Legally grounded fairness objectives

Dylan Holden-Sim [1] **Gavin Leech** [1] **Laurence Aitchison** [1]

Abstract

Recent work has identified a number of formally incompatible operational measures for the unfairness of a machine learning (ML) system. As these measures all capture intuitively desirable aspects of a fair system, choosing "the one true" measure is not possible, and instead a reasonable approach is to minimize a weighted combination of measures. However, this simply raises the question of how to choose the weights. Here, we formulate Legally Grounded Fairness Objectives (LGFO), which uses signals from the legal system to non-arbitrarily measure the social cost of a specific degree of unfairness. The LGFO is the expected damages under a putative lawsuit that might be awarded to those who were wrongly classified, in the sense that the ML system made a decision different to that which would have be made under the court's preferred measure. Notably, the two quantities necessary to compute the LGFO, the court's preferences about fairness measures, and the expected damages, are unknown but well-defined, and can be estimated by legal advice. Further, as the damages awarded by the legal system are designed to measure and compensate for the harm caused to an individual by an unfair classification, the LGFO aligns closely with society's estimate of the social cost.

1. Introduction

Automated decision making systems have not only become more prevalent, but are also being applied in increasingly sensitive contexts (Kamiran & Calders, 2009; Kamishima & Asoh, 2012; Cabitza et al., 2017; Gronlund, 2019). This has led to demand for more transparent systems, as well as tools for assuring fairness. A key result here is that systems can discriminate on protected characteristics such as race, religion and gender, even when protected attributes are not an input to the system (Žliobaitė & Custers, 2016; Veale & Binns, 2017).

The fundamental issue of algorithmic fairness is that no single definition of fairness captures the full phenomenon. Famously, the COMPAS recidivism prediction system was used in the criminal justice process in several US states. A ProPublica study argued that the COMPAS system was racially discriminatory, finding that African Americans labeled 'High risk' were in fact 50% less likely to reoffend than white defendants with the same label (Larson et al., 2016). Flores et al responded by arguing that defendants with the same COMPAS recidivism score had approximately equal probability of recidivism (Flores et al., 2016). Here we see a direct clash between operationalisations of fairness.

Attempts to formalise fairness have yielded many such reasonable definitions (Hardt et al., 2016; Barocas et al., 2019; Dwork et al., 2012); ideally, we would fulfill these simultaneously. But we can prove that some sets of commonsensical definitions are incompatible, outside trivial cases (Kleinberg et al., 2016; A.Chouldechova, 2017).

Since perfect multi-measure fairness is almost always impossible, we instead aim at systems which minimise violations. But it is unclear which fairness definitions to relax and to what extent: subjective decisions about relative importance are required. While it is clear that we cannot leave this task to the system implementors alone, the problem is actually far worse: as it is a question of values there may be irreconcilable differences between different individuals and there is no underlying well-defined but perhaps unknown "correct answer". Without a well-defined correct answer even in principle, what weights should we pick? We note that society can and must answer such question in other contexts using the legal system. As such, we propose using the legal system to operationalise society's estimate of social costs. The resulting Legally Grounded Fairness Objectives (LGFO) measure the damages awarded to those who were wrongly classified by the ML system under a putative lawsuit.

Minimising case cost as maximising social welfare

Our solution is to find the classifier which minimises the damages to people classified differently under different definitions. That is, we minimise the *legal* cost of choosing one measure over the other. This shifts the burden of selecting fairness measures onto the legal system and away from the programmer.

Our contention is that the social cost of an unfair classifier can be measured by the expected damages awarded to an individual given a false classification.

On first glance, setting the objective to minimised legal penalties looks inappropriate: as if privileging the interests of the system deployer. However, it is reasonable to view the size of legal damages is a proxy for social good, since: 1) in principle, the law is designed to reflect the values of a society, including the broadest reading of fairness. 2) legal damages are intended to reimburse an individual for harm caused, as assessed by a judge. Thus, by minimising total damages (for instance by reducing unfairness and so the number of associated lawsuits) we simultaneously minimise harm.

