially correlated with $\mathbf{x}_i$, y_i.

Modeler M wishes to train a model $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \{0,1\}$, keeping $\boldsymbol{\theta}$ private. The model $f_{\boldsymbol{\theta}}$ does not use $\mathbf{z}_i$ as input to prevent disparate treatment. For each user U_i, M observes or is provided $\mathbf{x}_i, y_i$. The sensitive information in $\mathbf{z}_i$ is required to ensure $f_{\boldsymbol{\theta}}$ meets a given fairness condition $\mathbb{F}$, but the users want to keep it private from all other parties. The regulator REG aims to ensure that M deploys only models that meet fairness condition $\mathbb{F}$. It has no incentive to collude with M. Further, M might be obliged to demonstrate to REG that condition $\mathbb{F}$ is met, before publicly deploying $f_{\boldsymbol{\theta}}$.

In this work we focus on a variant of balancing acceptance rates across demographic groups, formulated as a constrained optimization problem by Zafar et al. (2017) mimicking the $p\%$-rule: for any binary protected attribute $z \in \{0,1\}$, it aims to achieve

$$\min\left\{\frac{P(\hat{y}=1 \mid z=1)}{P(\hat{y}=1 \mid z=0)}, \frac{P(\hat{y}=1 \mid z=0)}{P(\hat{y}=1 \mid z=1)}\right\} \geq \frac{p}{100} \,. \quad (1)$$

We believe that a similar MPC approach could also be used for balancing accuracy, true positive rates, and true negative rates, i.e., all measures which, to our knowledge, have been addressed with efficient, non-private methods.

Fairness Certification. The modeler M works with the regulator REG to obtain a certificate that model $f_{\boldsymbol{\theta}}$ is fair. To do so, we propose that users send their non-sensitive data $\mathbf{D}$ to REG; and send *private* versions of their sensitive data $\mathbf{Z}$ to both M and REG. Neither M nor REG can read the sensitive data. However, we can design a secure protocol between M and REG to certify if the model is fair. This setup is shown in Figure 1 (*Left*).

Privacy constraints: (C1) *privacy of sensitive user data:* no one other than U_i ever learns $\mathbf{z}_i$ in the clear, (C2) *model secrecy:* only M learns $f_{\boldsymbol{\theta}}$ in the clear, and (C3) *minimal disclosure of* $\mathbf{D}$ *to* REG*:* only REG learns $\mathbf{D}$ in the clear.

Fair Model Training. A modeler M learns a fair model without access to users' sensitive data $\mathbf{Z}$. This is solved by having users send $\mathbf{D}$ to M and send *private* versions of $\mathbf{Z}$ to both M and REG. We shall describe a secure MPC protocol between M and REG to train a fair model $f_{\boldsymbol{\theta}}$ privately. This setup is shown in Figure 1 (*Center*).

Privacy constraints: (C1) privacy of sensitive user data, (C2) model secrecy, and (C3) minimal disclosure of $\mathbf{D}$ to M.

Decision Verification. We aim to avoid that a malicious M swaps a successfully certified model $f_{\boldsymbol{\theta}}$ for a potentially unfair model $f_{\boldsymbol{\theta}'}$ in the real world. Upon receiving a decision $\hat{y}$, a user can challenge it by asking REG for a verification. The verification consists of M and REG jointly verifying that $f_{\boldsymbol{\theta}'}(\mathbf{x}) = f_{\boldsymbol{\theta}}(\mathbf{x})$, where $\mathbf{x}$ are the user's non-sensitive features. While there is no simple technical way to prevent a malicious M from deploying an unfair model, they will get caught if a user challenges a decision that would differ under $f_{\boldsymbol{\theta}}$. This setup is shown in Figure 1 (*Right*).

Privacy constraints: (C1) privacy of sensitive user data, and (C2) model secrecy.

Design Choices. We use a regulator for several reasons. While in principle MPC can be carried out without a regulator, using all users as parties, this comes at a significantly greater computational cost and the requirement of all users being online simultaneously. Additionally, given that industry has traditionally resisted privacy-enhancing technologies (Brown, 2014), a coordinating body has been highlighted as practically important in this area (The Royal Society and the British Academy, 2017).

