Considering ‘Algorithimic decision making and the cost of fairness’ and Accompanying Normative Considerations

Jacob Thoma

2024-05-02

Introduction

The high chair facing the darkest corner of the dining room corralled what mother could only describe as a wild animal. Anger, resentment, and confusion tore apart three-year-old Jacob as he sat in tortuous isolation from his Hot Wheels. While mother hoped five minutes was enough time for Jacob to reflect on his actions, not more than five minutes later, Jacob was found around mother’s shoulder getting spanked yet again. In the context of punishment, a hand is to a mother as a gavel is to a judge. Judges are put in a similar position when contemplating a parole grant, making use of limited information to determine if an offender will be a repeat offender. Machine learning algorithms predicting recidivism risk (such as COMPAS) have been developed to aid judges for such life-altering judgement. But as a judge’s virtues lie in fairness and accuracy, recidivism risk algorithms must do the same to be used justly. The research paper *Algorithmic decision making and the cost of fairness*, authored by Corbett-Davies, Pierson, Feller, Goel, and Huq, makes use of formal proofs to explore an accuracy vs safety debate surrounding such optimization algorithms (such as COMPAS), and is linked here. This writing aims to analyze the methods and results of the above paper and consider its accompanying philosophical implications.

Approach to Formal Proofs (Methodology Analysis Part 1)

To evaluate fairness, the authors introduce three statistical metrics: *statistical parity*, *conditional statistical parity* and *predictive equality*. The notation for each can be seen below

Statistical parity is defined with the following notation

$\mathbb{E}[d(X)|g(X)] = \mathbb{E}[d(X)]$ **Eq.1**

for a decision rule $d(X)$ and indicator variable$g(X)$ of an individual’s attributes X. As race is not directly considered within the COMPAS algorithm, $g(X)$ is inferred from proxies of race in the COMPAS data set, such as zip code [1]. In the context of this research, statistical parity would be achieved if the expected probability of the COMPAS algorithm classifying someone as a repeat offender does not change based on knowing their inferred race.

Conditional statistical parity, notated

$\mathbb{E}[d(X)|l(X), g(X)] = \mathbb{E}[d(X)|l(X)]$ **Eq.2** is an extension of statistical parity that also controls for a set of "legitimate" risk factors $l(X)$ (such as prior offenses) that could influence $d(X)$ [1].

Predictive equality is measured by false positive rate (FPR) and notated$\mathbb{E}[d(X)|Y = 0, g(X)] = \mathbb{E}[d(X)|Y = 0]$ **Eq.3**

where $Y$is the true class of the individual with attributes X. As such, predictive equality in the COMPAS use case involves comparable false positive rates of COMPAS classification with and without the racial inference $g(x)$.

With definitions for fairness intact, the researchers developed mathematical framework for constrained and unconstrained optimization problems. The goal for both optimization problems was to maximize **immediate utility** $u(d, c)$, expressed below:

$u(d, c) = \mathbb{E}\left[d(X) \cdot \left(p_{Y|X} - c\right)\right]$ **Eq.4**

where c is the cost of detaining an individual in terms of crimes prevented between 0 and 1 and $p_{Y|X}$ is the probability of recidivism for a given X.

Under this definition, $u(d, c)$ can be thought of as a balance between the expected number of crimes prevented and the expected number of people detained [1]. To maximize $u(d, c)$ in this case is to find an optimal decision rule, $d(X)$ that allows for optimal classification of both repeat offenders and non-repeat offenders. The constrained optimization of $u(d, c)$ is based on the three restrictions of statistical fairness discussed earlier, while the unconstrained optimization lacks these restrictions.

Formal Proof Results and Applications (Methodology Analysis Part 2)

Now that $u(d, c)$ is explicitly defined, the authors have a clear goal to maximize $u(d, c)$ with adherence to the three separate fairness constraints. Theorem 3.2 in *Algorithmic Decision Making and the Cost of Fairness* separately defines optimal decision rules for optimized algorithms constrained on each of the three fairness conditions (see below) [1].