A key advantage of this perspective is that we make use of the canonical process for balancing values and estimating social costs: the law, in this case civil law. While we should expect persistent disagreement about the nature of the social good, the legal system is the working mechanism society uses to approximate it, when informal means fail. In well-functioning jurisdictions, the legal process has a degree of public accountability, adaptiveness, and consensus - or anyway more than an average IT department (Burri, 2016; Israni, 2017).

A second advantage is the relative availability of high-quality data. Our training signal is the monetary damages awarded to the plaintiff in algorithmic discrimination cases; in many jurisdictions, this data is openly available, e.g. (BAILII, 2020). We also need to elicit expert legal opinion on the type of fairness most applicable (or most often applied) in particular contexts, and on how much it would cost in a given case if a plaintiff's classification was changed. Given these, we can minimise a weighted combination of unfairness measures.

Related work

CFAθ (Zehlike et al., 2019) is a fairness algorithm used to map distributions of raw scores towards the barycenter, a distribution occupying "middle ground" between the distributions of the different groups. The algorithm takes parameter θ which gives the degree of the mapping. $\theta = 0$ leaves the raw scores unchanged, whereas $\theta = 1$ sets all group distributions equal to the barycentre. It thus operationalises the tradeoff between individual fairness (low θ) and group fairness (high θ). A value $0 < \theta < 1$ corresponds to a partial mapping of group distributions towards this barycenter. θ is normative: its tuning would ideally be left to some democratic process. The issue is that selecting an appropriate θ requires a nuanced understanding of the algorithm by the decision maker, and the value is still a decision rather than calculated based on concrete values (i.e. legal costs).

The method proposed in (Dwork et al., 2012) encapsulates the idea that 'similar people should be treated similarly', a view known as Individual Fairness. This is achieved by enforcing a Lipschitz condition on the classifier: For any two individuals x, y at a distance $d(x, y) \in [0, 1]$ and map to distributions $M(x)$ and $M(y)$ respectively, the statistical distance between $M(x)$ and $M(y)$ is at most $d(x, y)$. Or $D(M(x), M(y)) \leq d(x, y)$.

This is intuitive: the difference in group outcomes should be less than or equal to the difference in individuals. This method is effective if domain knowledge can be used in constructing the distance function, i.e. if the normative work can be shared by other parties. But we have only shifted the subjectivity problem onto the distance function: whoever is given the task of defining d still has to work in the absence of well-defined, unambiguous standards (Kim et al., 2018).

Our contributions

We propose a new perspective in algorithmic fairness, using legal costs as a proxy for the social cost of a given fairness measure.

We define a method to account for multiple fairness measures and give an overall degree of unfairness, allowing for fairness maximisation.

We report experiments on a real-world dataset, showing that fairness measure combinations can correct naive correlations between the response variable and protected attributes.

2. Methods

Our algorithm is a post-processing step for binary classifiers. We find the *cost-minimal* decision boundary for each group: the pair (t_0, t_1) where t_i denotes the decision boundary (i.e. threshold value) for group i.

The fairness measures we use are initially binary properties: either satisfied perfectly or violated perfectly. To find decision boundaries which maximise a given fairness definition, we translate these notions into measures: functions of outcomes which are minimal at 0 (where the property is perfectly satisfied) and increase as we deviate further from the definition.

Unfairness as cost measure

There are many proposed measures; here we focus on three, namely Sufficiency, Equalised Odds and Statistical Parity (Barocas et al., 2019; Hardt et al., 2016; Dwork et al., 2012). In (Kleinberg et al., 2016) it was proven that these measures are mutually incompatible outside of trivial cases - we cannot satisfy all three simultaneously (see Supplement, Proof 1). They include a positive and negative case, but we use only the positive case. Let G denote the group of the defendants (here, a boolean for ethnicity), Y be the ground truth label (here, actual recidivism risk category) and $\hat{Y}$ be

the classifier's predicted label; let 1 denote the high risk category and 0 low risk.

The sufficiency of a classifier (Suff) involves the difference in precision between groups (i.e. the probability of positive ground truth, given a positive prediction):
$\mathrm{Prec} = P(Y = 1 \mid \hat{Y} = 1)$ and

$$\mathrm{Suff} = |\,\mathrm{Prec}_{\,G=0} - \mathrm{Prec}_{\,G=1}\,|$$

Violation of Suff means that a positive prediction is more reliable for one group: and if positive classifications are less reliable for one group, then they cannot be used naively for decisions.

