Supplementary material for "The Cost of Fairness in Binary Classification"

Appendix A. Proofs of results in main body

Proof [Proof of Lemma 1] By definition,

$$\begin{aligned}
\mathrm{DI}(f) \geq \tau &\iff \frac{\mathrm{FPR}(f)}{1-\mathrm{FNR}(f)} \geq \tau \\
&\iff \mathrm{FPR}(f) \geq \tau - \tau \cdot \mathrm{FNR}(f)\text{ , since } \mathrm{FNR}(f) \leq 1 \\
&\iff \tau \cdot \mathrm{FNR}(f) + \mathrm{FPR}(f) \geq \tau \\
&\iff \frac{\tau}{1+\tau} \cdot \mathrm{FNR}(f) + \frac{1}{1+\tau} \cdot \mathrm{FPR}(f) \geq \frac{\tau}{1+\tau}\text{ , since } 1+\tau > 0 \\
&\iff \kappa \cdot \mathrm{FNR}(f) + (1-\kappa) \cdot \mathrm{FPR}(f) \geq \kappa\text{ , by definition of } \kappa \\
&\iff \mathrm{CS}_{\mathrm{bal}}(f; 1-\kappa) \geq \kappa.
\end{aligned}$$

■

Proof [Proof of Lemma 2] By definition,

$$\begin{aligned}
\mathrm{MD}(f) &= \mathrm{FNR}(f) + \mathrm{FPR}(f) - 1 \\
&= 2 \cdot (1/2 \cdot \mathrm{FNR}(f) + 1/2 \cdot \mathrm{FPR}(f)) - 1 \\
&= 2 \cdot \mathrm{CS}_{\mathrm{bal}}(f; 1/2) - 1.
\end{aligned}$$

The subsequent implication follows trivially. ■

Proof [Proof of Lemma 3] For any f and $\bar{c}$,

$$\begin{aligned}
\mathrm{CS}_{\mathrm{bal}}(1-f; \bar{c}) &= (1-\bar{c}) \cdot \mathrm{FNR}(1-f) + \bar{c} \cdot \mathrm{FPR}(1-f) \\
&= (1-\bar{c}) \cdot (1 - \mathrm{FNR}(f)) + \bar{c} \cdot (1 - \mathrm{FPR}(f)) \text{ by Equation 1} \\
&= (1-\bar{c}) + \bar{c} - (1-\bar{c}) \cdot \mathrm{FNR}(f) - \bar{c} \cdot \mathrm{FPR}(f) \\
&= 1 - \mathrm{CS}_{\mathrm{bal}}(f, \bar{c}).
\end{aligned}$$

Lemmas 1 and 2 imply that for any τ, there exists suitable $\bar{c}, \kappa$ such that

$$R_{\mathrm{fair}}(f) \geq \tau \iff \mathrm{CS}_{\mathrm{bal}}(f; \bar{c}) \geq \kappa.$$

For example, for the DI score, $\bar{c} = 1 - \kappa = 1/(1+\tau)$, while for the MD score, $\bar{c} = 1/2$ and $\kappa = (1+\tau)/2$. Since this is true for any classifier, we also have

$$\begin{aligned}
R_{\mathrm{fair}}(1-f) \geq \tau &\iff \mathrm{CS}_{\mathrm{bal}}(1-f; \bar{c}) \geq \kappa \\
&\iff 1 - \mathrm{CS}_{\mathrm{bal}}(f, \bar{c}) \geq \kappa \\
&\iff \mathrm{CS}_{\mathrm{bal}}(f, \bar{c}) \leq 1 - \kappa.
\end{aligned}$$

Consequently, we can rewrite the desired "symmetrised" fairness constraint as

$$\min(R_{\mathrm{fair}}(f), R_{\mathrm{fair}}(1-f)) \geq \tau \iff \mathrm{CS}_{\mathrm{bal}}(f; \bar{c}) \in [\kappa, 1-\kappa].$$

Consequently, for $\lambda_1, \lambda_2 \geq 0$, the corresponding Lagrangian version will be (see Appendix E for details)

$$\min_f \mathrm{CS}(f; D, c) + \lambda_1 \cdot (\mathrm{CS}_{\mathrm{bal}}(f; \bar{D}, \bar{c}) - (1-\kappa)) - \lambda_2 \cdot (\mathrm{CS}_{\mathrm{bal}}(f; \bar{D}, \bar{c}) - \kappa).$$

Letting $\lambda \doteq \lambda_1 - \lambda_2$ shows the result. ■

Proof [Proof of Proposition 4] By Lemma 9, the performance and fairness measures are

$$R_{\text{perf}}(f; D) = (1 - c) \cdot \pi + \mathop{\mathbb{E}}_{\mathsf{X}}[(c - \eta(\mathsf{X})) \cdot f(\mathsf{X})]$$
$$R_{\text{fair}}(f; \bar{D}_{\text{DP}}) = (1 - \bar{c}) \cdot \bar{\pi} + \mathop{\mathbb{E}}_{\mathsf{X}}[(\bar{c} - \bar{\eta}_{\text{DP}}(\mathsf{X})) \cdot f(\mathsf{X})]\,.$$

Ignoring constants independent of f, the overall objective is thus

$$\begin{aligned}
&\min_f R_{\text{perf}}(f; D) - \lambda \cdot R_{\text{fair}}(f; \bar{D}_{\text{DP}}) \\
&= \min_f \mathop{\mathbb{E}}_{\mathsf{X}}[((c - \eta(\mathsf{X})) - \lambda \cdot (\bar{c} - \bar{\eta}_{\text{DP}}(\mathsf{X}))) \cdot f(\mathsf{X})] \\
&= \min_f \mathop{\mathbb{E}}_{\mathsf{X}}[-s^*(\mathsf{X}) \cdot f(\mathsf{X})]\,.
\end{aligned}$$

Thus, at optimality, when $s^*(x) \neq 0$, $f^*(x) = [\![s^*(x) > 0]\!]$. When $s^*(x) = 0$, any choice of $f^*(x)$ is admissible. ■

Proof [Proof of Corollary 5] We simply apply Proposition 4 to $\bar{x} = (x, \bar{y})$. Note that

$$\begin{aligned}
\bar{\eta}_{\text{DP}}(x, \bar{y}) &= \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x, \bar{\mathsf{Y}} = \bar{y}) \\
&= [\![\bar{y} = 1]\!].
\end{aligned}$$

Then,

$$\begin{aligned}
f^*(x, \bar{y}) = 1 &\iff \eta(x, \bar{y}) > c + \lambda \cdot (\bar{\eta}_{\text{DP}}(x, \bar{y}) - \bar{c}) \\
&\iff \eta(x, \bar{y}) > c + \lambda \cdot ([\![\bar{y} = 1]\!] - \bar{c}).
\end{aligned}$$

■

Proof [Proof of Proposition 6] By Lemma 9, the fairness measure is

$$\begin{aligned}
R_{\text{fair}}(f; \bar{D}_{\text{EO}}) &= (1 - \bar{c}) \cdot \mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{Y} = 1) + \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=1}[(\bar{c} - \bar{\eta}_{\text{EO}}(\mathsf{X}, 1)) \cdot f(\mathsf{X})] \\
&= (1 - \bar{c}) \cdot \mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{Y} = 1) + \mathop{\mathbb{E}}_{\mathsf{X}}\left[\frac{\eta(\mathsf{X})}{\pi} \cdot (\bar{c} - \bar{\eta}_{\text{EO}}(\mathsf{X}, 1)) \cdot f(\mathsf{X})\right],
\end{aligned}$$

where the second line is from applying the importance weighting identity, and the fact that

$$\frac{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = 1)}{\mathbb{P}(\mathsf{X} = x)} = \frac{\mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x)}{\mathbb{P}(\mathsf{Y} = 1)} = \frac{\eta(x)}{\pi}.$$

Equivalently, for suitable λ, we seek

$$\begin{aligned}
&\min_f \mathop{\mathbb{E}}_{\mathsf{X}}\left[\left(c - \eta(\mathsf{X}) - \lambda \cdot \frac{\eta(\mathsf{X})}{\pi} \cdot (\bar{c} - \bar{\eta}_{\text{EO}}(\mathsf{X}, 1))\right) \cdot f(\mathsf{X})\right] \\
&= \min_f \mathop{\mathbb{E}}_{\mathsf{X}}[-s^*(\mathsf{X}) \cdot f(\mathsf{X})]\,.
\end{aligned}$$

Thus, at optimality, when $s^*(x) \neq 0$, $f^*(x) = [\![s^*(x) > 0]\!]$. When $s^*(x) = 0$, any choice of $f^*(x)$ is admissible. ■

Proof [Proof of Corollary 7] Plug in $\bar{\eta}(x, \bar{y}) = [\![\bar{y} = 1]\!]$ into Proposition 6. ■

Proof [Proof of Proposition 8] Let $R_{\text{perf}}(f; D) = \text{CS}(f; D, c)$. Observe that $f_0^* \in \operatorname{Argmin} R_{\text{perf}}(f; D)$. Consequently, the frontier is

$$F(\tau) = \text{reg}(f_\tau^*; D)$$

where the *regret* or *excess risk* of a classifier is

$$\text{reg}(f; D) = R_{\text{perf}}(f; D) - \min_{g\colon \mathcal{X}\to[0,1]} R_{\text{perf}}(g; D).$$

This lets us specify the form of $F(\cdot)$ analytically when R_{perf} is a cost-sensitive risk. By Lemma 11,

$$F(\tau) = \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[(c - \eta(\mathsf{X})) \cdot (f_\tau^*(\mathsf{X}) - [\![\eta(\mathsf{X}) > c]\!])\right].$$

Now, since f_τ^* is the solution to a linear program by Lemma 13, we can appeal to strong duality (see Appendix E) to conclude that there exists some λ for which the corresponding soft-constrained version of the problem (Equation 9) has the same optimal value. This means there is some Bayes-optimal classifier f_λ^* to Equation 9 for which

$$F(\tau) = \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[(c - \eta(\mathsf{X})) \cdot (f_\lambda^*(\mathsf{X}) - [\![\eta(\mathsf{X}) > c]\!])\right].$$

If additionally this f_λ^* is deterministic, then also by Lemma 11,

$$F(\tau) = \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[|\eta(\mathsf{X}) - c| \cdot [\![(\eta(\mathsf{X}) - c) \cdot (2f_\lambda^*(\mathsf{X}) - 1) < 0]\!]\right].$$

Now just plug in the definition of f_λ^* from Equation 15. ■

Appendix B. Helper results

We present some results here that are useful in establishing results in the body of the paper. In the following, let D be any distribution over $\mathcal{X} \times \{0, 1\}$.

Lemma 9 *Pick any randomised classifier f. Then, for any cost parameter $c \in (0, 1)$,*

$$\mathrm{CS}(f; c) = (1 - c) \cdot \pi + \mathop{\mathbb{E}}_{\mathsf{X}}[(c - \eta(\mathsf{X})) \cdot f(\mathsf{X})]$$

where $\pi = \mathbb{P}(\mathsf{Y} = 1)$, and $\eta(x) = \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x)$.

Proof [Proof of Lemma 9] By definition,

$$\begin{aligned}
\mathrm{CS}(f; c) &= (1 - c) \cdot \pi \cdot \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=1}[1 - f(\mathsf{X})] + c \cdot (1 - \pi) \cdot \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=0}[f(\mathsf{X})] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}}[(1 - c) \cdot \eta(\mathsf{X}) \cdot (1 - f(\mathsf{X})) + c \cdot (1 - \eta(\mathsf{X})) \cdot f(\mathsf{X})] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}}[(1 - c) \cdot \eta(\mathsf{X})] + \mathop{\mathbb{E}}_{\mathsf{X}}[(c \cdot (1 - \eta(\mathsf{X})) - (1 - c) \cdot \eta(\mathsf{X})) \cdot f(\mathsf{X})] \\
&= (1 - c) \cdot \pi + \mathop{\mathbb{E}}_{\mathsf{X}}[(c \cdot (1 - \eta(\mathsf{X})) - (1 - c) \cdot \eta(\mathsf{X})) \cdot f(\mathsf{X})] \\
&= (1 - c) \cdot \pi + \mathop{\mathbb{E}}_{\mathsf{X}}[(c - \eta(\mathsf{X})) \cdot f(\mathsf{X})] \,.
\end{aligned}$$

The second line is since $\mathbb{P}(\mathsf{X} \mid \mathsf{Y} = 1) \cdot \mathbb{P}(\mathsf{Y} = 1) = \mathbb{P}(\mathsf{X}) \cdot \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X})$. ■

Lemma 10 *Pick any cost parameter $c \in (0, 1)$. Let*

$$(\forall x \in \mathcal{X})\, s^*(x) = \eta(x) - c$$

where $\eta(x) = \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x)$. Then, any randomised classifier f^* satisfying

$$(\forall x \in \mathcal{X})\, s^*(x) \neq 0 \implies f^*(x) = [\![s^*(x) > 0]\!]$$

minimises $\mathrm{CS}(f; D, c)$.