Algorithmic Decision Making and the Cost of Fairness, Theorem 3.2

Suppose $D\left(p_{Y|X}\right)$ has positive density on [0,1].

Unconstrained Optimum:

- The optimal decision rule $d^*(X)$ that maximizes $u(d, c)$ is

$$d^*(X) = \begin{cases} 1 & \text{if } p_{Y|X} \geq c \\ 0 & \text{otherwise} \end{cases}$$

Optimum under Statistical Parity:

- Among rules satisfying statistical parity, the optimum is

$$d^*(X) = \begin{cases} 1 & \text{if } p_{Y|X} \geq t_g(X) \\ 0 & \text{otherwise} \end{cases}$$

where $t_g(X) \in [0,1]$ are constants depending only on group membership $g(X)$.

Optimum under Conditional Statistical Parity:

- Suppose $D\left(p_{Y|X|l(X)=l}\right)$ has positive density on [0,1]. Among rules satisfying conditional statistical parity, the optimum is

$$d^*(X) = \begin{cases} 1 & \text{if } p_{Y|X} \geq t_g\left(X, l(X)\right) \\ 0 & \text{otherwise} \end{cases}$$

where $t_g\left(X, l(X)\right) \in [0,1]$ are constants depending on group membership and "legitimate" attributes.

The important proposition from Theorem 3.2 is that different optimal decisions thresholds exist across different sets of race proxies $g(x)$ and legitimate risk factors $l(x)$ to ensure statistical fairness. While this proposition will be explored from a philosophical perspective further on in this research paper, we will first observe the mathematical framework.

For sake of analysis, this paper will give a more detailed mathematical analysis for the three fairness constraints. The following section provides an intuitive logical explain in Theorem 3.2 behind two of the fairness constraints: *statistical parity* and *conditional statistical parity* because they are direct consequences of the formulation of **Eq.1**. In addition, a formal mathematical proof is recreated and analyzed for *predictive equality*.

Unconstrained Optimization Logic

The first part of Theorem 3.2 is easily seen when understanding $u(d, c)$ (see **Eq.4**) and the objectives of the paper. In the unconstrained case, maximizing $u(d, c)$ means maximizing $d(X) \cdot \left(p_{Y|X} - c\right)$. Any cases where $p_{Y|X}$ is less than c have the potential to reduce $u(d, c)$, because a case where $p_{Y|X} < c$ and $d(X) = 1$ is strictly negative. Therefore, maximizing $u(d, c)$ means only cases where $p_{Y|X} \geq c$ are cases where $d(X) = 1$, and $d(X) = 0$ otherwise. This idea makes sense from the unconstrained perspective because we are essentially saying that the riskiest inmates (any with risk above the decision threshold c) should be detained while less risky inmates (below c) should be released.

Statistical and Conditional Statistical Parity Logic

The second part and third pieces of Theorem 3.2 are extremely similar to the unconstrained optimized algorithm with threshold c with one important change. For an optimized algorithm that is constrained on statistical parity, it is imperative to detain the same proportion of inmates across a set of indicator variables $g(x)$. This idea here instead of detaining c defendants is to detain the riskiest p^* defendants across groups $g(x)$ to ensure statistical parity. As admitted in the book, this is done under the assumptions of a continuous distribution of $p_{Y|X}$ that allow for existence of some threshold $t_{g(x)}$ that guarantee an equal proportion of detainees p^* across every $g(x)$. As such, the optimal algorithm constrained on statistical parity finds an optimal proportion of detained inmates p^* equal across all group memberships $g(x)$ by setting varying thresholds for $t_{g(x)}$ per each $g(x)$, guaranteeing the result of **Eq.1**. The same logic and distribution assumptions apply to constrained optimization with conditional statistical parity, the only addition being the set of "legitimate risk factors" $l(x)$ that is considered in addition to $g(x)$. There exist thresholds $t_{g(x),l(x)}$ that guarantee an proportion of detainees p^* across every membership of $g(x)$ and $l(x)$. Once the threshold are in for each fairness condition are found, the same logic for maximizing $u(d,c)$ applies with the new optimal threshold sets $t_{g(x)}$ and $t_{g(x),l(x)}$ for statistical and conditional statistical parity respectively.