If users are uncomfortable sharing $\mathbf{D}$ in the clear, we can keep all of $\mathbf{x}_i, y_i, \mathbf{z}_i$ private throughout the three tasks using recent MPC methods, with computational cost increasing only by a factor of 2 (Mohassel & Zhang, 2017). While this extension would sometimes be desirable, it hinders exploratory data analysis by the modeler and validation by the regulator that user-provided data is correct. In this work, privacy or secrecy constraints are separate from other attacks such as model extraction (Tramèr et al., 2016) or inversion (Fredrikson et al., 2015). If relevant, modelers may need to consider these separately.

3. Our MPC Solution

Fair Multi-Party Machine Learning. MPC protocols allow two parties P_1 and P_2 holding secret values x_1 and x_2 to evaluate a public function f such that the parties (either both or one of them) learn *only* $f(x_1, x_2)$. In our setting, f will be model training, verification, or certification and the parties involved are the modeler M and the regulator REG.

Because generic solutions do not yet scale to real-world tasks, custom tailored protocols have proved successful, for example in logistic and linear regression (Nikolaenko et al., 2013b; Gascón et al., 2017; Mohassel & Zhang, 2017), neural network training (Mohassel & Zhang, 2017) and evaluation (Juvekar et al., 2018; Liu et al., 2017), matrix factorization (Nikolaenko et al., 2013a), and principal component analysis (Al-Rubaie et al., 2017).

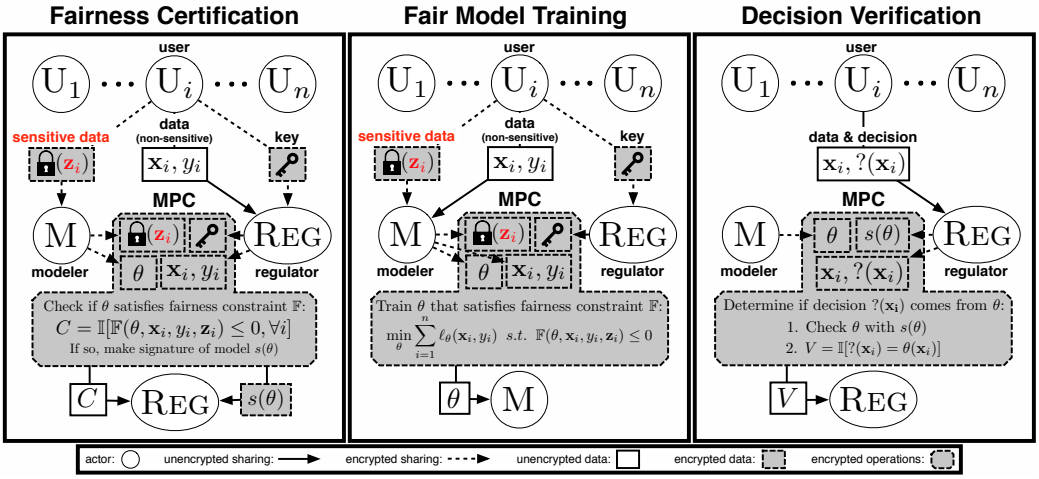

Figure 1. Our setup for *Fairness certification* (*Left*), *Fair model training* (*Center*), and *Decision verification* (*Right*).

All protocols consist of the two parties jointly evaluating arithmetic and/or Boolean circuits gate by gate. Because *floating point arithmetic* cannot be represented by practical circuits (see Demmler et al., 2015, Table 4), we are limited to *fixed-point arithmetic* and need to approximate non-linear functions by (piecewise) linear ones. We use *additive secret sharing* to keep intermediate values hidden from other parties, i.e., z lives in $\mathbb{Z}_q$ with $q = 2^{64}$. To share z, a user draws r uniformly at random from $\mathbb{Z}_q$, sends r to M and $z - r$ to REG. None of these values reveals anything about z, but z can easily be reconstructed within the MPC computation. In Figure 1, we thus reinterpret the key held by REG and the encrypted z by M as their corresponding shares.

For *fair training* we follow the techniques introduced by Mohassel & Zhang (2017) and extend them to handle linear constraints. This extension may be of independent interest, and has applications for privacy-preserving machine learning beyond fairness. *Certification* and *verification* partly correspond to sub-procedures of the *fair training* task, hence do not add technical difficulties. For certification, we check that $\boldsymbol{\theta}$ satisfies $\mathbb{F}$ in MPC, followed by computing a signature $s(\boldsymbol{\theta})$. We use cryptographic hash functions such as SHA-256 that can be evaluated quickly in MPC for $s(\boldsymbol{\theta})$ (see Keller et al., 2013, Figure 14). For verification, we compute the signature of the model provided by M and proceed only if it matches $s(\boldsymbol{\theta})$. We believe this application to machine learning model certification is novel.