The Equalised Odds measure (ΔF) involves the difference in false positive rate between groups:
$\mathrm{FPR} = P(\hat{Y} = 1 \mid Y = 0)$ and

$$\Delta\mathrm{F} = |\,\mathrm{FPR}_{\,G=0} - \mathrm{FPR}_{\,G=1}\,|$$

Violating ΔF means we are more likely to wrongly predict that one group will reoffend than another group. This was the allegation in the ProPublica analysis: African Americans were more likely to be incorrectly labelled 'high risk' than white Americans (Larson et al., 2016).

Finally, Statistical Parity (SP) involves the difference between groups in the probability of predicting a positive label:

$$\mathrm{SP} = |\,P(\hat{Y} = 1 \mid G = 0) - P(\hat{Y} = 1 \mid G = 1)\,|$$

(Note that these are really *unfairness* measures: that is, higher values indicate greater differences in handling different groups.)

Legally grounded fairness objectives

Using any set of fairness measures M and a set of example cases X, we can define the LGFO, the expected damages resulting from a hypothetical civil suit for wrongful classification:

$$\mathrm{LGFO} = \sum_{m \in M} P(m) \sum_{x \in X} C(\hat{y}, y_m)$$

where $\hat{y} = c(x)$ is the decision originally made by the ML system, y_m is the decision that would have been made under fairness measure m, $P(m)$ is the probability that the court prefers that measure, and $C(y, y_m)$ is the misclassification cost of y according to m.

The LGFO Algorithm

Choose a particularly simple approach to minimizing the LGFO: we fix the classifer and modify group-dependent thresholds (Algorithm 1). This finds a separate score threshold for each group, such that the thresholds minimise overall multi-measure cost.

Let the expected legal cost of changing an outcome from positive to negative be P2N and the cost of changing from negative to positive be N2P. Let $\mathbf{X}$ be the set of all inputs to the classifier. Let $\hat{y}_i(x) \in \{0, 1\}$ be the predicted label for x after the raw score is thresholded by s_i, which is a tuple (t_0, t_1) of per-group thresholds.

Let P^* be a target number of positive classifications, which we set in order to avoid trivially fair cases (such as classifying all inputs as positive). Ideally we would consider cases where we exactly achieve P^* positives; in practice this is not always possible.

Let M be the set of fairness measures to balance, $\mathbf{C}_m$ be the set of costs incurred by applying measure m alone over each threshold $s \in S$. The misclassification cost (of a threshold pair s_i relative to the best threshold pair s_j) O is, for example x:

$$O(x, s_i, s_j) = \begin{cases} \text{P2N}, & \text{if } \hat{y}_i(x) - \hat{y}_j(x) = 1 \\ \text{N2P}, & \text{if } \hat{y}_i(x) - \hat{y}_j(x) = -1 \\ 0, & \text{otherwise} \end{cases}$$

The output is the threshold pair which gives the minimum summed cost C_{sum}; that is, the lowest cost we can obtain under all measures.

We validated LGFO using the COMPAS dataset (Bellamy et al., 2018; Larson et al., 2016): we implemented a PyTorch binary classifier predicting the probability of belonging to the 'High chance of violent recidivism' class. The original COMPAS system used a scoring system; our approach mirrors ProPublica in merging Medium and High risk categories and constructing a binary classifier for this new group.

2.1. Cost sensitivity

A property that then naturally arises is *cost-sensitivity*. Consider a measure cost-sensitive if it leads to large changes in cost for small changes in absolute measure value.

Cost-insensitive measures can be relaxed to a much greater degree without incurring large social cost. This provides an opportunity to improve on other measures which are more sensitive to cost, leading to an output that is more fair under more definitions.