Predictive Equality -

The conclusions of Theorem 3.2 apply to predictive equality, but require more than intuition to derive. This paper attempts to expand upon the predictive equality proof related to Theorem 3.2.

Objective

Let d be a decision rule satisfying equal false positive rates (per **Eq.3**) that does not use multiple thresholds. We will prove that there is a decision rule d' satisfying equal false positive rates with strictly higher utility such that

$$u(d',c) > u(d,c)$$

Proof

Assume that the distribution of $p_{Y|X}$ is of a continuous distribution. (**assumption.1**)

Assume that there exists some a^* such that $P(Y|X, G(X) = a^*)$ is non-equal to any other conditional distribution on a group of a such that:

$P(Y|X, G(X) = a^*) \neq P(Y|X, G(X) = a)$ (**assumption.2**)

This assumption is plausible when considering that the conditional distribution of $p_{Y|X}$ across all race proxies $g(x)$ are impossible to be exactly identical.

Sub Proof - Mis-classification present in d

Assume for sake of contradiction that for every a where $g(X) = a$, the decision rule d detains all defendants with predicted probabilities above t_a with no error. Therefore, d is a perfect decision rule.

It is impossible to have predictive equality in single threshold rule d if it is not inherently a quality of the data which does not hold in **assumption.2**. This is a direct contradiction with **assumption.2** .

By contradiction, d is not a perfect decision rule free of error.

Therefore, there will be some mis-classification in d. (**result.1**)

Sub Proof Complete

By **assumption.1** $\mathrm{P}\left[p_{Y|X} = t_{a*}\right] = 0$

Define t_a such that the same proportion of people detained by d for group $g(X) = a$ are detained by the rule with this threshold.

$$\mathbb{E}[d(X)|g(X) = a] = \mathbb{E}\left[1\{p_{Y|X} \geq t_a\}|g(X) = a\right]$$

By the **assumption.1** and **result.1**, there must be some group a^* where the proportion of detainees are mis-classified such that there are high risk offenders released and low risk offenders detained like so.

$$E\left[1_{\{pY|X \geq t_{a^*}\}}\left(1 - d(X)\right) \mid g(X) = a^*\right]$$

$$E\left[1_{\{pY|X < t_{a^*}\}}d(X) \mid g(X) = a^*\right]$$

By **assumption.1**, both of these expectations are continuous and by **result.1** they are nonzero.

By **result.1** and the continuous nature of these expectations, there is some β such that

$$\beta = \mathbb{E}\left[1\{p_{Y|X} \geq t_a^*\}\left(1 - d(X)\right)|g(X) = a^*\right] = \mathbb{E}\left[1\{p_{Y|X} < t_a^*\}d(X)|g(X) = a^*\right] > 0$$

(**result.3**)

Let t_1 and t_2 be new thresholds such that $0 \leq t_1 < t_2 \leq 1$. Define the new threshold rule d' as follows

$$d'_{t1,t2}(X) = \begin{cases} 1 & \text{if } p_{Y|X} \geq t_2, g(X) = a^* \\ 0 & \text{if } p_{Y|X} < t_1, g(X) = a^* \\ d(X) & \text{otherwise} \end{cases}$$

