"Algorithmic Accountability" (Bartlett et al., 2019): Formalization and Comments

John Heilbron

March 13, 2020

1 Setting

The authors consider a setting in which one agent, j, makes a decision about how to treat another agent, i. This setting has many applications in the law, as they demonstrate with examples. I will consider an employer, j, and a prospective employee, i.

The law designates the following:

- $\mathbb{P}$, a set of "protected classes", observable characteristics of the prospective employee that an employer will not be permitted to subject to "disparate treatment" or "disparate impact" (defined later), unless there is a legal exemption.
- $G_i \in \mathbb{P}$, that gender is a "protected class".
- $\mathbb{N}$, a set of "legitimate business necessities", (potentially unobservable) qualification statuses of a worker, each comprised of a worker characteristic and threshold, for which the law grants a legal exemption from prohibitions on "disparate impact".
- $S_i \equiv \mathbb{1}\{A_i \geq \tau_s\} \in \mathbb{N}$, that it is a "legitimate business necessity" for a worker to be satisfactory, i.e. for the worker to have ability or productivity, A_i, in an amount greater than some minimum satisfactory threshold, τ_s.

The prospective employee has the following attributes:

- $X_i \in \mathbb{R}^n$, a suite of ex-ante observable characteristics with no legal designation.
- $G_i \in \{0, 1\}$, an ex-ante observable characteristic, gender.
- $A_i \in \mathbb{R}$, an ex-ante unobservable characteristic of interest to the employer, like ability or productivity.

- $S_i \equiv \mathbb{1}\{A_i \geq \tau_s\} \in \{0,1\}$, the ex-ante unobservable status of a worker as satisfactory according to whether their ability, legally-designated as a "business necessity", surpasses the legally-designated threshold.

The employer makes a hiring decision as follows:

- $f : \mathbb{R}^{(n+1)} \to \mathbb{R}$, a scoring method for evaluating any prospective employee.
- $\tau_h \in \mathbb{R}$, a threshold for determining the score above which any prospective employee will be hired.
- $H_i \equiv \mathbb{1}\{f(X_i, G_i) \geq \tau_h\} \in \{0,1\}$, a decision rule comparing the score of observable characteristics to the employment threshold, the hiring outcome for the employee.

2 Standards of Discrimination

The law restricts certain hiring practices:

- "Disparate treatment", $H_i \mid X_i, G_i \not\perp H_i \mid X_i$, is the differential application of hiring decision-rules according to the gender of the prospective employee. This is prohibited.
- "Disparate impact", $E[H_i \mid G_i] \neq E[H_i]$, is application of proxy variables and hiring decision-rules that result in differential hiring rates by gender. This is restricted in the sense that an employee may bring a suit by presenting evidence of this fact as part of the "first stage of the burden-shifting framework".
- "Disparate impact within qualification status", $E[H_i \mid S_i, G_i] \neq E[H_i \mid S_i]$, is an application of proxy variables and hiring decision-rules that result in differential hiring rates by gender, conditional on prospective employee qualification. This weakens the restrictions on "disparate impact" to the extent that the law designates "legitimate business necessities" relevant to a business, and defendants may demonstrate the relevance of such "business necessities" in the "second stage of the burden-shifting framework". This is still restricted in the sense that defendants must demonstrate that they do not disadvantage otherwise qualified applicants of a given gender. (This standard appears to be consistent with the case law surveyed by the authors, though, as shown below, it is somewhat different from the interpretation that they suggest.)

3 Insufficient Defenses of Disparate Impact

The paper uses case law to illustrate insufficient defenses of disparate impact, i.e. when responding to the demonstrated claim in the "first stage of the burden-shifting framework" that $E[H_i \mid G_i] \neq E[H_i]$.

- The hiring decision-rule selects on the qualification status of the worker, $E[S_i \mid H_i] \neq E[S_i]$. Note this is insufficient even when the qualification status varies with the protected class, $E[S_i \mid G_i] \neq E[S_i]$, and the qualification status, $S_i \equiv \mathbb{1}\{A_i \geq \tau_s\}$, is deemed a "legitimate business necessity", $S_i \in \mathbb{N}$. The example of the prison underscores that the law does not deem this a sufficient defense.