3. Results

Figure 1 compares the raw values of our chosen fairness measures at different thresholds; Figure 2 shows the cost of violating the fairness measure. Figure 3 then shows the

Algorithm 1 Minimizing LGFO

Input: M : set of fairness measures to balance,
$\mathbf{X}$: examples,
$\hat{\mathbf{Y}}$: classifier scores, $c(x)\ \forall x \in X$
P^*: target number of positives

S = `get_thresholds` $(\mathbf{X}, \hat{\mathbf{Y}}, P^*)$
$\mathbf{C}_m = [\,]$
forall $m \in M$ **do**
  $s_{\min} = \underset{s \in S}{\operatorname{argmin}}(m(s))$
  forall $s_i \in S,\ s_i \neq s_{\min}$ **do**
    $C = \sum_{x \in X} O(x, s_i, s_{\min})$
    $\mathbf{C}_m[s_i] = C$
forall $s_i \in S$ **do**
  $\mathbf{C}_{\text{sum}}[s_i] = \sum_{m \in M} \mathbf{C}_m[s_i]$
return $\underset{s \in S}{\operatorname{argmin}}(\mathbf{C}_{\text{sum}})$

Procedure `get_thresholds` $(\mathbf{X}, \hat{\mathbf{Y}}, P^*)$
  $S = [\,]$
  forall $t \in [0, 0.02, ..., 1]$ **do**
    $n_{p'},\ n_{q'} = \infty$
    forall $t_0,\ t_1 \in [0, 0.02, ..., 1]$ **do**
      $s_p = (t, t_1)$
      $s_q = (t_0, t)$
      $n_p = \sum_{x \in \mathbf{X}}\ \hat{y}_p(x)$
      $n_q = \sum_{x \in \mathbf{X}}\ \hat{y}_q(x)$
      if $|n_p - P^*| < n_{p'}$ **then**
        $n_{p'} = n_p$
        $s_{p'} = s_p$
      if $|n_q - P^*| < n_{q'}$ **then**
        $n_{q'} = n_q$
        $s_{p'} = s_q$
    append $s_{p'}$ to S
    append $s_{q'}$ to S
  return S

summed cost which yields the minimal-cost fair configuration.

It is helpful to visualise costs as a curve by ordering threshold pairs from highly preferential treatment for one group to highly preferential treatment for the other, with the midpoint being equal treatment for both groups. The 'Threshold pair index' then represents the index of this ordered collection.

Fairer COMPAS predictions

In Figure 1 we see that SP and ΔF roughly agree on the fairest region. This is because both encapsulate a similar notion of fairness, penalising discrepancies between group outcomes. The cost-optimal thresholds found are [0.54,0.41], corresponding to slightly favourable treatment for African Americans.

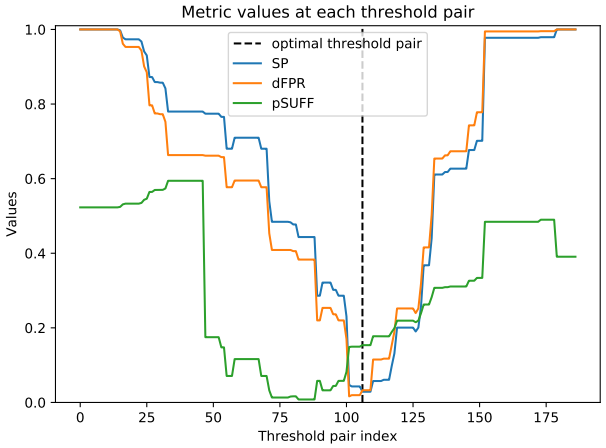

Figure 1. Unfairness values over threshold choices.

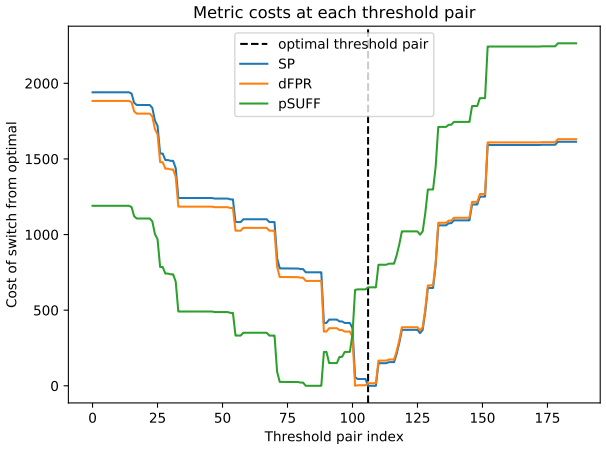

Figure 2. Costs of deviating from the optimal configuration of each measure in LGFO.