Proof [Proof of Lemma 10] By Lemma 9, we need to find, for each $x \in \mathcal{X}$

$$\min_{f(x) \in [0,1]} (c - \eta(x)) \cdot f(x) = \min_{f(x) \in [0,1]} -s^*(x) \cdot f(x),$$

observing that the minimisation may be done pointwise. Clearly, it is optimal to predict $f^*(x) = 1$ when $s^*(x) > 0$, and $f^*(x) = 0$ when $\eta(x) < c$. When $\eta(x) = c$, any prediction is optimal. ■

Lemma 11 *Pick any cost parameter $c \in (0, 1)$. Then, for any randomised classifier f,*

$$\mathrm{CS}(f; c) - \min_{g\colon \mathcal{X} \to [0,1]} \mathrm{CS}(g; c) = \mathop{\mathbb{E}}_{\mathsf{X}}[(c - \eta(\mathsf{X})) \cdot (f(\mathsf{X}) - [\![\eta(\mathsf{X}) > c]\!])] \,.$$

If further $f \in \{0, 1\}^{\mathcal{X}}$,

$$\mathrm{CS}(f; c) - \min_{g\colon \mathcal{X} \to [0,1]} \mathrm{CS}(g; c) = \mathop{\mathbb{E}}_{\mathsf{X}}[|\eta(\mathsf{X}) - c| \cdot [\![(\eta(\mathsf{X}) - c) \cdot (2f(x) - 1) < 0]\!]] \,.$$

Proof [Proof of Lemma 11] By Lemma 10, an optimal classifier for $R_{\text{perf}}(g; D)$ is the deterministic $f^*(x) = [\![\eta(x) > c]\!]$. Thus, plugging this into Lemma 9,

$$\text{CS}(f; c) - \min_{g\colon \mathcal{X}\to[0,1]} \text{CS}(g; c) = \mathop{\mathbb{E}}_{\mathsf{X}} \left[(c - \eta(\mathsf{X})) \cdot (f(\mathsf{X}) - [\![\eta(\mathsf{X}) > c]\!])\right].$$

The second statement follows from a simple case analysis. The difference $f(x) - [\![\eta(x) > c]\!]$ takes on the value $+1$ when $f(x) = 1$ and $\eta(x) < c$, and -1 when $f(x) = 0$ and $\eta(x) > c$, i.e. the value $\text{sign}(c - \eta(x))$ when $2f - 1$ and $\eta(x) - c$ disagree in sign. Since $|z| = z \cdot \text{sign}(z)$, the result follows. ■

Lemma 12 *Pick any cost parameter $c \in (0, 1)$. Then,*

$$\min_{f\colon \mathcal{X}\to[0,1]} \text{CS}(f; c) = -\mathbb{I}_\varphi(\mathbb{P}(\mathsf{X} \mid \mathsf{Y} = 1), \mathbb{P}(\mathsf{X} \mid \mathsf{Y} = 0))$$

where $\mathbb{I}_f(\cdot, \cdot)$ denotes the f-divergence between distributions, and

$$\varphi(t) = -\min\left((1 - c) \cdot \pi \cdot t, c \cdot (1 - \pi)\right).$$

Proof [Proof of Lemma 12] This follows from Reid and Williamson (2011, Theorem 9), applied as follows. Let $f^*(x) = [\![\eta(x) > c]\!] + \frac{1}{2} \cdot [\![\eta(x) = c]\!]$, which is an optimal classifier for $\text{CS}(f; c)$ by Lemma 10. Then,

$$\text{CS}(f^*; c) = \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D} \left[\ell(\mathsf{Y}, \eta(\mathsf{X}))\right]$$

where ℓ is the *cost-sensitive loss* given by

$$\ell(1, v) = (1 - c) \cdot \left([\![v < c]\!] + \frac{1}{2}[\![v = c]\!]\right)$$
$$\ell(0, v) = c \cdot \left([\![v > c]\!] + \frac{1}{2}[\![v = c]\!]\right).$$

Now, ℓ is *proper* in the sense of Reid and Williamson (2010). Consequently,

$$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D} \left[\ell(\mathsf{Y}, \eta(\mathsf{X}))\right] = \min_{\hat{\eta}\colon \mathcal{X}\to[0,1]} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D} \left[\ell(\mathsf{Y}, \hat{\eta}(\mathsf{X}))\right].$$

The right hand side above is the *Bayes-risk* for the proper loss ℓ in the sense of Reid and Williamson (2011). Consequently, by Reid and Williamson (2011, Theorem 9),

$$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D} \left[\ell(\mathsf{Y}, \eta(\mathsf{X}))\right] = -\mathbb{I}_\varphi(\mathbb{P}(\mathsf{X} \mid \mathsf{Y} = 1), \mathbb{P}(\mathsf{X} \mid \mathsf{Y} = 0)),$$

where

$$\varphi(t) = -\min\left((1 - c) \cdot \pi \cdot t, c \cdot (1 - \pi)\right).$$

This may be verified easily, since

$$
\begin{aligned}
&\mathrm{CS}(f^*; c) \\
&= \mathop{\mathbb{E}}_{\mathsf{X}}\left[(1-c)\cdot\eta(\mathsf{X})\cdot\left([\![\eta(x)<c]\!]+\frac{1}{2}\cdot[\![\eta(x)=c]\!]\right)c\cdot(1-\eta(\mathsf{X}))\cdot\left([\![\eta(x)>c]\!]+\frac{1}{2}\cdot[\![\eta(x)=c]\!]\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}}\left[\min((1-c)\cdot\eta(\mathsf{X}), c\cdot(1-\eta(\mathsf{X})))\right],
\end{aligned}
$$

while, if $P = \mathbb{P}(\mathsf{X} \mid \mathsf{Y}=1), Q = \mathbb{P}(\mathsf{X} \mid \mathsf{Y}=0)$ with densities p, q,

$$
\begin{aligned}
-\mathbb{I}_\varphi(P,Q) &= -\mathop{\mathbb{E}}_{\mathsf{X}\sim Q}\left[\varphi\left(\frac{p(\mathsf{X})}{q(\mathsf{X})}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim Q}\left[\min\left((1-c)\cdot\pi\cdot\frac{p(\mathsf{X})}{q(\mathsf{X})}, c\cdot(1-\pi)\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[\min\left((1-c)\cdot\pi\cdot\frac{p(\mathsf{X})}{m(\mathsf{X})}, c\cdot(1-\pi)\cdot\frac{q(\mathsf{X})}{m(\mathsf{X})}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[\min\left((1-c)\cdot\eta(\mathsf{X}), c\cdot(1-\eta(\mathsf{X}))\right)\right] \\
&= \mathrm{CS}(f^*; c).
\end{aligned}
$$

■

Appendix C. On anti-classifiers

Employing a statistical risk for R_{fair} in Equation 8 constrains the false-positive and negative rates in some manner. However, these constraints may assume our classifier is non-trivial on $\bar{D}$. As an example, suppose a classifier f has $\text{MD}(f;\bar{D}) = \tau$. Then, it is easy to check that $\text{MD}(1-f;\bar{D}) = 1-\tau$. Consequently, when τ is small, simply by flipping around predictions, one achieves a poor fairness score.

Intuitively, one wishes to disallow such a trivial transformation from adversely affecting fairness. A simple option is to work with the *symmetrised* fairness measure $R_{\text{fair}}(f) \wedge R_{\text{fair}}(1-f)$, where $a \wedge b = \min(a,b)$. For cost-sensitive fairness measures, it is an easy calculation that for any $c \in (0,1)$ and classifier f,

$$\text{CS}_{\text{bal}}(1-f;c) = 1 - \text{CS}_{\text{bal}}(f;c).$$

Consequently,

$$\text{CS}_{\text{bal}}(f;c) \wedge \text{CS}_{\text{bal}}(1-f;c) = \text{CS}_{\text{bal}}(f;c) \wedge (1 - \text{CS}_{\text{bal}}(f;c));$$

further, this implies that constraints on the symmetrised fairness measure can be translated as

$$\text{CS}_{\text{bal}}(f;c) \wedge \text{CS}_{\text{bal}}(1-f;c) \geq \tau \iff \text{CS}_{\text{bal}}(f;c) \in [\tau, 1-\tau].$$

When converting this hard constraint to a soft constraint on the Lagrangian, one will thus obtain two non-negative Lagrange (KKT) multipliers λ_1 and λ_2, so that

$$\begin{aligned} &\min_f \text{CS}_{\text{bal}}(f;D,c) \colon \text{CS}_{\text{bal}}(f;\bar{D},\bar{c}) \in [\tau, 1-\tau] \\ \equiv &\min_f \text{CS}_{\text{bal}}(f;D,c) + (\lambda_2 - \lambda_1) \cdot \text{CS}_{\text{bal}}(f;\bar{D},\bar{c}) + \lambda_2 \cdot (1-\tau) + \lambda_1 \cdot \tau. \end{aligned}$$

One can then define $\lambda \doteq \lambda_2 - \lambda_1$ and use this to obtain Equation 9, noting that λ is now unconstrained.

Appendix D. The frontier problem as a linear program

Lemma 13 *For finite $\mathcal{X}$, pick any costs $c, \bar{c} \in (0,1)$, and $\tau \in \mathbb{R}_+$. Then, the problem*

$$\min_{f\colon \mathcal{X}\to[0,1]} \mathrm{CS}(f; D, c)\colon \mathrm{CS}(f; \bar{D}, \bar{c}) \geq \tau$$

is expressible as a linear program.

Proof [Proof of Lemma 13] By Lemma 9, cost-sensitive risks are linear in the randomised classifier. In particular, for discrete $\mathcal{X}$,

$$\begin{aligned}\mathrm{CS}(f; c) &= (1-c)\cdot\pi + \mathop{\mathbb{E}}_{\mathsf{X}}[(c-\eta(\mathsf{X}))\cdot f(\mathsf{X})] \\ &= (1-c)\cdot\pi + \sum_{x\in\mathcal{X}} m(x)\cdot(c-\eta(x))\cdot f(x),\end{aligned}$$

where $m(x) = \mathbb{P}(\mathsf{X} = x)$. Similarly,

$$\begin{aligned}\mathrm{CS}(f; \bar{c}) &= (1-\bar{c})\cdot\pi + \mathop{\mathbb{E}}_{\mathsf{X}}[(\bar{c}-\bar{\eta}(\mathsf{X}))\cdot f(\mathsf{X})] \\ &= (1-\bar{c})\cdot\pi + \sum_{x\in\mathcal{X}} m(x)\cdot(\bar{c}-\bar{\eta}(x))\cdot f(x).\end{aligned}$$

Now let

$$\begin{aligned}(\forall x \in \mathcal{X})\, a(x) &\doteq m(x)\cdot(c-\eta(x)) \\ (\forall x \in \mathcal{X})\, b(x) &\doteq m(x)\cdot(\bar{c}-\bar{\eta}(x))\,.\end{aligned}$$

Then, the optimisation is

$$\begin{aligned}\min_f a^T f\colon\ & -b^T f \leq -\tau \\ & 0 \leq f \leq 1.\end{aligned}$$

This is a linear objective with linear constraints. We thus may find the optimal random classifier by the solution to a linear program. ■