Define β_1 , β_2 , γ_1, and γ_2 with the following notation

Let β_1 represent the expected number of new detainees below t_1 released by d' compared to d

$$\beta_1(t_1, t_2) = \mathbb{E}\left[1\{p_{Y|X} < t_1\}d(X)|g(X) = a^*\right]$$

Let γ_1 represent the expected number of new innocent detainees released by d' based on β_1

$$\gamma_1(t_1, t_2) = \mathbb{E}\left[1\{p_{Y|X} < t_1\}d(X)\left(1 - p_{Y|X}\right)|g(X) = a^*\right]$$

Let β_2 represent the expected number of new detainees below t_2 detained by d' compared to d

$$\beta_2(t_1, t_2) = \mathbb{E}\left[1\{p_{Y|X} \geq t_2\}\left(1 - d(X)\right)|g(X) = a^*\right]$$

Let γ_2 represent the expected number of new innocent detainees by d' based on β_2

$$\gamma_2(t_1, t_2) = \mathbb{E}\left[1\{p_{Y|X} \geq t_2\}\left(1 - d(X)\right)\left(1 - p_{Y|X}\right) | g(X) = a^*\right]$$

By **result.1** and **assumption.1** (presence of mis-classification in d)

$$\gamma_1(t_1, t_2) \geq (1 - t_1)\beta_1(t_1, t_2)$$

$$\gamma_2(t_1, t_2) \leq (1 - t_2)\beta_2(t_1, t_2)$$

Now, choose t_a^* such that $t_1 < t_a^* < t_2$.

By **assumption.1**, **result.3** and the definition of the values t_1 and t_2 could take, β_1 and β_2 are continuous

Therefore, there exists a t_a^* such that $\beta_2(t_1, t_2) = \beta_1(t_1, t_2) = \beta$

In this case,

$$\gamma_1(t_1, t_2) \geq (1 - t_1)\beta \quad \text{and} \quad \gamma_2(t_1, t_2) \leq (1 - t_2)\beta$$

And because $t_1 < t_2$,

$$\gamma_1(t_1, t_2) > \gamma_2(t_1, t_2)$$

(**result.4**)

result.4 implies that that decision rule d' releases more innocent low-risk offenders than it decides to detain innocent high risk defendants. The moral implications of this will be explored later in the paper.

Equalizing false positive rates between d and d' means ensuring equality between γ_1 and γ_2 . This can be thought of as redistributing the errors made across d' to ensure the equal false positive rates to d. Decreasing t_1 means decreasing the threshold for detainment for lower-risk offenders.

By the continuity of B_1, B_2, t_1, t_2, and the formulaic definitions of γ_1 and γ_2, γ_1 and γ_2 are continuous.

The definition of d' , γ_1 and γ_2 are also convenient in the fact that γ_1 only depends on t_1 and γ_2 only depends on t_2, allowing the following logic to hold.

By **result.4** there exists a threshold t'_1 in the continuous range between 0 and t_1 such that $t'_1 < t_1$ where the following holds

$$\gamma_1(t'_1, t_2) = \gamma_2(t_1, t_2) = \gamma_2(t'_1, t_2)$$

Because t'_1 is less than t_1 , it follows that

$$\beta_1(t'_1, t_2) < \beta_1(t_1, t_2) = \beta_2(t_1, t_2)$$

(**result.5**)

Which essentially means that a decision rule exists with the thresholds t'_1 and t_2 which maintains the false positive rate present in d while releasing fewer offenders total. Below is the formulation of this improved decision rule.

$$d_{t'_1,t_2}(X) = \begin{cases} 1 & \text{if } p_{Y|X} \geq t_2, g(X) = a^* \\ 0 & \text{if } p_{Y|X} < t'_1, g(X) = a^* \\ d(X) & \text{otherwise} \end{cases}$$

Now, we can directly compare the difference between $u\left(d'_{t'_1,t_2}, c\right)$ and $u(d, c)$. If this difference is nonzero, than the new and improved decision rule offers higher utility than d by detaining more people with the same false positive rate.