The optimization problem for fair learning is to minimize some classification loss $\mathcal{L}(\mathbf{X}, \mathbf{y}, \boldsymbol{\theta})$ subject to a (often convex) fairness constraint $\mathbb{F}(\boldsymbol{\theta}) \leq 0$. Zafar et al. (2017) use a convex approximation of the $p\%$-rule, see eq. (1), for linear classifiers to derive the constraint $\mathbb{F}(\boldsymbol{\theta}) = \frac{1}{n}|\hat{\mathbf{Z}}^\top \mathbf{X}\boldsymbol{\theta}| - \mathbf{c}$, where $\hat{\mathbf{Z}}$ is the matrix of all $\hat{\mathbf{z}}_i := \mathbf{z}_i - \bar{\mathbf{z}}$ and $\mathbf{c} \in \mathbb{R}^d$ is a constant vector corresponding to the tightness of the fairness constraint. Here, $\bar{\mathbf{z}}$ is the mean of all inputs $\mathbf{z}_i$. With $\mathbf{A} := {}^1\!/_n \hat{\mathbf{Z}}^\top \mathbf{X}$, the $p\%$ *constraint* reads $\mathbb{F}(\boldsymbol{\theta}) = |\mathbf{A}\boldsymbol{\theta}| - \mathbf{c}$, where the absolute value is taken element-wise.

Table 1. Dataset sizes and online timing results of MPC certification and training over 10 epochs with batch size 64.

	Adult	Bank	COMPAS	German	SQF
n examples	2^{14}	2^{15}	2^{12}	2^{9}	2^{16}
d features	51	62	7	24	23
p sensitive attr.	1	1	7	1	1
certification	802 ms	827 ms	288 ms	250 ms	765 ms
training	43 min	51 min	7 min	1 min	111 min

Technical Challenges. Zafar et al. (2017) use Sequential Least Squares Programming (SLSQP), which requires solving a sequence of quadratic programs and non-integer divisions by non-constant numbers. These operations are currently infeasible within MPC.

Instead, we run stochastic gradient descent and use Lagrangian multipliers, which performed consistently better than projected gradient descent and interior point log barrier (Boyd & Vandenberghe, 2004) in our experiments. Hence, we minimize $\mathcal{L} := \mathcal{L}^{\text{BCE}}(\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) + \boldsymbol{\lambda}^\top \max\{\mathbb{F}(\boldsymbol{\theta}), \mathbf{0}\}$ for the standard logistic regression loss $\mathcal{L}^{\text{BCE}}$ with alternating updates $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \mathcal{L}$ and $\boldsymbol{\lambda} \leftarrow \max\{\boldsymbol{\lambda} + \eta_{\boldsymbol{\lambda}} \nabla_{\boldsymbol{\lambda}} \mathcal{L}, \mathbf{0}\}$, where $\eta_{\boldsymbol{\theta}}, \eta_{\boldsymbol{\lambda}}$ are learning rates.

The gradients only require matrix multiplications and a single evaluation of the logistic function, which we approximate by 0 and 1 for $x < -0.5$ and $x > 0.5$ respectively, and linear in between (Mohassel & Zhang, 2017). The largest number representable in fixed-point format with m integer and m fractional bits is roughly $2^m + 1$. Since we whiten the features $\mathbf{X}$ column-wise, we need to be careful whenever we add more than 2^m numbers, limiting the minibatch size. For large n, evaluating $\hat{\mathbf{Z}}^\top \mathbf{X}/n$ in the fairness function $\mathbb{F}$ is particularly problematic. To avoid under- and overflows, we perform it in blocks of size $b < 2^m$, divide each block by b, and multiply the sum of all blocks by ${}^b\!/_n > 2^{-m}$. Moreover, using powers of two for n and the minibatch size, we can use fast bit shifts to avoid prohibitively expensive divisions.