Appendix E. Relating the constrained and unconstrained objectives

Pick some cost-sensitive performance and fairness measures. Suppose $\mathrm{CS}^\diamond(f;\bar{c}) = \min(\mathrm{CS}(f;\bar{c}), \mathrm{CS}(1-f;\bar{c}))$. Consider the constrained version of the fairness problem,

$$f^* \in \operatorname*{Argmin}_{f\in[0,1]^{\mathcal{X}}} \mathrm{CS}(f;D,c)\colon \mathrm{CS}^\diamond(f;\bar{D},\bar{c}) \geq \tau.$$

By Lemma 13, for finite $\mathcal{X}$, this is expressible as the solution to a linear program

$$\min_{f\in\mathcal{F}} a^T f$$

where

$$\mathcal{F} = \{f \mid b^T f \in [\tau, 1-\tau], 0 \leq f(x) \leq 1\}$$

and

$$\begin{aligned}(\forall x \in \mathcal{X})\, a(x) &\doteq m(x)\cdot(c-\eta(x))\\ (\forall x \in \mathcal{X})\, b(x) &\doteq m(x)\cdot(\bar{c}-\bar{\eta}(x))\,.\end{aligned}$$

Now, by strong duality for linear programs[5], we have

$$\min_{f\in\mathcal{F}} a^T f = \max_{\lambda_1,\lambda_2\geq 0}\left(\min_{f\in[0,1]^{\mathcal{X}}} (a-\lambda_1 b+\lambda_2 b)^T f\right) + \lambda_1\tau - \lambda_2(1-\tau). \tag{22}$$

Observe now that the inner optimisation is

$$\begin{aligned}&\min_{f\in[0,1]^{\mathcal{X}}} (a-\lambda_1 b+\lambda_2 b)^T f\\ &= \min_{f\in[0,1]^{\mathcal{X}}} (a-(\lambda_1-\lambda_2) b)^T f\\ &= \min_{f\in[0,1]^{\mathcal{X}}} \sum_{x\in\mathcal{X}} m(x)\cdot[c-\eta(x)-(\lambda_1-\lambda_2)(\bar{c}-\bar{\eta}(x))]\cdot f(x)\\ &= \min_{f\in[0,1]^{\mathcal{X}}} \mathop{\mathbb{E}}_{\mathsf{X}}[(c-\eta(\mathsf{X})-(\lambda_1-\lambda_2)(\bar{c}-\bar{\eta}(\mathsf{X})))\cdot f(\mathsf{X})]\\ &= \min_{f\in[0,1]^{\mathcal{X}}} \mathrm{CS}(f;D,c)-(\lambda_1-\lambda_2)\cdot\mathrm{CS}(f;\bar{D},\bar{c}).\end{aligned}$$

That is, we solve Equation 9 for $\lambda = \lambda_1 - \lambda_2$. By sweeping over λ, we can thus in principle find the one which achieves the highest value of the objective in Equation 22, and consequently find the solution to the constrained problem for a fixed τ.

Note that strong duality guarantees agreement of the objective functions. In general, it does not mean that every optimal solution to the inner problem (for optimal λ_1, λ_2) will also be optimal for the original constrained problem. As an extreme case, suppose that $\bar{\eta} = \eta$, and $c = \bar{c}$. Then, the constrained problem has optimal solution any f for which $\mathrm{CS}(f;\bar{D},\bar{c}) = \tau$, so that the frontier is linear. On the other hand, we will find that for the optimal λ_1, λ_2, the inner optimisation is simply of the constant 0, in which case every f is deemed optimal.

5. This implicitly assumes feasibility of the primal problem, i.e. that we pick τ such that it is possible to find a randomised classifier with symmetrised fairness at least τ.

Appendix F. Relating disparate impact and balanced error

Following Feldman et al. (2015), we explore the relationship between the balanced error and disparate impact. Intuitively, we expect that when the balanced error of a classifier is low – meaning that it accurately predicts the sensitive variable – we will have disparate impact. Conversely, we might hope that possessing disparate impact implies low balanced error. Can we formalise a relationship akin to Lemma 2?

In what follows, for an implicit distribution $\bar{D}$, let

$$\mathrm{BER}(f) \doteq \frac{\mathrm{FPR}(f) + \mathrm{FNR}(f)}{2}.$$

We have the following relations between the two quantities.

Lemma 14 *Pick any randomised classifier* $f\colon \mathcal{X} \to [0,1]$ *with* $\mathrm{FNR}(f) \neq 1$. *Then,*

$$\begin{aligned} \mathrm{DI}(f) &= \frac{\mathrm{FPR}(f)}{1 - 2 \cdot \mathrm{BER}(f) + \mathrm{FPR}(f)} \\ &= \frac{2 \cdot \mathrm{BER}(f) - \mathrm{FNR}(f)}{1 - \mathrm{FNR}(f)}, \end{aligned} \tag{23}$$

and similarly

$$\begin{aligned} \mathrm{BER}(f) &= \frac{1}{2} \cdot \mathrm{FNR}(f) + \frac{1}{2} \cdot (1 - \mathrm{FNR}(f)) \cdot \mathrm{DI}(f) \\ &= \frac{1}{2} \cdot \mathrm{FPR}(f) + \frac{1}{2} \cdot \left(1 - \frac{\mathrm{FPR}(f)}{\mathrm{DI}(f)}\right). \end{aligned} \tag{24}$$

Proof [Proof of Lemma 14] These are trivial consequences of the fact that, by definition of $\mathrm{DI}(f)$,

$$\mathrm{FPR}(f) = (1 - \mathrm{FNR}(f)) \cdot \mathrm{DI}(f).$$

■

We now turn to relating a bound on the balanced error to a bound on the disparate impact factor. The following is a minor generalisation of Feldman et al. (2015, Theorem 4.1) to account for disparate impact at any level.

Lemma 15 *Pick any randomised classifier* $f\colon \mathcal{X} \to [0,1]$ *with* $\mathrm{FNR}(f) \neq 1$. *Then, for any* $\epsilon \in [0, \frac{1}{2}]$,

$$\mathrm{BER}(f) \leq \epsilon \iff \mathrm{DI}(f) \leq \frac{\mathrm{FPR}(f)}{1 - 2 \cdot \epsilon + \mathrm{FPR}(f)} \wedge \frac{2 \cdot \epsilon - \mathrm{FNR}(f)}{1 - \mathrm{FNR}(f)},$$

and for any $\tau \in [0,1]$,

$$\mathrm{DI}(f) \leq \tau \iff \mathrm{BER}(f) \leq \left(\frac{\tau}{2} + \frac{1-\tau}{2} \cdot \mathrm{FNR}(f)\right) \wedge \left(\frac{1}{2} - \frac{1-\tau}{2 \cdot \tau} \cdot \mathrm{FPR}(f)\right).$$

Proof [Proof of Lemma 15] The first equivalence follows from the two expressions in Equation 23, and the fact that the dependence on $\mathrm{BER}(f)$ is monotone increasing.

This second equivalence follows from the two expressions in Equation 24, and the fact that the dependence on $\mathrm{DI}(f)$ is monotone increasing. ■

When $\tau = 0.8$, as considered in Feldman et al. (2015), this means that

$$\mathrm{DI}(f) \leq 0.8 \iff \mathrm{BER}(f) \leq \left(\frac{2}{5} + \frac{1}{10} \cdot \mathrm{FNR}(f)\right) \wedge \left(\frac{1}{2} - \frac{1}{8} \cdot \mathrm{FPR}(f)\right).$$

F.1. Low balanced error implies disparate impact

It is of interest to remove the dependence of the above bounds on the false positive and negative rates of f. For one direction, this is possible.

Corollary 16 *Pick any randomised classifier* $f : \mathcal{X} \to [0, 1]$ *with* $\mathrm{FNR}(f) \neq 1$. *Then, for any* $\epsilon \in [0, \frac{1}{2}]$,

$$\mathrm{BER}(f) \leq \epsilon \implies \mathrm{DI}(f) \leq 2 \cdot \epsilon,$$

or for any $\tau \in [0, 1]$,

$$\mathrm{DI}(f) \geq \tau \implies \mathrm{BER}(f) \geq \frac{\tau}{2}.$$

Proof The first bound follows from Lemma 15 and the fact that if $\mathrm{BER}(f) \leq \epsilon$, it must be true that $\mathrm{FPR}(f) \vee \mathrm{FNR}(f) \leq 2 \cdot \epsilon$. The second bound is the contrapositive of the first. ■

Corollary 16 says that with a balanced error of $\frac{\tau}{2}$ or less, we are guaranteed a disparate impact of level at least τ, though possibly worse. So, if we want to guarantee a lack of disparate impact at level τ, it is *necessary* that the balanced error be at least $\frac{\tau}{2}$. But is this condition also *sufficient*? Unfortunately, it is not.

F.2. Disparate impact does not imply low balanced error

It is evident from Lemma 15 that regardless of the precise level of impact τ, we *could* have a classifier with balanced error arbitrarily close to $\frac{1}{2}$. The basic issue is that by driving the false positive rate to 0, we trivially have disparate impact. By further driving the false negative rate to 0 (i.e. by predicting everything negative), we trivially have a balanced error rate of $\frac{1}{2}$.

Corollary 17 *Pick any* $\tau \in [0, 1]$ *Then, there exists a classifier* $f : \mathcal{X} \to \{0, 1\}$ *with*

$$\mathrm{BER}(f) = \frac{1}{2}$$

$$\mathrm{DI}(f) \leq \tau.$$

Proof Consider the trivial classifier with

$$\mathrm{FPR}(f) = 0$$

$$\mathrm{FNR}(f) = 1.$$

Clearly, this has balanced error $\frac{1}{2}$. Evidently, this classifier also has disparate impact at level τ. ■

Corollary 17 says that even if we have a classifier with high balanced error, there is no guarantee it will not have disparate impact. This is a worst case analysis over all possible classifiers we *might* have obtained. However, if we happen to know the false positive and negative rates we *actually* have obtained, we might be able to conclude there is no disparate impact. This is used in Feldman et al. (2015, Section 4.2) to certify the lack of disparate impact for a particular classifier.

Corollary 18 *Pick any randomised classifier* $f : \mathcal{X} \to [0, 1]$. *For any* $\tau \in [0, 1]$,

$$\mathrm{BER}(f) \geq \left(\frac{\tau}{2} + \frac{1 - \tau}{2} \cdot \mathrm{FNR}(f)\right) \wedge \left(\frac{1}{2} - \frac{1 - \tau}{2 \cdot \tau} \cdot \mathrm{FPR}(f)\right) \iff \mathrm{DI}(f) \geq \tau.$$

Proof [Proof of Corollary 18] This is the contrapositive of Lemma 15. ■

Appendix G. Experimental illustration

We present experiments that illustrate each of our three contributions **C1**–**C3**, showing that

- relating fairness measures to cost-sensitive risks (**C1**) can be used to certify a dataset as free of disparate impact.
- the plugin approach inspired by the form of the optimal fairness-aware classifiers (**C2**) is practically viable.
- the inherent accuracy-fairness tradeoff (**C3**) behaves as predicted by our quantification in terms of probability alignment.

G.1. Certifying lack of disparate impact

We present an experiment inspired by Feldman et al. (2015), who aimed to certify whether a dataset admits disparate impact (i.e. for fixed τ, one can achieve $\mathrm{DI}(f) \leq \tau$) by testing if the minimal achievable balanced error is below a classifier dependent threshold (see Appendix F). Following Remark 4.2, we verify that such certification is equivalently possible via testing if the minimal achievable balanced cost-sensitive risk for $\bar{c} = 1/(1+\tau)$ is below $1 - \bar{c}$.

Specifically, following Feldman et al. (2015), we consider the UCI `german` dataset with $\bar{\mathsf{Y}}$ denoting whether or not the age of a person is above 25, and fix $\tau = 0.8$. For a number of train-test splits to be specified, we train models to minimise the cost-sensitive logistic loss with parameter c (per Equation 17), and evaluate on the test set the disparate impact, as well as the gap $\Delta(f) \doteq \mathrm{CS}_{\mathrm{bal}}(f; \bar{c}) - (1 - \bar{c})$. Our Lemma 1 indicates that we should find the latter to be positive only when the former is larger than $\tau = 0.8$.