- The hiring decision-rule does not generate disparate impact within qualification status, $E[H_i \mid \hat{S}_i, G_i] = E[H_i \mid \hat{S}_i]$, but the qualification status, $\hat{S}_i \equiv \mathbb{1}\{\hat{A}_i \geq \hat{\tau}_s\}$, is not legally designated a "legitimate business necessity", $\hat{S}_i \notin \mathbb{N}$. The example of profitability in lending underscores this point.

- The hiring decision-rule does not generate disparate impact within qualification status, $E[H_i \mid S_i^*, G_i] = E[H_i \mid S_i^*]$, even when the law designates a "legitimate business necessity", $S_i \equiv \mathbb{1}\{A_i \geq \tau_s\} \in \mathbb{N}$, if the employer uses a higher standard for determining qualification status, $S_i^* \equiv \mathbb{1}\{A_i \geq \tau_s^*\}$ and $\tau_s^* > \tau_s$. The example of hospital chaplains underscores this point.

4 Input Accountability Test: Comment on Orthogonality Conditions

The paper suggests the use of an "input accountability test" (IAT) to determine which proxy variables, $x_i = X_i^{[\ell]}$, may be used in the scoring function. The test is a one-by-one projection of proxy variables on the qualification status and protected class, $x_i \sim \alpha^{IAT} + \beta^{IAT} * S_i + \gamma^{IAT} * G_i + \varepsilon_i^{IAT}$. Characteristics passing the test, i.e. with $\gamma^{IAT} = 0$, may be used in the scoring function.

The article imposes certain orthogonality conditions on this projection that are different from the standard OLS orthogonality conditions. We could imagine a test based on a more standard OLS projection, $x_i \sim \alpha^{OLS} + \beta^{OLS} * S_i + \gamma^{OLS} * G_i + \varepsilon_i^{OLS}$. The specific orthogonality conditions suggested by the article appear not to matter in the sense that a variable passes the IAT if and only if it passes the OLS-IAT, $\gamma^{IAT} = 0 \iff \gamma^{OLS} = 0$.

The intuition for why this is the case comes from considering the orthogonality conditions imposed by the IAT and OLS projections. OLS imposes that $\varepsilon_i^{OLS} \perp 1, S_i, G_i$. The IAT imposes

that $\varepsilon_i^{IAT} \perp G_i$ and that $(\gamma^{IAT} * G_i + \varepsilon^{IAT}) \perp 1, S_i$. Note that if $\gamma^{IAT} = 0$ then this second condition basically reduces to $\varepsilon^{IAT} \perp 1, S_i$, and the remainder of the OLS conditions will also be satisfied. (N.B. The intuition runs the other way as well.)

We can prove this definitively by constructing γ^{OLS} and γ^{IAT} according to the projection restrictions. Beginning with the OLS projection, we have:

$$0 = E\left[\begin{bmatrix} 1 \\ S_i \\ G_i \end{bmatrix}\varepsilon_i^{OLS}\right] = E\left[\begin{bmatrix} 1 \\ S_i \\ G_i \end{bmatrix}\left(x_i - \begin{bmatrix} 1 & S_i & G_i \end{bmatrix}\begin{bmatrix} \alpha^{OLS} \\ \beta^{OLS} \\ \gamma^{OLS} \end{bmatrix}\right)\right] = \begin{bmatrix} E[x_i] \\ E[x_iS_i] \\ E[x_iG_i] \end{bmatrix} - \begin{bmatrix} 1 & E[S_i] & E[G_i] \\ E[S_i] & E[S_i^2] & E[S_iG_i] \\ E[G_i] & E[S_iG_i] & E[G_i^2] \end{bmatrix}\begin{bmatrix} \alpha^{OLS} \\ \beta^{OLS} \\ \gamma^{OLS} \end{bmatrix}$$