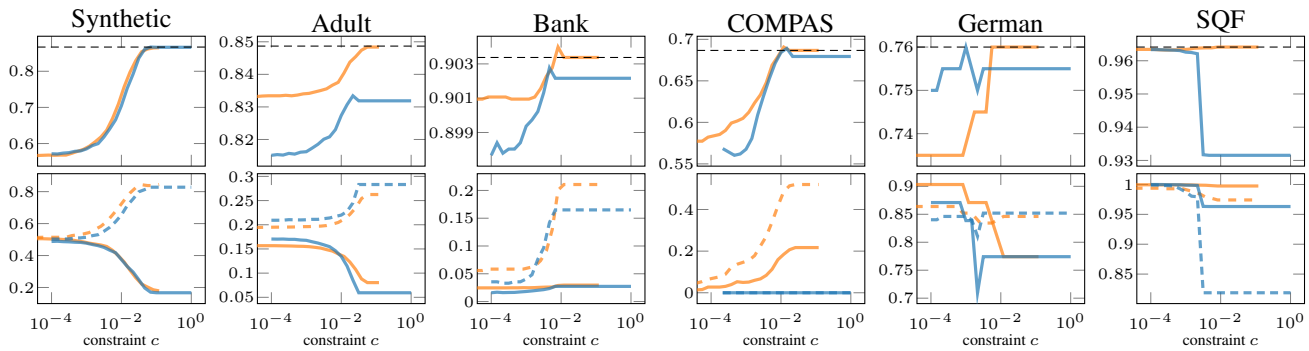

Figure 2. **First row:** Test set accuracy of our method (blue) and the SLSQP baseline (orange) (with approximations and fixed point). The dashed line is unconstrained logistic regression (from scikit-learn). **Second row:** The fraction of people with $z = 0$ (*continuous*) and $z = 1$ (*dashed*) who get assigned positive outcomes (blue: our method, orange: SLSQP baseline).

4. Experiments

We now show that our approach is feasible for all three tasks on 5 real-world datasets, namely the Adult, Bank, and German datasets from the UCI machine learning repository (Lichman, 2013), the stop, question and frisk 2012 dataset (SQF),[1] and the COMPAS dataset (Angwin et al., 2016), see Table 1. Moreover, we also run on synthetic data, generated as described by Zafar et al. (2017), because it allows us to control the correlation between the sensitive attributes and the class labels. We compare to SLSQP as a baseline. Our constraint value covers the range $[10^{-4}, 10^{0}]$ and a corresponding range for SLSQP.

Accuracy and Fairness. The first row of Figure 2 shows test set accuracies over the fairness constraint (smaller is tighter). Like the synthetic dataset, Adult, Bank, and COMPAS exhibit a visible trade-off between accuracy and fairness. The German dataset contains only 512 training examples, which explains the varying and seemingly discrete accuracy. For SQF, accuracy slightly improves as the constraints become active. Further investigation is needed to determine the cause of this behavior. We suspect the constraint to act as a regularizer for SGD. Despite fixed point arithmetic, approximate non-linearities, and SGD, our method retains comparable accuracy across all datasets.

The second row of Figure 2 shows how well disparate impact is mitigated as the fractions of users with positive outcomes in each group gradually approach as we decrease c, i.e., increase p. The effect is most pronounced for the first three datasets. The COMPAS dataset has 7 sensitive attributes (some with only 10 positive instances) quickly leading to infeasible constraints. Consequently, the $p\%$-rule needs careful interpretation when applied for multiple sensitive attributes. A random sensitive attribute is chosen in the second row for COMPAS. Variations in the German dataset are due to its small size. In COMPAS, SQF, and arguably in Bank, the classifier tends to collapse to negative or positive outcomes as c decreases, i.e., we require fractions to be balanced exactly. Our method is competitive in removing disparate impact while retaining high accuracy on all but the challenging COMPAS dataset.

Runtime. In Table 1 we show the online running times on a laptop computer. While training takes orders of magnitudes longer than a non-MPC implementation, our approach still remains feasible and realistic. We use the offline precomputation of multiplication triples as described and timed in Mohassel & Zhang (2017, Table 2). Certification consists of checking $\mathbb{F}(\boldsymbol{\theta}) > 0$, which is already done for each gradient update during training and only takes negligible computation time, see Table 1. Similarly, the operations required for verification stay well below one second.

5. Conclusion

Real world fair learning suffers from a dilemma: in order to enforce fairness, sensitive attributes must be examined; yet in many situations, users may feel uncomfortable in revealing these attributes, or modelers may be legally restricted in collecting and utilizing them. By introducing and extending recent methods from MPC, we have demonstrated that it is practical for certain outcome-based notions of group fairness on real-world datasets to: (i) certify and sign a model as fair; (ii) learn a fair model; and (iii) verify that a fair-certified model has indeed been used; all while maintaining cryptographic privacy of all users' sensitive attributes. Connecting concerns in privacy, algorithmic fairness and accountability, our proposal reduces the barrier for regulators to provide better oversight, modelers to develop fair and private models, and users to retain control over data they consider highly sensitive. Issues of social fairness are complex. We propose our approach as one tool which may be useful in some settings to mitigate concerns around machine learning and society.