$$u\left(d'_{t'_1,t_2}, c\right) - u(d, c) = E\left[d'_{t_1,t_2}(X)\left(p_{Y|X} - c\right)\right] - E\left[d(X)\left(p_{Y|X} - c\right)\right]$$

By the linearity of expectation

$$u\left(d'_{t_1,t_2}\right) = E\left[d'_{t'_1,t_2}(X)(1 - c)\right] - E\left[d'_{t'_1,t_2}(X)\left(1 - p_{Y|X}\right)\right]$$

$$u(d, c) = E[d(X)(1 - c)] - E\left[d(X)\left(1 - p_{Y|X}\right)\right]$$

Therefore,

$$u\left(d'_{t'_1,t_2}, c\right) - u(d, c)$$
$$= E\left[d'_{t'_1,t_2}(X)(1 - c)\right] - E\left[d'_{t'_1,t_2}(X)\left(1 - p_{Y|X}\right)\right] - E[d(X)(1 - c)] + E\left[d(X)\left(1 - p_{Y|X}\right)\right]$$

Because d'_{t_1,t_2} and d' have equalized false positive rates,

$$E\left[d'_{t'_1,t_2}(X)\left(1 - p_{Y|X}\right)\right] = E\left[d(X)\left(1 - p_{Y|X}\right)\right]$$

Allowing a simplification to

$$u\left(d'_{t'_1,t_2}, c\right) - u(d, c) = E\left[d'_{t'_1,t_2}(X)(1 - c)\right] - E[d(X)(1 - c)]$$

Also by the linearity of expectation

$$u\left(d'_{t'_1,t_2}, c\right) - u(d, c) = (1 - c)E\left[d'_{t'_1,t_2}(X)\right] - (1 - c)E[d(X)]$$

$$u\left(d'_{t'_1,t_2}, c\right) - u(d, c) = (1 - c)\left(E\left[d'_{t'_1,t_2}(X)\right] - E[d(X)]\right)$$

By definition of $E\left[d'_{t'_1,t_2}(X)\right]$, it is an expected number of new detainees detained under $d'_{t'_1,t_2}$, which can be written as $\beta_2(t'_1, t_2)$ like so: $E\left[d'_{t'_1,t_2}(X)\right] = \beta_2(t'_1, t_2)$

$\beta_1(t'_1, t_2)$ only depends on the value of t'_1 . As such, varying t'_1 over its entire range is equivalent to finding the expectation at different decision thresholds for $d(X)$. As such, $\beta_1(t'_1, t_2) = E[d(X)]$.

Therefore,

$$u\left(d'_{t'_1,t_2}, c\right) - u(d, c) = (1 - c)\left(\beta_2(t'_1, t_2) - \beta_1(t'_1, t_2)\right)$$

Using **result.5** and the fact that $\beta_2(t'_1, t_2)$ is only a function of t_2 we know $\beta_2(t'_1, t_2) = \beta_2(t_1, t_2)$

$$\beta_1(t'_1, t_2) < \beta_2(t'_1, t_2)$$

Using the fact that $0 < c < 1$ and the above expression

$$u\left(d'_{t'_1,t_2}, c\right) - u(d, c) > 0$$

(**result.6**)

Therefore, there exists a decision rule $d'_{t'_1,t_2}$ with thresholds t'_1 and t_2 with strictly higher utility than d' that maintains predictive equality. As such, the optimal decision rule for an algorithim with a predicitive equality constraint is a multiple threshold rule.

Proof Complete

Once the optimized decision framework was in place, the researchers applied the theory to a subset of the COMPAS dataset. Three constrained optimization frameworks (one for each fairness constraint) were compared directly to an unconstrained optimization algorithm. The constrained optimization algorithms, while satisfying one conditional of statistical fairness each, detained more low-risk offenders in the process. Unfortunately, the detainment of more low-risk offenders in these cases led to estimated increases in violent crime of 9% (statistical parity constraint), 7% (predictive equality constraint), and 4% (conditional statistical parity constraint) in reference to the unconstrained algorithm [1].