LFGO vs uncorrected classification

To evaluate our algorithm, we compare the LGFO classification to the uncorrected classification (equivalent to the threshold pair $[0.5, 0.5]$).

We see our corrected model makes the trade off of Sufficiency for Statistical Parity and Equalised Odds. There is also a small accuracy decrease of 2%. Accuracy is maintained, since LGFO mostly changes classifications only for defendants receiving uncertain predictions. Inputs with predictive values close to 0 or 1 will only see changes to their outcomes in extreme decision boundaries, i.e. those that are

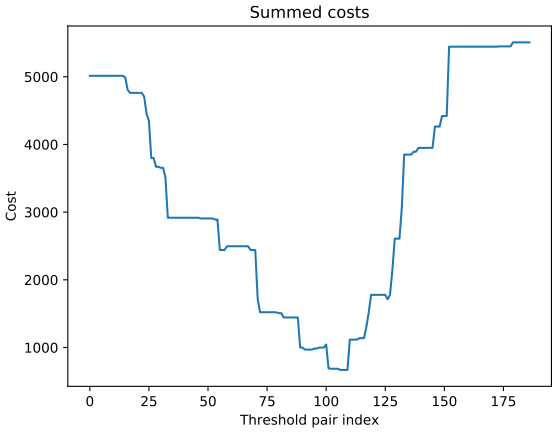

Figure 3. Summed individual measure costs. The minima is our cost-optimality configuration. This aligns with the optimal for ΔF.

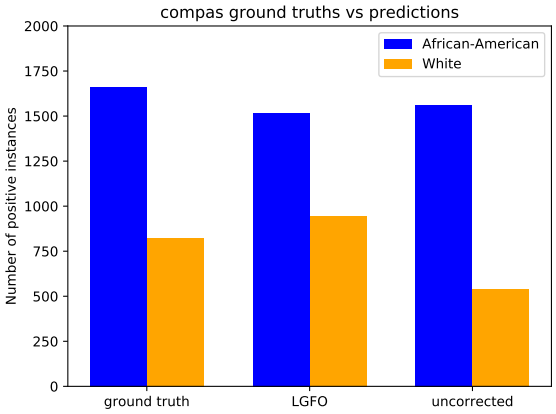

Figure 4. Comparing corrected and uncorrected per-group positive predictions to ground truth.

Table 1. measure results for the uncorrected and LGFO classifiers

MEASURE	UNCORRECTED	LGFO
STATISTICAL PARITY	0.236	**0.029**
SUFF	**0.062**	0.154
ΔF	0.162	**0.033**
ACCURACY	**67.0%**	65.7%

unfair by our measures. Our algorithm only changes outcomes for a fraction of individuals, those the raw classifier is more likely to mislabel.

Looking at Figure 4, we see that the Uncorrected predictor significantly under-represents white plaintiffs in positive predictions versus the ground truth data. LGFO corrects for this, bringing the number of predictions closer to the ground truth.

Cost-sensitivity

We see that Suff is the most cost-sensitive measure: we see the same cost incurred when Suff $= 0.2$ as we do when $\Delta F = 0.5$. This explains the result in Table 1, in which Sufficiency actually decreases after applying LGFO: this is a principled trade for greatly reduced unfairness on the other measures. The key result is that inspecting the measure values themselves is insufficient to judge the actual relative fairness of two classifiers: taking the damages into account shows that lower measure fairness can occur when damages are reduced.

Illustrative scenarios

We now investigate LFGO with counterfactual scenarios.