Let $\mathcal{D} \equiv \det\begin{bmatrix} 1 & E[S_i] & E[G_i] \\ E[S_i] & E[S_i^2] & E[S_iG_i] \\ E[G_i] & E[S_iG_i] & E[G_i^2] \end{bmatrix}$ and use matrix algebra to solve for the projection coefficients. We recover:

$$\gamma^{OLS} = \frac{E[x_i]\Big(E[S_i]E[S_iG_i] - E[G_i]E[S_i^2]\Big) - E[x_iS_i]Cov(S_i, G_i) + E[x_iG_i]Var(S_i)}{\mathcal{D}}$$

Alternately, for the IAT projection, we have:

$$0 = E\Big[G_i\varepsilon_i^{IAT}\Big] = E\left[G_i\Big(x_i - \begin{bmatrix} 1 & S_i & G_i \end{bmatrix}\begin{bmatrix} \alpha^{IAT} \\ \beta^{IAT} \\ \gamma^{IAT} \end{bmatrix}\Big)\right] = E[G_ix_i] - \begin{bmatrix} E[G_i] & E[G_iS_i] \end{bmatrix}\begin{bmatrix} \alpha^{IAT} \\ \beta^{IAT} \end{bmatrix} - E[G_i^2]\gamma^{IAT}$$

$$0 = E\Big[\begin{bmatrix} 1 \\ S_i \end{bmatrix}(\gamma^{IAT}G_i + \varepsilon_i^{IAT})\Big] = E\left[\begin{bmatrix} 1 \\ S_i \end{bmatrix}\Big(x_i - \begin{bmatrix} 1 & S_i \end{bmatrix}\begin{bmatrix} \alpha^{IAT} \\ \beta^{IAT} \end{bmatrix}\Big)\right] = \begin{bmatrix} E[x_i] \\ E[x_iS_i] \end{bmatrix} - \begin{bmatrix} 1 & E[S_i] \\ E[S_i] & E[S_i^2] \end{bmatrix}\begin{bmatrix} \alpha^{IAT} \\ \beta^{IAT} \end{bmatrix}$$

Solve for γ^{IAT} in the first equation, plug in solutions for $\begin{bmatrix} \alpha^{IAT} \\ \beta^{IAT} \end{bmatrix}$ from the second, and multiply terms by $\frac{Var(S_i)}{Var(S_i)} = 1$ as necessary for clarity:

$$\gamma^{IAT} = \frac{E[x_iG_i]Var(S_i) - E[G_i]\Big(E[x_i]E[S_i^2] - E[x_iS_i]E[S_i]\Big) - E[G_iS_i]\Big(E[x_iS_i] - E[x_i]E[S_i]\Big)}{E[G_i^2]Var(S_i)}$$

Note that, rearranging terms, the numerators of γ^{IAT} and γ^{OLS} are equal. Therefore, $\gamma^{IAT} = 0 \iff \gamma^{OLS} = 0$.

5 Input Accountability Test: A Proof

In the spirit of the paper, I consider an "input accountability test" (IAT) to determine which proxy variables, X_i, may be used in the scoring function. The test is a one-by-one projection of proxy variables on the "business necessity" and protected class, $X_i^{[\ell]} \sim \alpha^\ell + \beta^\ell * S_i + \gamma^\ell * G_i + \varepsilon_i^\ell$. Characteristics with $\gamma^\ell = 0$ may be used in the scoring function. I ignore the specific orthogonality

conditions proposed by the paper, both because the equivalence result shown above and because the choice of orthogonality conditions appears to have no bearing on the result below.