To construct training sets, we make an initial 2:1 train-test split of the full data, treating $\bar{\mathsf{Y}}$ as the label to predict. To obtain models with varying levels of accuracy in predicting $\bar{\mathsf{Y}}$, we inject symmetric label noise of varying rates into the training set. Figure 1(*a*) shows that for the resulting models, as per Lemma 1, there is perfect agreement of disparate impact at $\tau = 0.8$ and $\mathrm{sign}(\Delta(f))$, evidenced by there being no points in the top-left and bottom-right quadrants.

G.2. A plugin approach to the fairness problem

We next present an experiment inspired by Zafar et al. (2016), where on the same `german` dataset we learn a classifier that respects a MD score constraint, while being accurate for predicting the target feature in the sense of balanced error (BER). We employ the plugin estimator proposed in §5.2 (which we term `2LR`), training logistic regression to predict the target and sensitive feature and combining these models via Equation 15 for some $\lambda \in \mathbb{R}$. On the test set, we compute the BER for the target feature, and the MD score for the sensitive feature. We compare this to the `COV` method of Zafar et al. (2016), which uses a surrogate to the MD constraint as discussed in §5.3, with tuning parameter $\tau \in \mathbb{R}_+$.

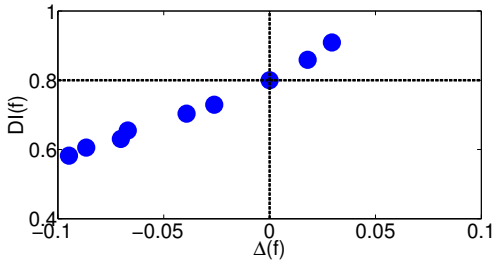
Varying $\lambda \in \{-1, -0.495, \ldots, 0\}$ and $\tau \in \{0, 0.005, \ldots, 1\}$ yields tradeoff curves for both methods. Figure 1(*b*) shows these curves at high fairness value, where we see that our plugin approach is generally competitive with `COV`, resulting in lower BER at higher fairness levels. (Note that `COV` does not sweep out the same range of lower fairness values as `2LR` owing to the use of a surrogate constraint.) Further, as mentioned in §5.2, tuning of λ with `2LR` requires no model training – to generate the curves in Figure 1(*b*) takes `2LR` **0.25 seconds**, as opposed to **25 seconds** for `COV`.

We present a further experiment on the synthetic dataset considered by Zafar et al. (2016), where $\mathbb{P}(\mathsf{Y} = 1) = 0.5$, each $\mathsf{X} \mid \mathsf{Y} = y \sim \mathcal{N}(\mu_y, \Sigma_y)$ where

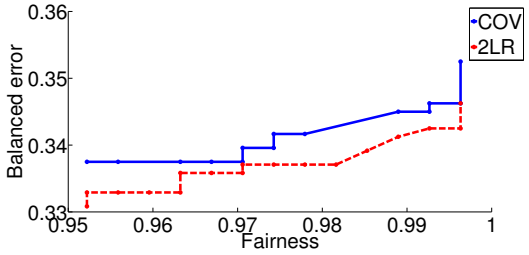
$$\mu_1 = \begin{bmatrix} 2 & 2 \end{bmatrix} \quad \mu_0 = \begin{bmatrix} -2 & -2 \end{bmatrix}$$
$$\Sigma_1 = \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix} \quad \Sigma_0 = \begin{bmatrix} 10 & 1 \\ 1 & 3 \end{bmatrix},$$

and

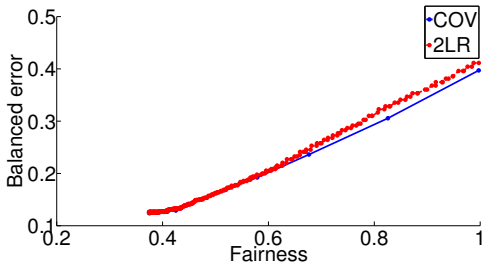
$$\mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{X} = x) = \frac{\mathbb{P}(\mathsf{X} = Rx \mid \mathsf{Y} = 1)}{\mathbb{P}(\mathsf{X} = Rx \mid \mathsf{Y} = 1) + \mathbb{P}(\mathsf{X} = Rx \mid \mathsf{Y} = -1)}$$



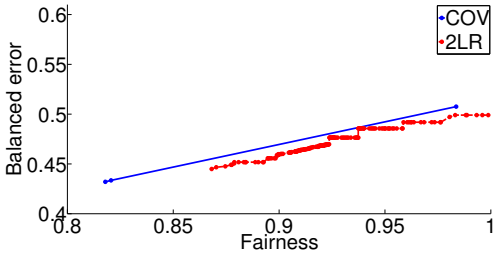
(*a*) Certification of disparate impact via cost-sensitive risk on `german` dataset. Each point is a different learned classifier.



(*b*) Comparison of plugin (`2LR`) and `COV` methods as tuning parameters for both are varied on `german` dataset.

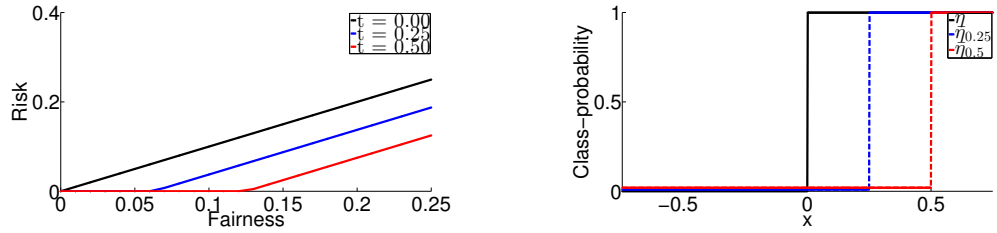


(*c*) Raw features.

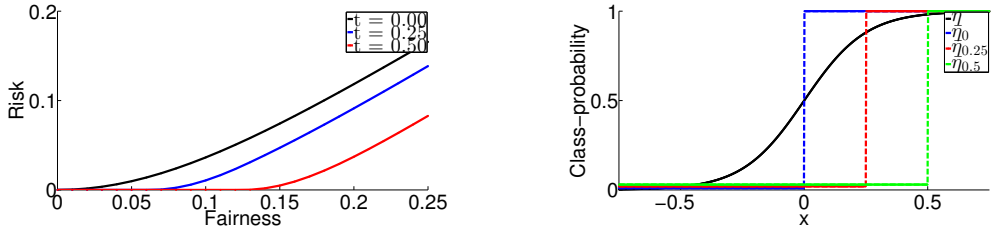


(*d*) Quadratic features.

Figure 1: Comparison of plugin (`2LR`) and `COV` methods as tuning parameters for both are varied, synthetic 2D Gaussian data.



(*a*) $\eta, \bar{\eta}$ given by indicator functions



(*b*) η given by sigmoid function, $\bar{\eta}$ by indicator function.

Figure 2: Fairness frontiers and probability disalignment; see text for description of parameter t.

for rotation matrix $R = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$. We pick $\phi = 0.5$.

We generated $N = 10^4$ samples from this distribution, and followed the same setup as before: we constructed a 2:1 train-test split, and compared COV and our plugin (2LR) approach in terms of the balanced error of predicting Y, versus the MD score in predicting $\bar{\mathsf{Y}}$.

Figure 1(*c*) shows the tradeoff curves of the methods closely track each other. However, the plugin approach performs slightly worse at higher fairness levels. We conjectured this is due to the fact that logistic regression is not suitable for $\eta, \bar{\eta}$, as the class-conditionals have non-isotropic covariance and thus possess *quadratic* boundaries. Indeed, when we explicitly include quadratic features as input to both methods, Figure 1(*d*) shows that the plugin approach performs consistently (albeit slightly) better than COV.

While a more exhaustive experimentation is needed to make firm conclusions, the above illustrates that our Bayes-optimal analysis of Problem 3.3 is not *solely* of theoretical interest, and may be useful in designing fairness-aware classifiers.

G.3. Illustrations of the fairness frontier

We illustrate the fairness frontier (Equation 19) for some simple distributions. Consider first the case where $\mathcal{X}$ comprises 100 linearly spaced points in $[-1, 1]$, D is such that $\eta(x) = [\![x > 0]\!]$ and the marginal over instances is uniform. Suppose also that $\bar{\eta}(x) = [\![x > t]\!]$, for tuning parameter t. Consider a cost-sensitive performance and fairness measure with $c = \bar{c} = 1/2$. We can explicitly compute the frontier here, shown in Figure 2(*a*) for a range of t. As t increases, η and $\bar{\eta}$ grow increasingly dissimilar, and so the fairness constraint does not affect performance as dramatically: this is manifest in the fact that the benign fairness value τ^* increases with t. Further, for every t, when there is a tradeoff, it is linear.

Suppose we instead have $\eta(x) = (1 + \exp(-x))^{-1}$, and retain the same $\bar{\eta}$. Here again, Figure 2(*b*) shows that as t increases, η and $\bar{\eta}$ grow increasingly dissimilar. The impact of changing the shape of η is the effect on the frontier when it is nonzero: as per Equation 21, this depends on the deviation of η from c, and hence the frontier here is nonlinear.

Appendix H. Fairness in Context

In this appendix we tentatively survey some of the broader literature that bears upon the problem of fairness in machine learning, especially as it relates to the perspective of the present paper (with a focus on measures such as Disparate Impact and Mean Difference that can be equivalently represented as a difference in risks[6]). The appendix does not claim to be comprehensive. And neither does it stake out a "position." It *does* illustrate the rich set of connections that remain to be explored in further developing an analytical theory of fairness in data analytics. We believe that machine learning has much to learn from the extant work on fairness, from a range of diverse disciplines.

We start by examining the meaning of fairness and its connections to justice (§H.1). We then argue that two of the most famous proposals for justice in terms of fairness are intrinsically information-theoretic (§H.2). The use of protected attributes (designating group membership) is central to most approaches to fair ML; we survey the significance of groups, the choice of groups, and their use and abuse in statistics (§H.3). We then briefly survey the literature on fairness and ethics as a natural phenomenon (§H.4). We conclude this appendix with a list of challenges for fair ML which we believe are novel (§H.5).

H.1. Fairness and Justice

"Fairness" is challenging to define and comprehend because (as a word) it carries significant cultural heritage, mostly originating in the Anglo culture (Wierzbicka, 2006, pages 141–167). It is *not* a cultural universal. Indeed, "fair" is an innocuous sounding word of a kind that "may be invisible to native speakers, who simply take them for granted and assume that they must have their equivalents in other languages" (Wierzbicka, 2006, pages 10–11), but often there are not equivalent words because the very concept is missing from the culture in which the language developed[7]. There is a considerable richness and complexity to what, exactly, is "fair" and there seems little hope of a single simple definition that works in all situations, and for all people.

The Anglo notion of Fairness is intimately related to Justice, (arguably) "the first virtue of social institutions" (Rawls, 1971, p586); "when Justice Potter Stewart first joined the Supreme Court in 1958, he said, 'fairness is what justice really is'" (Dorans and Cook, 2016). But there is no consensus that this is actually the case; see the series of tables in (Kolm, 1996, pages 77–83) summarising the (largely incommensurate) proposals that have been made to date. One of the most famous proposals for defining Justice is (somewhat surprisingly) information-theoretic. John Rawls' famous theory of "justice as fairness" has at its heart an information throttling device (the "veil of ignorance") whereby a given person's position in society is held to be unknown (to the person) while designing the rules of a just society. (This information theoretic observation is expanded in the next subsection). While his theory as stated is philosophical, and not mathematical, Rawls recognised the necessity of a mathematical treatment of fairness:

> A correct account of moral capacities will certainly involve principles and theoretical constructions which go beyond the norms and standards cited in everyday life; it may eventually require fairly sophisticated mathematics as well (Rawls, 1971, p47).