[1] https://perma.cc/6CSM-N7AQ

Acknowledgments

The authors would like to thank Chris Russell and Phillipp Schoppmann for useful discussions and help with the implementation, as well as the anonymous reviewers for helpful comments. AG and MK were supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. MV was supported by EPSRC grant EP/M507970/1. AW acknowledges support from the David MacKay Newton research fellowship at Darwin College, The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074, and the Leverhulme Trust via the CFI.

References

Al-Rubaie, M., Wu, P. Y., Chang, J. M., and Kung, S. Privacy-preserving PCA on horizontally-partitioned data. In *DSC*. IEEE, 2017.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: There is software used across the country to predict future criminals. and it is biased against blacks. *ProPublica*, 2016.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *California Law Review*, 104, 2016.

Bennett Capers, I. Blind justice. *Yale Journal of Law & Humanities*, 24:179, 2012.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Brown, I. Britain's smart meter programme: A case study in privacy by design. *International Review of Law, Computers & Technology*, 2014.

Calders, T. and Žliobaitė, I. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society*. 2012.

Demmler, D., Dessouky, G., Koushanfar, F., Sadeghi, A., Schneider, T., and Zeitouni, S. Automated synthesis of optimized circuits for secure computation. In *ACM Conference on Computer and Communications Security*. ACM, 2015.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *ITCS*, ITCS '12, 2012.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *SIGKDD*, 2015.

Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *SIGSAC*, 2015.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. On the (im)possibility of fairness. *arXiv:1609.07236*, 2016.

Gascón, A., Schoppmann, P., Balle, B., Raykova, M., Doerner, J., Zahur, S., and Evans, D. Privacy-Preserving Distributed Linear Regression on High-Dimensional Data. *Privacy Enhancing Technologies*, 2017, 2017.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., and Weller, A. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *AAAI*, 2018.

Handel, P., Skog, I., Wahlstrom, J., Bonawiede, F., Welch, R., Ohlsson, J., and Ohlsson, M. Insurance telematics: Opportunities and challenges with the smartphone solution. *IEEE Intelligent Transportation Systems Magazine*, 2014.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NIPS*, 2016.

Juvekar, C., Vaikuntanathan, V., and Chandrakasan, A. Gazelle: A Low Latency Framework for Secure Neural Network Inference. *IACR Cryptology ePrint Archive*, 2018, 2018.

Keller, M., Scholl, P., and Smart, N. P. An architecture for practical actively secure MPC with dishonest majority. In *ACM Conference on Computer and Communications Security*, 2013.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. Accountable algorithms. *University of Pennsylvania Law Review*, 165, 2016.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Lipton, Z. C., Chouldechova, A., and McAuley, J. Does mitigating ml's disparate impact require disparate treatment? *arXiv:1711.07076*, 2017.

Liu, J., Juuti, M., Lu, Y., and Asokan, N. Oblivious neural network predictions via minionn transformations. In *CCS*. ACM, 2017.

Mohassel, P. and Zhang, Y. SecureML: A system for scalable privacy-preserving machine learning. In *IEEE Symposium on Security and Privacy (SP)*, 2017.

Nikolaenko, V., Ioannidis, S., Weinsberg, U., Joye, M., Taft, N., and Boneh, D. Privacy-preserving matrix factorization. In *Conference on Computer and Communications Security*, 2013a.

Nikolaenko, V., Weinsberg, U., Ioannidis, S., Joye, M., Boneh, D., and Taft, N. Privacy-preserving ridge regression on hundreds of millions of records. In *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2013b.

Schreurs, W., Hildebrandt, M., Kindt, E., and Vanfleteren, M. *Cogitas, Ergo Sum*. The Role of Data Protection Law and Non-discrimination Law in Group Profiling in the Private Sector. In *Profiling the European Citizen*. Springer, 2008.

The Royal Society and the British Academy. Data management and use: Governance in the 21st Century, 2017.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., and Ristenpart, T. Stealing machine learning models via prediction apis. In *USENIX Security Symposium*, 2016.

Veale, M. and Binns, R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2), 2017.

Veale, M. and Edwards, L. Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling. *Computer Law & Security Review*, 2018. doi: 10.1016/j.clsr.2017.12.002.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*, 2017.

Žliobaitė, I. and Custers, B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 2016.