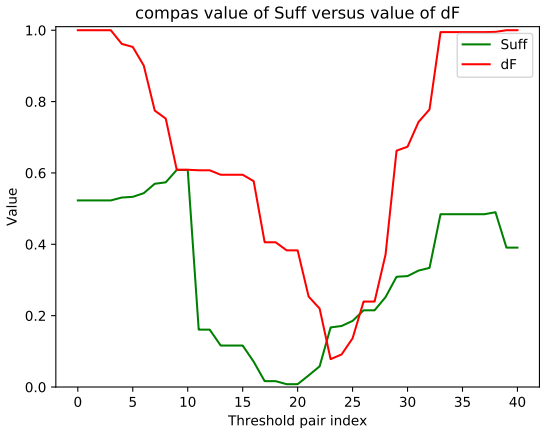

Figure 5. Value of the measures at each threshold pair.

INTERMEDIATE FAIRNESS

In place of true legal costs, this scenario sets P2N $= 0$ and N2P $= 1$. This corresponds to the cost of changing a prediction to 'highly likely to reoffend' (a false positive) being higher than the converse (false negative).

Next compare the cost of violating a measure to the degree of violation (C_m vs $m(X)$). In Figure 5 we see that to minimise Suff (around $x = 20$), we incur high ΔF; but conversely a very low ΔF (around $x = 23$) results in moderate Suff. When comparing the costs, however (Figure 6), we see that costs are actually equal for both thresholds.

Figure 7 shows the minimum summed-cost point. This cost-minimal configuration occurs in-between the optimal Suff or ΔF, corresponding to partial unfairness on both accounts; but this simultaneous relaxation yields a better

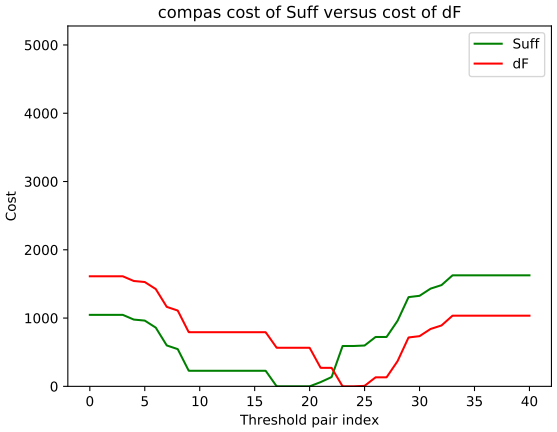

Figure 6. Costs of deviating from the optimal configuration for each measure.

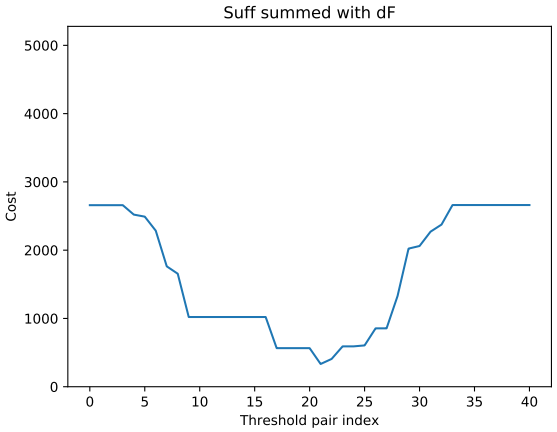

Figure 7. Summed individual measure costs. The minimum corresponds to our cost-optimal fair configuration - a trade off between both measures.

social outcome than optimising for either alone.

SINGLE-TYPE FAIRNESS

Figure 8 is from a scenario with $\mathrm{P2N} = 1$ and $\mathrm{N2P} = 0$. Intermediate thresholds yield higher cost than optimising for either measure individually. The local maximum occurs at the same location as the minimum in Figure 7, which implies that while the measure values are reasonable, we actually cause more harm in trying to balance both than when a single measure is used.

4. Discussion

LGFO has several virtues: it brings fairness into business operations, by setting an unambiguous, hard-to-game monetary incentive towards fair systems. It also allows for stakeholders other than the technical team to contribute to the system design, and makes use of long-standing legal expertise on decision-making in complex social situations.