Now after observation of the methodology in *Algorithmic Decision Making and the Cost of Fairness*, this paper will explore and take a position on some of the normative concerns in the following section.

Normative Considerations

There are several direct moral implications of Algorithmic *Decision Making and the Cost of Fairness*. In order to evaluate these implications, this paper will adopt the perspective of a judge, as judges are often the ones that have to make the final ruling on an offender's extended detainment or release.

Statistical fairness cannot be looked at one angle at a time

A main objective of the methodology is to mathematically prove that an optimal decision algorithm constrained by any of the three statistical fairness defintions uses a multiple threshold rule. While the authors of *Algorithmic Decision Making and the Cost of Fairness* do successfully show that multiple threshold decision rules are optimal to meet each of these fairness conditions, they fail to consider the moral implications of not meeting all of these fairness conditions at once.

From the perspective of a judge, fairness is a virtue. How is a fair and honorable judge supposed to choose between the three fairness constraints at hand when there are different sets of decision thresholds for each? This point is especially important for inmates that are very close to decision thresholds in any of the algorithms. The difference between a judge abiding by predictive equality or statistical fairness could be the deciding factor in whether or not an inmate is classified as a "recidivist", and this choice is objectively unfair to the inmate because it has nothing to do with their case. A fair judge should not let any factors outside of the inmates case facts control their decision to grant parole. As such, a judge that honors true fairness would rather not arbitrarily choose to follow one of the fairness constraints over another if it means arbitrarily changing peoples lives in the process.

A judge that emphasizes fairness would rather use an optimized algorithm that considers all fairness conditions as constraints. It remains to be proven if an algorithm constrained on all three fairness conditions is usable, let alone possible. But neglecting other aspects of fairness and accuracy just to meet one of these fairness constraints disproportionately and arbitrarily affects some inmates more than others, and is thus unfair.

The immorality of using of multiple decision thresholds in a recidivism algorithm

While there is plenty of moral concern surrounding the idea of abiding by only one of the three fairness constraints, there are also significant moral concerns surrounding the fundamental idea of multiple decision thresholds that vary across different proxies for race $g(X)$.

Let's present an example:\
Our current decision algorithm is set as follows:

- *Black inmates have a decision threshold of 0.3*
- *White inmates have a decision threshold of 0.2*
- *We have two inmates, one black and one white, with the exact same calculated recidivism risk probability of 0.25*

By the nature of the decision algorithms referenced earlier, this algorithm would dictate that the black inmate is released while the white inmate is further detained and labeled as a recidivist despite having the same recidivism risk probability of 0.25.

For a judge, consistency is one of the most important virtues. In the case of sentencing inmates, a judge should sentence offenders with the same risk the exact same. Regardless if the optimized multiple decision thresholds are based on fairness or optimized on accuracy, multiple threshold rules fundamentally go against the idea of consistency by considering race as a component of the decision. In the case of the example above, the white inmate has a lower threshold of risk to being labeled a recidivist. The following propositions are potential reasons why white inmates could have a lower threshold in a recidivism risk algorithm:

1. White inmates actually have a lower overall recidivism risk than black inmates.
2. White inmates and black inmates have different levels of recidivism risk because of injustice towards black inmates.

If proposition 1 had any remote truth whatsoever, than giving white inmates a lower threshold for recidivism risk to obtain statistical fairness is essentially a punishment to all white inmates for the favorable, obedient behavior of some white inmates who did not repeat offend. On the flip-side, black inmates are intrinsically rewarded for the recidivism of other black inmates by having a higher recidivism risk threshold. A judge that values fairness should only consider factors in the control of the inmate, and as such, should not let the recidivism of other offenders affect their judgement on a current case. Therefore, if proposition 1 is true, multiple decision threshold rules are immoral on the grounds of unfairness towards different races.