Variants of Rawls' theory, especially due to Harsanyi (1955), who anticipated him, *are* mathematical, and are well aligned with the approach of the present paper. The notion of group membership is often assumed to be essential in defining fairness; indeed, a "protected attribute" typically denotes group membership. But one should not take this group-theoretic notion of fairness for granted (see below, subsection H.3). Certain notions of fairness (wealth redistribution and a social safety net) were key to the New Zealand elections in the 1990s. But was it fairness "for" groups or "for" individuals? Fischer (2012)[8] said that the New Zealand Royal

6. There is a wide range of fairness measures considered in the ML literature. They are usually incommensurate with each other in the sense that they can not be optimized simultaneously (Kleinberg et al., 2016). The recent survey by Žliobaitė (2017) enumerates many of the fairness measures proposed to date, and while drawing some measures from literature outside of machine learning, does not deal with the problem of how one might select from these measures in a principled fashion. Some of the literature we survey in this appendix attempts to do that. We do not claim they are successful in doing so, nor even that it is possible in general to do so!
7. See the discussion in (Sen, 2009, p72) regarding the distinction between the concepts of justice and fairness.
8. Fischer credibly claims (Appendix — Fairness in Other Disciplines) that his is the "first book to be published on the history of fairness". His appendix contains a valuable survey of fairness in other disciplines, including linguistics, moral philosophy,

Commission on Social Policy (1988) "argued that a fair society should do justice not to groups or classes of people, but to individuals according to their rights, needs, and just 'deserts.'" The crucial point is the demand was for justice for *individuals*, not *groups*. The tension between fairness for groups or individuals recurs throughout the literature, and is especially relevant to the problem of fairness in machine learning.

Much of the machine learning literature on fairness is based on the notion that membership of a given category should not cause statistical decisions to "discriminate". This is sometimes taken to mean that outcomes need to be the same regardless of membership in a category. Analogous theories such as that of (Harsanyi, 1955) and (Sen, 2009) differ in what precisely "ignorance" means (uniform prior over what role one has, or the perspective of a separate impartial observer). In all cases, the general idea is that a just outcome should not depend (either way) on membership of a particular category, but should focus upon the individual; Rawls (1971, page 26) motivates his approach by saying "utilitarianism does not take seriously the distinction between persons" because it is only concerned with *average* welfare. These ideas have had profound impact on political philosophy, and there already exist a range of mathematical theories motivated by them (Binmore, 1994, 2005; Fleurbaey and Maniquet, 2011; Harsanyi, 1977c; Sen, 1986).

Rawls argued the advantages of "pure procedural justice" where the attainment of justice (hence fairness) is a consequence of the process followed, not the outcome obtained. This is, of course, taken for granted in machine learning and statistics, where the statistician typically analyses a procedure (an estimator), rather than the results of the procedure on a given set of data. Taking this principle seriously means that protected attributes have to be identified ahead of time, and not after the fact because you did not like the outcome; confer the analogous problem in designing electoral districts in subsection H.2 below.

Much of the philosophical literature on fairness (including that of Rawls) is motivated by the ultimate question of designing a perfect society. The intrinsic pragmatism and empiricism of machine learning suggests that instead it is better to follow the precept of Sen (2009, Chapter 18) that mere identification of "fully just social arrangements is neither necessary nor sufficient" (to make the present world a better place). As with any other engineering problem, one never attains perfection, but rather one identifies the fundamental trade-offs. We embrace Sen's pragmatism by focussing on the *quantifiable trade-offs* one might make to approach (certain notions of) fairness and hence justice. Rawls (1971, p37) acknowledges that there will have to be trade-offs between overall social utility and fairness, but he tries to argue what the "right" trade-off is. We do not try to solve that question, believing instead there is unlikely to be a universal right trade-off, and instead focus upon merely quantifying what the trade-offs might be. Our goal is to merely parametrise this (i.e. define the Pareto frontier), not to pick a particular point on the frontier[9].

H.2. Rawls' and Harsanyi's Notions of Fairness are Information-Theoretic

Whate'er the passion, knowledge, fame or pelf,
Not one will change his neighbour with himself.
—Alexander Pope, *Essay on Man* (II)

Rawls (1971) and Harsanyi (1958, 1977a,b,c, 1978, 1979) independently proposed approaches to fairness. While usually explained in terms of a "veil of ignorance" or an "impartial observer,"[10] from an engineer's

behavioural sceinces, genetics and evolution, neuroscience, social and cultural sciences, economics, and mathematics. It is a fine starting point to explore this diverse threads of fairness.

9. Focussing upon quantifying the trade-off between utility and fairness has drawn attention in the machine learning literature only recently (Johnson et al., 2016). Zafar et al. (2017), in talking about fairness in pretrial risk assessments (where one might be required to make the false positive rates equal for all groups), quote William Blackstone's famous aphorism "[T]he law holds it is better that ten guilty persons escape, than that one innocent party suffer" (Blackstone, 1765). This is a trade-off, of sorts, between utility and fairness. It begs the question: Why 10? Why not some other value n? Of course there is no especial reason to choose 10, as has been documented by Volokh (1997) with admirable scholarship. He cites seperate authorities claiming that the right choice of n is, respectively, $\frac{1}{54}$, $\frac{1}{10}$, 1, 2, 3, 4, 5, 9, 10, 12, 20, 99, 100, 599, 900, 1000, 5000, ∞, "a few", "some", "several", "many", "a considerable amount", or "a goodly number"! (Numbers less than 1 mean, taking $\frac{1}{10}$ for example, that it is considered better that 10 innocent men suffer than one guilty man escape). In the spirit of data analytics, it is possible to empirically estimate the value of n at a given time and place (by looking at various court statistics). Volokh (1997) estimated the value in 1997 in the US was 59.72.

10. "As Harsanyi has often and forcefully argued, an important basis for utilitarianism's claim to be morally compelling is that it can claim to represent the judgment of a sympathetic but impartial observer — one who accepts the interests of other persons and seeks

perspective it is perhaps better described as an information throttling device — a "bottleneck" of sorts. The idea is that in order to determine what is fair, one needs to *not* know one's place in the system one is designing. If designing the rules of a society, one must not know one's place in the society, else it is impossible to prove one's design choices were not in fact motivated by self interest. This "veil of ignorance" hides the morally relevant information. The particular bottlenecks or veils chosen by Rawls and Harsanyi differ:

> Behind Rawls's 'thick' veil of ignorance, individuals do not know the likelihood that they may end up in any given person's position. Behind Harsanyi's 'thin' veil of ignorance, they know that they have an equal chance of being in any person's position (Kurtulmus, 2012, p41).

Harsanyi (2008) has argued that his 'thin' veil is a better choice, and that the choice of a uniform prior is not only a more rational choice, but is more robust too. Rawls argues that since one does not know one's position at all (Knightian uncertainty), and thus one can reason in terms of the *worst case* — his maximin rule. Harsanyi, on the other hand, by assuming the uniform prior, can reason probabilistically (Knightian risk) and one should reason in terms of the *average* case, and in doing so throttles the relevant information:

> The idea of deriving substantive principles of morality based on rational decision making in a hypothetical situation in which the decision maker is *deprived of morally irrelevant information* is one of Harsanyi's greatest achievements (Fleurbaey et al., 2008, p15).

One interesting aspect of Harsanyi's solution is that his utilities (which correspond roughly to our loss functions) are 'cardinal' (in the terminology of economists), meaning they are what we would call scores, rather than just rank orderings. Furthermore, his fairness proposal relies upon the "interpersonal comparison" of these utilities (Hammond, 1991). The form of our overall objective function, a difference of two loss functions, implicitly makes such a comparison. (We are glossing over the issue of exactly what the utilities represent in the ML fairness problem; they are usually *not* the utilities of the people who are the subject of the data analysis, although the fairness loss $\bar{\ell}$ is arguably correlated with that.) The richer information conveyed by cardinal utilities completely bypass vexing negative results such as Arrow's impossibility theorem precisely because this comparison becomes possible; see (Harsanyi, 1977c) for details.

We speculate that by looking at a richer class of information throttling devices (especially, for example, those that only *partially* destroy information, such as utilising noisy labels and devices such as used in differential privacy) we may be able to develop a richer class of theories of fairness in machine learning. The hope is that the impact of an information-theoretic view will be as signifcant as that view was in economics (Stiglitz, 2002). We suspect that focussing on the information-theoretic limits may be more helpful than concentrating on different notions of equity (Young, 1994).

H.2.1. An Analogy: Fairness and the Prevention of Gerrymandering

The information throttling aspect of Harsanyi's model of fairness can be illustrated by an analogy. Consider the problem of the "fair" subdivision of a region into electoral districts. The importance of the problem is demonstrated by the fact that there is a name for what happens when this is not done fairly — gerrymandering.

Much of the literature on the prevention of gerrymandering attempts to solve the problem *ex post* — that is, to detect that gerrymandering has occurred by detecting unusual or "bizarre" shapes of the boundaries drawn. There are various proposal for the "compactness" (Young, 1988) or "bizarreness" (Chambers and Miller, 2010) of boundaries. But this is akin to judging whether overfitting has occurred in a statistical estimation problem by *only* examining the resulting hypothesis. We know, however, that this cannot work. One needs to control the complexity of all possible hypothesis *before* the inference step is done. Such a proposal is precisely what was made by William Vickrey[11]:

to resolve conflicts among them in a disinterested way" (Wagner, 1980). But this way of looking at matters is tricky! See the discusson of the limits of personal identity in (Adler, 2014).

11. Fleurbaey et al. (2008) (and indeed Harsanyi himself) credit the idea of limiting the information available in order to guarantee fairness to even earlier work by Vickrey (1945). The information-hiding idea also arises in the neuroscience of fairness in human behaviour; see §H.4 below.

> [I]t means that the selection of the process, which must itself be at least initiated by human action, should be as far removed from the results as possible, in the sense that it should not be possible to predict in any detail the outcome of the process (Vickrey, 1961, p106).

Vickrey extolls the virtue of randomization as an information destroying device:

> This means that at some stage in the process a random element should be introduced which, while not affecting the general principles that are to be applied, will substantially affect the result in the particular instance, so that legislators in selecting one process rather than another will not be able to base their choice on particular detailed outcomes and will be compelled to base their choice on general principles (Vickrey, 1961).

Vickrey's point is that the only way to guarantee no untoward manipulation occurs is to *hide the relevant information from the decision maker* (the person drawing up the boundary). Thus if they are required to draw the boundaries *before* they know the outcomes of the way individual people vote, this would suffice. A proxy would be to choose a class of shapes of boundaries *a priori* as a way to control the degree of strategic overfitting that is allowed.

H.2.2. Fairness *Ex Ante* or *Ex Post* — Temporal Inconsistencies in Human Behaviour

The veil of ignorance and other such information throttling devices are intrinsically *ex ante*. While people accept the logic of this in the abstract, when put into practice it is not so simple. Andreoni et al. (2016) found in a lab experiment that:

> [M]ost people view fairness from an *ex ante* perspective when making decisions *ex ante*, and from an *ex post* perspective when making decisions *ex post*.[12]

The matter is subtle; Trautmann and van de Kuilen (2016), who used the different distinction of *process* vs *outcome* fairness found "a significant share of people subscribe to process fairness both before and after the resolution of uncertainty." Trautmann and van de Kuilen (2016) discuss an example due to Machina (1989, p1643) which makes it clear what the issue is: the difference between fairness in *a priori* probabilities (*ex ante*) and unfair mutually exclusive outcomes once the random outcome occurs (*ex post*).

This phenomonenon has been called "moral luck" (Nagel, 1979) — the idea that one's blame should be in terms of one's choices (things under one's control) rather than exogenous things not under one's control:

> Anything which is the product of happy or unhappy contingency is no proper object of moral assessment, and no proper determinant of it, either (Williams, 1981, p20).

Vickrey's solution to gerrymandering, and Harsanyi's solution to fairness are both *ex ante* and defined in terms of probabilities not outcomes. It seems there is no alternative in the design of fair systems. But the evidence above suggests we should not be surprised nevertheless at complaints regarding the *outcomes* of such "fair" systems.

H.3. Taking Groups for Granted

The approach to fairness in the machine learning literature, which we follow in this paper, focusses on the notion of a protected attribute (typically indicating membership of a group), and assumes that both its choice is manifest, and indeed it is a sensible categorisation. But such categories to which people are assigned are not intrinsic to the world, but a choice that we make (Lakoff, 1987), and can be highly ambiguous and contested (Bowker and Star, 1999). In this section we will consider some of the complexities and subtleties regarding group membership that are largely taken for granted in the machine learning literature.