Suppose the employer is choosing a decision rule for hiring. Suppose, for the sake of legal compliance, the decision rule does not exhibit disparate treatment. For simplicity, suppose also that the employer is choosing within a class of linear scoring functions, $f(X_i, G_i) = \Gamma' X_i$. Suppose that the legally-designated "business necessity" is related to the protected class, $E[S_i \mid G_i] \neq E[S_i]$. But suppose that each characteristic entering the decision rule, $X_i^{[\ell]}$ for $\ell \in \{1, ..., n\}$, satisfies the IAT. We also suppose, for convenience, that the residual of the projection among those satisfying the "business necessity" is invariant to gender, $\varepsilon_i \mid S_i \sim \varepsilon_i \mid S_i, G_i$. Then the employer can choose an arbitrary scoring function, Γ, and threshold, τ_h, as the basis for its decision rule without fear of violating standing discrimination case law even if it knowingly violates the "disparate impact" standard .

Under decision-rule $H_i \equiv \mathbb{1}\{\Gamma' X_i \geq \tau_h\}$, we have:

$$E[\mathbb{1}\{\Gamma' X_i \geq \tau_h\} \mid S_i]$$

$$= E[\mathbb{1}\{\Gamma'(\alpha + \beta S_i + \gamma G_i + \varepsilon_i) \geq \tau_h\} \mid S_i]$$

$$\overset{\gamma=\overline{0}\,;\, \varepsilon_i | S_i \sim \varepsilon_i | S_i, G_i}{=} E[\mathbb{1}\{\Gamma'(\alpha + \beta S_i + \gamma G_i + \varepsilon_i) \geq \tau_h\} \mid S_i, G_i]$$

$$= E[\mathbb{1}\{\Gamma' X_i \geq \tau_h\} \mid S_i, G_i]$$

But this says that, even in the event of disparate impact by protected class, $E[H_i \mid G_i] \neq E[H_i]$, we will have satisfied no disparate impact by protected class among satisfactory workers, in particular, $E[H_i \mid S_i, G_i] = E[H_i \mid S_i]$. In this sense, the employer will not have run afoul of the case law cited in the article.

6 Input Accountability Test: Comment on Test Limitations

The use-value the authors ascribe to the IAT is clear (and formalized in at least in some limited set of circumstances). It is worth noting, though, that the authors recommend throwing out any instance of a proxy variable, $X_i^{[\ell]}$, that does not satisfy the IAT. They emphasize that an advantageous feature of machine learning is the ability to search for signals that do, in fact, satisfy this criterion.

What doesn't seem clear, though, is whether any scoring procedure using a variable that fails

to satisfy the IAT will necessarily fail to satisfy the standard of case law they cite, as formalized in the above definition. In fact, under the same setting as described above, where I provide (an admittedly simple) proof of their concept, the contrary seems to be true.

Suppose the above conditions prevail but that now each proxy variable fails the IAT, i.e. $\gamma^{\ell} \neq 0 \ \forall \ell \in \{1, ..., n\}$. Suppose, though, that there is some linear score function, Γ, orthogonal to γ. Then let $\Gamma' X_i$ be the score function and choose some arbitrary τ_h. Now, we see there is a score function that would violate the IAT but that would, notwithstanding, not violate the case law as interpreted in the article and formalized above.

In particular, again let $H_i \equiv \mathbb{1}\{\Gamma' X_i \geq \tau_h\}$, and we have:

$$E[\mathbb{1}\{\Gamma' X_i \geq \tau_h\} \mid S_i]$$

$$= E[\mathbb{1}\{\Gamma'(\alpha + \beta S_i + \gamma G_i + \varepsilon_i) \geq \tau_h\} \mid S_i]$$

$$\overset{\Gamma'\gamma=0 \ ; \ \varepsilon_i | S_i \sim \varepsilon_i | S_i, G_i}{=} E[\mathbb{1}\{\Gamma'(\alpha + \beta S_i + \gamma G_i + \varepsilon_i) \geq \tau_h\} \mid S_i, G_i]$$

$$= E[\mathbb{1}\{\Gamma' X_i \geq \tau_h\} \mid S_i, G_i]$$

Certainly, machine learning has the power to identify signals that pass the IAT. Constraining lenders to rely solely on the IAT, however, overlooks the power of machine learning to identify combinations of signals that might separately constitute impermissible disparate impact, but are jointly tolerable.