If proposition 2 was true, than a plausible point could be that there is some need for statistical fairness by way of multiple thresholds to make up for inconsistent recidivism risk scores across race proxies $g(X)$. While this point attempts to undermine a single decision threshold in favor of multiple decision thresholds, it is counterproductive to its goal. The problem of inconsistency in risk scores of recidivism algorithms cannot simply be accounted for by introducing more inconsistency in a judge's decision. One could argue that the one of the main causes in inconsistent recidivism risk scores is based on historical inequalities in treatment of different races on the streets and in the court rooms. How could a judge that values consistency justify inconsistent rulings of inmates now to make up for past failures of the justitice system, especially when the problem of inconsistent risk scores could be dealt with closer to its source with more fairness in policing or statistical adjustments to the risk scores themselves? Instead, a judge that values consistency and honor would like to be part of the solution by being consistent in their rulings regardless of race. Simply put, two wrongs do not make a right.

Statistical fairness vs. increased violent crime

Another philosophical consideration of adhering to statistical fairness are the consequences of violent crime on the public. The estimated impact of meeting a single fairness constraint has already been mentioned, leading to estimated increases in violent crime of 9% (statistical parity constraint), 7% (predictive equality constraint), and 4% (conditional statistical parity constraint) compared to an optimized unconstrained algorithm. These estimated increases in crime are in some part due to an increase in the percentage of low-risk inmates that remain detained, which is itself a moral issue.

Another fundamental virtue for a judge is confidence. When a judge has to make a decision regarding a parole case, they need to be sure it is the right one because peoples lives are at stake. A confident judge must be confident in fairly assinging the risk of the offender in consideration. When a judge identifies someone as a "low risk" to be a recidivist, they should be confident enough to put their own life down on it, because they indirectly are along with the lives of others. Relying on some arbitrary notion of statistical fairness to release “high risk” individuals instead of “low risk” individuals is the opposite of confidence. A confident and fair judge should not feel the need to mathematically adhere to statistical fairness constraints, because their decisions are already fairly made on a case by case basis. In other words, they should not care about statistical parity or predictive equality, they should decide fairly and confidently inmate by inmate and hope that the patterns of behavior in society are consistent enough to reflect statistical fairness intrinsically.

Conclusion

Algorithmic Decision Making and the Cost of Fairness provides detailed mathematical insight into the debate between accuracy of a decision algorithm and fairness. While there is proof that optimized algorithms with multiple decision thresholds are optimal even under constraints of fairness, there deserves to be philosophical debate behind its use. In starting this conversation, this paper brings up questions about statistical fairness, what it means, and if it is being used properly. The key virtues of judge: fairness, consistency, and confidence are used to take moral positions on the implications of multiple decision threshold rules. Ultimately, this paper takes the stance that statistical fairness metrics should be used as checks for fairness, not as constraints to achieve. At the end of the day, a judge must trust his gavel as a mother trusts her spanking hand in hopes for a better tomorrow.

While the moral stances of this paper do not align with the proposed multiple decision threshold rules in *Algorithmic Decision Making and the Cost of Fairness*, these stances do not aim to demerit the paper. If anything, *Algorithmic Decision Making and the Cost of Fairness* should be applauded for its use of several statistical fairness conditions as landmarks for success in the form of true justice. While these statistical landmarks do not currently line up well with the optimized unconstrained algorithm, the authors help expose deeper philosophical concerns in the data being used, such as inconsistencies in assigning recidivism risk scores. While the intention of this writing is to analyze *Algorithmic Decision Making and the Cost of Fairness,* the author hopes it is seen as a step in the right direction towards fairness simply by creating discussion around important normative concerns.

References

[1] Corbett-Davies, Sam, et al. "Algorithmic Decision Making and the Cost of Fairness." In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM, 2017. https://doi.org/10.1145/3097983.3098095.