12. They continue:

> As a result, they exhibit the hallmark of time-inconsistency: after making an initial plan that is fully state-contingent, they revise it upon learning that certain states will not occur. These patterns are robust and persist even when people are aware of their proclivities. Indeed, subjects who switch from *ex ante* fair to *ex post* fair choices, and who are aware of this proclivity, generally avoid precommitments and intentionally retain the flexibility to manifest time inconsistency.

H.3.1. The Significance of Groups — Categorical Nominalism

Singling out *particular* attributes to be protected (as opposed to the rather more sweeping requirements of Rawls' full theory that requires no specific attributes be taken into account) opens the door to the problems inherent in the "ecological fallacy" (Kramer, 1983) — making inferences about individuals based on membership in a category, which is precisely what some argue one should not do (the right to be treated as an individual) (Lippert-Rasmussen, 2011).

The notions of fairness considered in the present paper (and indeed in most of the machine learning literature on the topic) take for granted that the are groups to which people belong (and which are the subject of fair learning algorithms) really exist in some sense. But it can be argued that these groups are purely conventions; this is the philosophical position called "categorical nominalism", and it has a long history. Desrosières (1998), in the context of "averages and the realism of aggregates", explained how William of Occam famously argued the nominalist position against the existence of aggregates in the context of the pope wanting to give back property to the Franciscan Order: Occam argued (presaging a famous remark of Margaret Thatcher[13]) "that it was impossible to return these possessions to the order as a whole, since the Franciscan Order was only a name designating individual Franciscans" (Desrosières, 1998, p71).

The reification of groups to which people belong is central to the notion of prejudice. Gordon Allport, in his famous book on the subject, said

> *Overcategorization* is perhaps the commonest trick of the human mind. Given a thimbleful of facts we rush to make generalizations as large as a tub (Allport, 1954, p8).

Of course some categorization is essential for normal functioning:

> The human mind must think with the aid of categories (the term is equivalent here to generalization). Once formed, categories are the basis for normal prejudgement. We cannot possibly avoid this process. Orderly living depends upon it (Allport, 1954, p20).

The relevant question is thus not so much whether groups "exist" and in what sense they do, but rather *which groups are singled out for "protection" and why*? We turn to this question now.

H.3.2. The Choice and Construction of Groups — Entitativity and Similarity

Humans form themselves into groups all the time. Academic researchers do this when they proudly quote the "impact factors" of journals in which they have published[14]. But how are these groups chosen? One can create bizarre groups that nobody could argue should be protected (people whose names have a number of letters that is prime, for example). But there are other groups (e.g. membership of a racial category) which are widely accepted as being legitimate and worthy of "protection." How can one distinguish between these two different types of groups? One obvious source is anthropology, "the *science of groups of men* and their behaviour and productions" (Kroeber, 1948, p1) which has made substantial study of the social groupings underpinning kinship systems (Lévi-Strauss, 1969; Stone, 2010; Dziebel, 2007; Radcliffe-Brown, 1952).

13. Thatcher famously said "There is no such thing as society. There is living tapestry of men and women and people" Keay (1987). Interestingly, rather than the common (mis)-interpretation (that she was dismissing all societal considerations), her actual point was methodological. This is clearly apparent in the context of the interview, and via her subsequently clarifactory statement

    > Society as such does not exist except as a concept. Society is made up of people. It is people who have duties and beliefs and resolve. It is people who get things done (Thatcher, 1988).

    She was, in effect, arguing for what is now called "methodological individualism," the formal statement and naming of which is due to Popper (1945, p91), and which, to continue the otsogian spirit of this footnote (Merton, 1965), was succinctly espoused by Karl Marx (the very subject of Popper's book) in 1844: "What is to be avoided above all is the re-establishing of 'Society' as an abstraction vis-à-vis the individual" (Marx, 1959) (quoted in (Sen, 2009)).

14. They do so in the hope of deriving individual benefit by virtue of belonging to the group. This example is particularly apposite because the grouping made rational sense for its original purpose (for librarians to choose which journals to subscribe to on the basis of the *average* impact of all the papers in the journal; the librarians do not care about the individual papers, just the aggregate — the journal). The use by individual researchers makes far less sense (Seglen, 1997): why should the average number of times *other people's* papers are cited imply anything about the significance of *your* paper that happens to be published in the same venue? The only way it makes sense is as an example of the freerider problem Hardin (2013)!

Even a grouping such as ethnicity or race (Wade, 2002; Barth, 1969), widely accepted as being worthy of protection, rests on rather uncertain foundations. Alfred Kroeber, whose famous book *Anthropology* quoted above lists "race" as the first word in its sub-title, attempts to clarify race in terms of the "traits on which classification rests" (Kroeber, 1948, p126). These traits include hair texture, skin color, nose shape, cranial shape, stature, mongoloid eyes etc. Kroeber complains of the difficulty of constructing a "trustworthy classification of the human races" and observes

> A race is only a sort of average of a large number of individuals; and averages differ from one another much less than individuals. Popular impression exaggerates the differences, accurate measurements reduce them (Kroeber, 1948, p126).

He claims that his trait maps (Kroeber, 1948, p145) are more accurate and reliable than (self-declared) race. Thus, ironically, the "protected attribute" is suggested to be best defined in terms of a feature vector! He also presents (p.148) a range of other racial categorizations, with the explicit goal of pointing out the impossibility of there being one single "true" racial categorization. Kroeber's difficulty arises becase there is no notion of similarity that cleanly draws the boundaries in question. This is not unique to race: "nothing out there in nature answers to the name 'overall similarity'" (Hull, 1996).

The point is not whether a given grouping can be given an objective and unassailable definition. (And even when it cannot, it is common that people believe it can through an essentialist fallacy (Haslam and Rothschild, 2000).) Rather, the question is whether enough people believe in the group, as occurs with people self-identifying with a racial group, even if this contradicts measurable features and exaggerates differences with non-members of the group. If a grouping is widely believed to exist, then, socially, it can be said to exist:

> Social categories differ from classifications of the natural world in an undeniable aspect. Classifications of the natural world are one-way relationships in the sense that only people categorise plants: plants are in no position to protest. People, however, have their own ideas about group membership — not only ideas, but strong sentiments (Starr, 1992, p158).

Starr (1992, p160) goes on to observe that official classification is a political act and the "the principle of classification varies with our purposes and intents." (Goodman, 1992, p16) The very question of whether one wishes to be considered to be in a given group is prone to self-serving biases in the same way that one's view of the value of affirmative action programs depends upon whether you are in the group on whose behalf the action is taken (Graves and Powell, 1994).

This phenomenon of the social construction of groups seems quite significant for the problem of fair machine learning. Fortunately there is a substantial literature on the topic (Turner et al., 1994; McGarty et al., 1995; McGarty, 1999; McGarty et al., 2002). The property of perceiving a social group as an entity in its own right was named "entitativity" (the property of being an entity) by Campbell (1958), who recognised that a significant factor affecting the entitativity of a group is the degree of "common fate" that its members share. Entitativity is central to notions of prejudice (Brown, 2010; Fishbein, 2002; Dovidio et al., 2005).

As well as assimilating the substantial knowledge regarding social categorization in the works cited above, there remains the significant challenge, which to our knowledge is not yet addressed in the machine learning literature, of handling partial or graded membership of social categories, and to manage the complexity of simultaneous membership of multiple categories, which is the most common situation in practice (Sen, 2009, p145), (Sen, 2006, p22).

H.3.3. THE INDIVIDUAL AND THE GROUP IN STATISTICS — FROM THE ECOLOGICAL FALLACY TO ADVERSE SELECTION

In statistical affairs . . . the first care before all else is to lose sight of the man taken in isolation in order to consider him only as a fraction of the species. It is necessary to strip him of his individuality to arrive at the elimination of all accidental effects that individuality can introduce into the question.

— Siméon Denis Poisson, 1835[15]

Whenever an efficiency-promoting but emotionally or morally suspect variable is rendered inadmissible, a redistribution of risk occurs. (Abraham, 1985, p441).

Classical statistics traditionally deals with individuals, not groups, but in doing so endeavours to erase notions of individuality as the above quotation illustrates. When the individuals (about whom the statistical inference is being performed) can be aggregated into groups, there arises the danger of the "ecological fallacy" (where one attempts to infer something statistical about an individual on the basis of only pre-aggregated data) (Robinson, 1950; Freedman, 1999; Kramer, 1983); a particular case of this is known as Simpson's paradox (Feld and Grofman, 2010, p.140, footnote 3). The problem arises when a conditional statistic (per group) is used to infer something about an individual:

> We believe that the investigator is never justified in interpreting the results of ecological analyses in terms of the individuals who give rise to the data. This may seem to many readers to be an overstatement; however, our theoretical and empirical analyses offer no consistent guidelines for the interpretation of ecological correlations or regressions when data on individuals are unavailable (Piantadosi et al., 1988).

As Barocas and Selbst (2016, footnote 68 and associated text) explain, even when the intent is that people are all treated as individuals, as ethicists argue is right (Lippert-Rasmussen, 2011), resource constraints will sometimes require treatment in terms of group membership. Obviously many of the methods proposed for fair machine learning sail very close to this problem.

A domain of statistical thinking where group membership questions loom large is insurance. In this case, the potential downside is often not to the individual (the insured) but to the the organisation (the insurer). The problem is known as "adverse selection" whereby members of high risk groups preferentialy buy insurance and those in low risk groups do not, thus increasing the aggregate risk for the insurer (Rees and Wambach, 2008).

Insurance is an asymmetric information market (Rees and Wambach, 2008; Dionne and Rothschild, 2014) (not unlike the veil of ignorance, which has the information asymmetry, but without necessarily a market structure). The individuals seeking insurance know more about their risk than the insurer, who in the extreme case knows only very broad population averages. The individual can condition on many variables that remain hidden to the insurer (Doherty and Thistle, 1996). It is thus (from the perspective of an insurer) a "market for lemons" Akerlof (1970). (To be sure, there are informational advantages on the insurer's side sometimes too (Keogh and Otlowski, 2013).)

The relevance to fairness in machine learning becomes clear when one considers the issue of genetic testing in insurance, an issue first raised by Fisher in 1933 (Harper, 1993). Those who study insurance have concerns about "discrimination" and "unfairness" Billings et al. (1992) without making it clear what they mean by "fair." Some authors (Hudson et al., 1995) have argued extreme positions against *any* conditioning on genetic tests.[16].

The economic impacts of banning such forms of risk classification (as it is so called) were studied by Dionne and Rothschild (2014) for classifying risk in terms of gender and race. (Gender based risk classification has been prohibited in the EU since late 2012, notwithstanding the significant differences in mortality on average.) Health insurance typically conditions on "age, sex, smoking behaviour, parent's health history information, current medical conditions,. . . information about lifestyle, diet, and exercise." Risk classification is usually

15. Quoted in (Hacking, 1990, p81).
16. Hudson et al. (1995) say that

> 1) Insurance providers should be prohibited from using genetic information, or an individual's request for genetic services, to deny or limit any coverage or establish eligibility, continuation, enrollment, or contribution requirements. 2) Insurance providers should be prohibited from establishing differential rates or premium payments based on genetic information or an individual's request for genetic services.

associated with increased efficiency (as one would expect). They conclude that prohibition of the use of predictive features (such as those listed above) is likely to cause significant problems.[17] The issue of the cost of such risk classification bans (note the similarity to the cost of fairness considered in the present paper) has been considered from the perspective of social welfare function (average utilities), but not fairness directly (Hao et al., 2016), although (Hoy, 2006) adopts "an explicit welfare function approach of the sort inspired by Harsanyi's veil of ignorance."

The information-throttling nature of the problem is made clear by current Australian law (Centre for Genetics Education, 2016), which draws a distinction between a subject having had a genetic test done, but not knowing the result (the law then says that the result cannot be taken into account in pricing the insurance) versus having had a test done and the subject being told the result, in which case they are obliged to share the result with the insurer, who may price the insurance differentially as a consequence. What matters is what information is known to whom at what time.