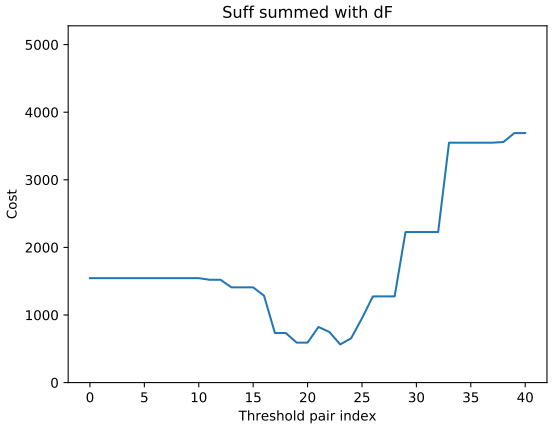

Figure 8. Summed individual measure costs. Here cost-optimality occurs under optimisation of a single measure. In this case partial satisfaction of both measures yields a higher cost.

We found that our algorithm was able to correct for erroneous bias in a neural classifier using a real-world dataset, while conserving performance. Fairness algorithms should remain performance-competitive, to make it more likely that they are actually implemented. On our dataset and classifier we noted a small (2% relative) loss of overall accuracy from applying LGFO; however, as shown in Figure 4, this minor performance cost impacts each group differently.

Limitations

Our approach cannot be applied immediately to arbitrary data, but requires a careful elicitation step. This is due to the lack of explicit use of formal fairness measures in most legal systems (Xiang & Raji, 2019).

In the present experiments, we use ProPublica's binarised risk groups, which does not reflect the actual use of deployed systems, nor the finer-grained information in the original scores.

We only handle binary classification; however, extensions of the method to other learning settings is a possible area for future experimentation. We could also extend the method to other stages of the ML pipeline. The LGFO value could, for instance, be used in the training stage of the classifier.

Clearly, the legal system is also an imperfect estimator of social cost, and has its own biases (Zamir & Ritov, 2012; Berrey et al., 2012). But it seems unlikely to be more biased

(or unaccountable) than a lone technical team with no clear incentives towards fairness. LGFO ties ML into an existing democratic process with greater domain knowledge of the tradeoffs involved.

The LGFO estimation process is unavoidably local: the distribution over fairness definitions, and the damages involved, will vary greatly between jurisdictions. But this is just the converse of the method's strength: that it makes use of actual domain knowledge.

References

A.Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2017.

BAILII. BAILII databases. https://www.bailii.org/databases.html, 2020.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and machine learning: Limitations and Opportunities.* 2019. URL https://fairmlbook.org.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. https://github.com/IBM/AIF360, 2018.

Berrey, E., Hoffman, S. G., and Nielsen, L. B. Situated justice: A contextual analysis of fairness and inequality in employment discrimination litigation. *Law & Society Review*, 46(1):1–36, 2012.

Burri, T. Machine learning and the law: Five theses. In *Proceedings of the NIPS 2016 Workshop on Machine Learning and the Law*, 2016.

Cabitza, F., Rasoini, R., and Gensini, G. F. Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518, 2017.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. *Innovations in Theoretical Computer Science*, 2012.

Flores, A. W., Lowenkamp, C. T., and Bechtel, K. False positives, false negatives, and false analyses: A rejoinder. http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf, 2016.

Gronlund, K. State of ai: Artificial intelligence, the military and increasingly autonomous weapons. https://futureoflife.org/2019/05/09/state-of-ai/, 2019.

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

Israni, E. Algorithmic due process: Mistaken accountability and attribution in state v. loomis. https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1, 2017.

Kamiran, F. and Calders, T. Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication*, pp. 1–6, 2009.

Kamishima, A. and Asoh, S. Fairness-aware classifier with prejudice remover regularizer. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012.

Kim, M. P., Reingold, O., and Rothblum, G. N. Fairness through computationally-bounded awareness, 2018.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores, 2016.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm, 2016. URL `https://github.com/propublica/compas-analysis`.

Veale, M. and Binns, R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

Xiang, A. and Raji, I. On the legal compatibility of fairness definitions, 11 2019.

Zamir, E. and Ritov, I. Loss aversion, omission bias, and the burden of proof in civil litigation. *The Journal of Legal Studies*, 41(1):165–207, 2012. doi: 10.1086/664911. URL `https://doi.org/10.1086/664911`.

Zehlike, M., Hacker, P., and Wiedemann, E. Matching code and law: Achieving algorithmic fairness with optimal transport, 2019.

Žliobaitė, I. and Custers, B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2):183–201, 2016.