This perspective, and the close connection to the "cost of fairness" is made clear by Tabarrok (1994) who asks "What rules governing genetic policy would rational agents agree to if placed behind a veil of ignorance". He quotes Maddox (1991):

> The contention that people unlucky enough to carry identifiable genetic abnormalities should not be denied insurance on the same terms as other people begs the question why people relatively free from identifiable genetic abnormality should therefore pay more than would otherwise be necessary.

The point is that some groupings are widely considered to be socially acceptable: Maddox (1991) goes on to say "In any case, the principle of discrimination has already been considered both in respect of those who smoke cigarettes and, for much longer, those with a familiar history of heart disease."

The issue of fairness (in the sense of the present paper) and risk classification in insurance is most systematically studied by Abraham (1985, 1986), who also recognises the trade-off inherent in excluding certain protected attributes (see the quotation at the beginning of this subsection). Akin to the notion or "moral luck" mentioned above, Abraham (1985, p424) considered the incentive effects of properly priced insurance (in encouraging the reduction of risk). But this only works if the risk factors are under the individual's control: "if classification sends messages about features outside the insured's control — age or sex, for example — the messages cannot be converted to action." This is particularly pertinent for risk factors one may be born with (Keogh and Otlowski, 2013).

H.4. Naturalising Fairness — Evolution, Ethology, Psychology, Neuroscience, & Immunology

Fairness can be studied as an empirical science, examining the way humans actually deal with notions of fairness, as well as trying to understand how such behaviours evolved. There is a range of work at many different levels that attempts to "naturalize fairness."

Trivers (1971) has famously demonstrated how simple notions of fairness (reciprocal altruism) can naturally evolve. There is a vast literature building upon this work. Binmore (2005, 2008) has focussed on naturalizing the proposals of Harsanyi and Rawls. Naturalists have argued that the broader foundation of ethics can be develop as an applied science (Ruse and Wilson, 1986) following the tenets of evolutionary naturalism (Ruse, 1995, 2009) and sociobiology (Alcock, 2001; Singer, 1981). The contractarian theories have also been examined from an evolutionary perspective (Skyrms, 1996). Such attempts do not so much as ignore the so called naturalistic fallacy (that "is" cannot imply "ought"), as deliberately repudiate it (Walter, 2006).

17. Dionne and Rothschild (2014) observe that

> Risk-pooling arising from legal restrictions on risk classification may lead to a situation in which lower-risk individuals are charged higher than actuarially fair premiums and higher-risk individuals are charged lower than actuarially fair premiums. While these financial inequities (may) reduce classification risk (and/or improve social equity), the higher-than-fair premiums for lower-risk individuals may cause them to forgo insurance entirely, particularly when the proportion of high-risk individuals is large. This reduced pool of insured individuals reflects a decrease in the efficiency of the insurance market. These negative efficiency consequences of limits on risk classification can, in principle, be quite severe: exit of low risks can lead to a "death spiral" of rising premiums and lower-risk exit that ends up unravelling and destroying the entire market.

Similar conclusions are reached by Porrini (2015).

The psychology of ethical behaviour has been studied in both humans (Rottschaefer, 2000) and other primates. de Waal et al. (2006) have argued from primate behaviour studies how human morality has evolved, and capuchin monkeys have been shown to have a strong sense of inequity aversion (Brosnan and de Waal, 2003; Brosnan, 2014). Scholars such as Haidt (2008) have turned morality into an empirical psychological discipline (the "new synthesis" (Haidt, 2007; Graham et al., 2011)), which provides a five-fold taxonomy for the basis of morality: harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect and purity/sanctity. People of differing political persuasions (liberal vs conservative) systematically differ in the relative weight they give to these 5 dimensions when considering whether something is right or wrong (Haidt, 2007), and women pay more attention to the fairness dimension than men (Graham et al., 2011, p.379). Haidt argues that "individuals are the fundamental units of moral value" (not groups)[18].

Even more mechanistically, there is now work on the neuroscience underpinning fair behaviour in humans (Pfaff, 2007). Like the philosophical theories of Harsanyi and Rawls, it seems our brains engender fairness through information throttling[19], with a clear and significant role played by the neuropeptide oxytocin (De Dreu, 2012; Sheng et al., 2013; Shalvi and De Dreu, 2014) which underpins the development of trust and in-group identification (Kosfeld et al., 2005).

We need not stop there. Our immune system's primary role is the information theoretic pattern recognition of self from non-self, the self being the ultimate in-group (Anderson and Mackay, 2014), with all individuals being immunologically unique (Medawar, 1957)[20]. This was the fundamental insight of (Burnet, 1956) who early on noted that "Recognition of 'self' from 'not-self' . . . is probably the basis of immunology" (Burnet, 1948).

We conjecture that much can be gleaned from this literature (briefly skteched above) to guide the further development of fairness in data analytics.

H.5. Challenges

God loves from Whole to Parts: But human soul
Must rise from Individual to the Whole.
Self-love but serves the virtuous mind to wake,
As the small pebble stirs the peaceful lake;
The centre moved, a circle straight succeeds,
Another still, and still another spreads;
Friend, parent, neighbour, first it will embrace;
His country next; and next all human race;
— Alexander Pope, *Essay on Man* (IV, 361-372)

We conclude this appendix with a list of some interesting challenges for fair machine learning that is 1) based on solid evidence but 2) not significantly treated in the machine learning community. Analogous to the sentiment expressed by Pope above, this requires us to "expand the circle[21]" of our moral concern.

Machine learning researchers know that any prediction of future performance is only ever probabilistic. But this in itself can be a problem. Rebecca Zwick, in studying the various schemes that have been devised for

18. "We might call it an *individualist approach to morality because individuals are the fundamental units of moral value*. In this approach, selfishness is suppressed by encouraging individuals to empathize with and care for the needy and vulnerable (Gilligan) and to respect the rights of others and fight for justice (Kohlberg). Authority and tradition have no value in and of themselves; they should be questioned and altered anew in each generation to suit society's changing needs. *Groups also have no special value in and of themselves. People are free to form voluntary cooperatives, but we must always be vigilant against the ancient tribal instincts that lead to group-based discrimiation.* (Haidt, 2008, p70) (italics added). See also (Gibbs, 2014; Sinnott-Armstrong, 2008a,b).
19. Pfaff (2007, p) says (italics added):

    > What circuitry in the brain could possibly produce behavior as universal, as exible, and yet as reliable as 'fair play?' Believe it or not, according to my theory, this circuitry does not require fancy tricks of learning and memory but instead makes use of the easiest brain process of all —– *the blurring and forgetting of information*.

    His point is that by blurring the distinction between self and non-self, one's self-interest transfers to interest in the welfare of others.
20. "So far from being one of his higher or nobler qualities, his individuality shows man nearer kin to mice and goldfish than to the angels; it is not his individuality but only his awareness of it that sets man apart" (Medawar, 1957, p185).
21. Observe that Pope anticipated the tag line of Singer (1981)!

college admissions (selecting who gets to attend) noted that psychologists "found this unpredictability hard to accept and rejected statistical prediction because it made the reality of poor prediction explicit" (Zwick, 2017, p85). Any acceptable data analytic notion of fairness will have to grapple with this point (related to the *ex ante* vs *ex post* distinction made earlier).

H.5.1. The Reality of Different Base Rates Across Protected Categories

Anecodotally, one of the most significant challenges arises, ironically enough, *from the underlying data itself*, which gets exposed to full view only when one grapples with the fairness problem in the manner of the present paper. Specifically the issue is that of "different base rates" between members and non-members of protected categores. That is, where the desired attribute being classified by the ML method is highly correlated with the chosen protected attribute. One just has to think of the heat generated in discussions about the correlation of race with recidivism[22] or about sex differences in cognitive abilities (exhaustively documented by Halpern (2012)) or sex differences in susceptibility to mental illnesses (Hartung and Widiger, 1998). Thus if one were to find that a predictor for some cognitive task resulted in a sex-ratio different to the population, emotions could run high. Although the evidence for these differences (which are *not* unidirectional) is overwhelmingly solid, merely drawing attention to them can be fraught:

> But among many professional women the existence of sex differences is still a source of discomfort. As one colleague said to me, "Look, I know that males and females are not identical. I see it in my kids, I see it in myself, I know about the research. I can't explain it, but when I read claims about sex differences, steam comes out of my ears" (Pinker, 2003, p352).

The challenge is that perhaps in response to such reactions, a practice has developed that attempts to deny what the data says:

> In contrast to such anthropological and sociological bases for expecting real group differences to provide in some sense 'accurate' components for stereotypes, there has grown up in social and educational psychology a literature and teaching practice which says that all stereotypes of group differences are false, and, implicitly, that all groups are on the average identical. (Campbell, 1967, p823).

We offer no simple solution, but note that until we move away from the practice described by Campbell, progress is impossible, and that the oft proposed device of inter-generational remediation (assuaging an injustice to an ancestor by providing recompense to a descendent) could well create more problems than it solves, especially noting the imponderable and opaque causative mechanisms involved (Sen, 2009).

H.5.2. The Partial Truth of Stereotypes and Prejudgements

Stereotypes are sometimes partially true. This too seems hard for some people to accept, perhaps because they have assimilated Campbell's point above that all sterotypes are wrong and harmful. But generalisations such as accurate stereotypes are not harmfully prejudicial. Indeed, Allport, in his foundational work on prejudice, actually defined prejudice as a *faulty* generalisation.

> Prejudgements become prejudices only if they are not reversible when exposed to new knowledge. . . . Ethnic prejudice is an antipathy based on a faulty and inflexible generalization. It may be felt or expressed. It may be directed toward a group as a whole, or toward an individual because he is a member of that group. (Allport, 1954, p9)

His point is refined in a more recent book on stereotype accuracy:

22. Berk et al. (2017, Section 8) complain that

> Perhaps the most challenging problem in practice for criminal justice risk assessments is that different base rates are endemic across protected group categories.

> Any probabilistic cue is potentially inaccurate. Using an accurate stereotype is not unfair, however, unless it is used when more diagnostic cues are ignored. (Lee et al., 1995, p308)

But such a viewpoint is far from universal. In fact there seems to be a widespread instinctive aversion to *any* probabilistic prediction of behaviour based on statistical behaviour (Schauer, 2003), even though the more individualised one makes the basis of prediction, the less one is likely to fall prey to unfair stereotypes.

H.5.3. Conflict with Folk Notions of Discrimination and Fairness

There are two "folk" notions that conflict with the more principled approaches that are the focus of the present paper. One is the the perceived difference between *identified* and *statistical* lives (Cohen et al., 2015; Russell, 2014). An *identified* life is a particular named person (think of newspaper human interest stories) versus a *statistical* life which is merely a statistic. One challenge that needs to be faced in the (statistical) approach to fairness is that it is impersonal in the sense of not identifying people. Given that processes are often judged *ex post*, the individuals concerned will often become identified and thus this bias could come into play.

Not unrelated to the previous point is the folk notion of discrimination. Using the Wikipedia article on "discrimination" as a proxy for this folk idea makes this clear:

> In human social affairs, discrimination is treatment or consideration of, or making a distinction in favor of or against, a person based on the group, class, or category to which the person is perceived to belong rather than on individual attributes (Wikipedia, 2017).

By that definition, the only non-discriminatory approach is to ignore the protected attributes *entirely* and take the outcome as it comes. That is, to make no attempt to ensure the outcome is "fair." In particular, the Bayes optimal schemes we derived would be inadmissible precisely because they *do* take consideration of a category to which a person is perceived to belong (the protected attribute). The difficulty is not so much with the proposed practice, but with clumsy definitions such as that above which are obviously unworkable. Ironically (since the whole point of a classifier is to "discriminate"), any ML method for fair data analytics will need to have a defence against such a charge of discrimination.

H.5.4. Ultimate Sources of Conflict — Different Frames of Sociality and Morality

Trading of utility for fairness is a fraught endeavour: witness the vehement response (Dorff, 2002) to the proposal of Kaplow and Shavell (2002). This particular debate was scholarly and rational[23], but alas that is not something one can rely upon.

We do not expect that even the most careful mathematical analysis will resolve the problem of fairness in data analytics for everyone. Beyond the obvious fact that different loss functions and utilities lead to different solutions (hardly a new idea in data analytics), there are more deep-seated disagreements, such as the varying opinions regarding whether groups have rights themselves[24], or that prior injustice to an ancestor justifies remediation to a descendant. As Fiske (1992) has shown, there are four quite different types of human sociality: community sharing, authority ranking, equality matching, and market pricing[25], and these underpin different moral motives (Rai and Fiske, 2011). Typically, notions of fairness are construed within the frame of equality matching. But this runs into problems, especially when considering probabilistic notions of fairness as is necessary for fairness in ML where some notion of utility (cost) is needed to even define what a satisfactory solution is — to talk of cost immediately pushes one into the market pricing type.

23. One concern, pertinent for the present paper, was the degree to which Pareto optimality is violated; see (Chang, 2000, p196). Our proposal does not do this: as one trades off fairness to utility, one does not make *everyone* worse off (as Kaplow and Shavell (2002) argue); instead only those who's predicted outcome was very uncertain would have their outcome changed for fairness concerns, as is evident by the form of the Bayes optimal solution we derived.
24. This is itself a complex issue which we can hardly do justice to here. The theories of Rawls and Harsanyi are entirely focussed upon the rights of individuals and we believe this is the appropriate stance to take. For an (in our opinion unconvincing) contrary view, namely that groups themselves have rights, see (Van Dyke, 1975; Fiss, 1976).
25. These have a correlation with, but are not identical to, the new synthesis typology mentioned earlier (harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect and purity/sanctity) (Graham et al., 2011).

Fiske (1991, p372) argued that "since the motives, the 'utilities' of each type are distinct and incommensurable, there is no transcendant value that encompasses all four types of social motives." Thus there is no single foundation on which to rest. We speculate that the present paper, which conjoined notions from equality matching (fairness) and market pricing (cost), uncovers this difficulty. Perhaps the revulsion that sometimes occurs when these moral domains come into conflict could explain the negative reaction that some people exhibit to the very point of this paper — evaluating the *cost* of fairness.

Additional References Cited in Appendices

Kenneth S. Abraham. Efficiency and fairness in insurance risk classification. *Virginia Law Review*, 71(3):403–451, April 1985.

Kenneth S. Abraham. *Distributing Risk: Insurance, Legal Theory, and Public Policy*. Yale University Press, 1986.

Matthew D. Adler. Extended preferences and interpersonal comparisons: A new account. *Economics and Philosophy*, 30:123–162, 2014.

George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics*, pages 488–500, 1970.

John Alcock. *The Triumph of Sociobiology*. Oxford University Press, 2001.

Gordon W. Allport. *The Nature of Prejudice*. Addison-Wesley, 1954.

Warwick Anderson and Ian R. Mackay. Fashioning the immunological self: The biological individuality of F. Macfarlane Burnet. *Journal of the History of Biology*, 47:147–175, 2014.

James Andreoni, Deniz Aydin, Blake Barton, B. Douglas Bernheim, and Jeffrey Naecker. When fair isn't fair: Sophisticated time inconsistency in social preferences. Technical report, UCSD and Stanford University, March 2016.

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

Frederik Barth, editor. *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*. Little, Brown and Company, 1969.

Richard Berk, Hoda Heidari, Shain Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessment: The state of the art. Technical report, ArXiV:1703.09207v1, March 2017.

Paul R. Billings, Mel A. Kohn, Margaret de Cuevas, Jonathan Beckwith, Joseph S. Alper, and Marvin R. Natowicz. Discrimination as a consequence of genetic testing. *American Journal of Human Genetics*, 50:476–482, 1992.

Ken Binmore. *Game Theory and the Social Contract Volume 1: Playing Fair*. MIT Press, 1994.

Ken Binmore. *Natural Justice*. Oxford University Press, 2005.

Ken Binmore. Naturalizing Harsanyi and Rawls. In Marc Fleurbaey, Maurice Salles, and John A. Weymark, editors, *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, chapter 13, pages 303–333. Cambridge University Press, 2008.

William Blackstone. *Commentaries on the Laws of England*. Clarendon Press, Oxford, 1765.

Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences*. MIT press, 1999.

Sarah F. Brosnan. Justice- and fairness-related behaviours in nonhuman primates. In Camila J. Cela-Conde, Raúl Gutiérrez Lombardo, John C. Avise, and Francisco J. Ayala, editors, *In the Light of Evolution: Volume VII: The Human Mental Machinery*, pages 191–210. The National Academies Press, 2014.

Sarah F. Brosnan and Frans B.M. de Waal. Monkeys reject unequal pay. *Nature*, 425:297–299, 2003.

Rupert Brown. *Prejudice: Its Social Psychology*. Wiley-Blackwell, 2nd edition, 2010.

F. Macfarlane Burnet. The basis of allergic diseases. *Medical Journal of Australia*, 29-35, 1948.

F. Macfarlane Burnet. *Enzyme, Antigen and Virus: A Study of Macromolecular Pattern in Action*. Cambridge University Press, 1956.

Donald T. Campbell. Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Systems Research and Behavioral Science*, 3(1):14–25, 1958.

Donald T. Campbell. Stereotypes and the perception of group differences. *American psychologist*, 22(10):817, 1967.

Centre for Genetics Education. Life insurance products and genetic testing in australia. NSW Government, Department of Health, February 2016.

Christopher P. Chambers and Alan D. Miller. A measure of bizarreness. *Quarterly Journal of Political Science*, 5:27–44, 2010.

Howard F. Chang. A liberal theory of social welfare: Fairness, utility, and the Pareto principle. *The Yale Law Journal*, 110:173–235, 2000.

I. Glenn Cohen, Norman Daniels, and Nir Eyal, editors. *Identified versus Statistical Lives: An Interdisciplinary Perspective*. Oxford University Press, 2015.

Carsten K.W. De Dreu. Oxytocin modulates cooperation within and competition between groups: An integrative review and research agenda. *Hormones and Behaviour*, 61:419–428, 2012.

Frans de Waal, Robert Wright, Christine M. Korsgaard, Philip Kitcher, and Peter Singer. *Primates and Philosophers: How Morality Evolved*. Princeton University press, 2006.

Alain Desrosières. *The Politics of Large Numbers: A History of Statistical Reasoning*. Harvard University Press, 1998.

Georges Dionne and Casey Rothschild. Economic effects of risk classification bans. *The Geneva Risk and Insurance Review*, 39:184–221, 2014.

Neil A. Doherty and Paul D. Thistle. Adverse selection with endogenous information in insurance markets. *Journal of Public Economics*, 63(1):83–102, 1996.

Neil J. Dorans and Linda L. Cook, editors. *Fairness in Educational Assessment and Measurement*. Routledge, 2016.

Michael. B. Dorff. Why welfare depends on fairness: A reply to Kaplow and Shavell. *Southern California Law Review*, 75:847–899, 2002.

John F. Dovidio, Peter Glick, and Laurie A. Rudman, editors. *On the Nature of Prejudice: Fifty Years after Allport*. Blackwell, 2005.

German V. Dziebel. *The Genius of Kinship: The Phenomenon of Human Kinship and the Global Diversity of Kinship Terminologies*. Cambria Press, 2007.

Scott L. Feld and Bernhard Grofman. Puzzles and paradoxes involving averages: An intuitive approach. In *Collective Decision Making*, pages 137–150. Springer, 2010.

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

David Hackett Fischer. *Fairness and Freedom: A History of Two Open Societies: New Zealand and the United States*. Oxford University Press, 2012.

Harold D. Fishbein. *Peer Prejudice and Discrimination: The Origins of Prejudice*. Lawrence Erlbaum Associates, 2nd edition, 2002.

Alan Page Fiske. *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing*. Free Press, 1991.

Alan Page Fiske. The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological review*, 99(4):689, 1992.

Owen M. Fiss. Groups and the equal protection clause. *Philosophy and Public Affairs*, 5(2):107–177, 1976.

Marc Fleurbaey, Maurice Salles, and John A. Weymark. *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*. Cambridge University Press, 2008.

Mark Fleurbaey and François Maniquet. *A Theory of Fairness and Social Welfare*. Cambridge University Press, 2011.

David Freedman. Ecological inference and the ecological fallacy. Technical Report 549, Department of Statistics, University of California, Berkeley, October 1999.

John C. Gibbs. *Moral Development and Reality: Beyond the Theories of Kohlberg, Hoffman, and Haidt*. Oxford University Press, 2014.

Nelson Goodman. Seven strictures on similarity. In Mary Douglas and David Hull, editors, *How Classification Works: Nelson Goodman among the Social Sciences*, pages 13–23. Edinburgh University Press, 1992.

Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. Mapping the moral domain. *Journal of Personal Psychology*, 101(2):366–385, 2011.

Laura M. Graves and Gary N. Powell. Effects of sex-based preferential selection and discrimination on job attitudes. *Human Relations*, 47:133–157, 1994.

Ian Hacking. *The Taming of Chance*. Cambridge University Press, 1990.

Jonathan Haidt. The new synthesis in moral psychology. *Science*, 316:998–1002, 18 May 2007.

Jonathan Haidt. Morality. *Perspectives on Psychological Science*, 3(1):65–72, 2008.

Diane F. Halpern. *Sex Differences in Cognitive Abilities*. Psychology Press, 4th edition, 2012.

Peter J. Hammond. Interpersonal comparisons of utility: Why and how they are and should be made. In Jon Elster and John E. Roemer, editors, *Interpersonal Comparisons of Well-Being*, chapter 7. Cambridge University Press, 1991.

MingJie Hao, Angus S. Macdonald, Pradip Tapadar, and R. Guy Thomas. Insurance loss coverage and social welfare. Technical report, University of Kent, 2016.

Russell Hardin. The free rider problem. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2013 edition, 2013.

Peter S. Harper. Insurance and genetic testing. *The Lancet*, 341:224–227, January 1993.

John C. Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *The Journal of Political Economy*, 63(4):309–321, 1955.

John C. Harsanyi. Ethics in terms of hypothetical imperatives. *Mind*, 67(267):305–316, July 1958.

John C. Harsanyi. Rule utilitarianism and decision theory. *Erkenntnis*, 11:25–53, 1977a.

John C. Harsanyi. Morality and the theory of rational behaviour. *Social Research*, 44(4):623–656, 1977b.

John C. Harsanyi. *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, 1977c.

John C. Harsanyi. Bayesian decision theory and utilitarian ethics. *The American Economic Review*, 68(2):223–228, 1978.

John C. Harsanyi. Bayesian decision theory, rule utilitarianism, and Arrow's impossibility theorem. *Theory and Decision*, 11:289–317, 1979.

John C. Harsanyi. John Rawl's Theory of Justice: Some critical comments. In Marc Fleurbaey, Maurice Salles, and John A. Weymark, editors, *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, pages 71–79. Cambridge University Press, 2008.

Cynthia M. Hartung and Thomas A. Widiger. Gender differences in the diagnosis of mental disorders: Conclusions and controversies of the DSM-IV. *Psychological Bulletin*, 123(3):260–278, 1998.

Nick Haslam and Louis Rothschild. Essentialist beliefs about social categories. *British Journal of Social Psychology*, 39: 113–127, 2000.

Michael Hoy. Risk classification and social welfare. *The Geneva Papers*, 31:245–269, 2006.

Kathy L. Hudson, Karen H. Rothenberg, Lori B. Andrews, Mary Jo Ellis Kahn, and Francis S. Collins. Genetic discrimination and health insurance: An urgent need for reform. *Science*, 229(5235):391–393, October 1995.

David L. Hull. Rainbows in retrospect: L.A.S. Johnson's contributions to taxonomic philosophy. *Telopea*, 6(4):527–539, 1996.

Kory D. Johnson, Dean P. Foster, and Robert A. Stine. Impartial predictive modeling: Ensuring fairness in arbitrary models. Preprint, University of Vienna, October 2016.

Louis Kaplow and Steven Shavell. *Fairness versus Welfare*. Harvard University Press, 2002.

Douglas Keay. Interview of Margaret Thatcher. *Woman's Own Journal*, pages 8–10, 31 October 1987.

Louise A. Keogh and Margaret F.A. Otlowski. Life insurance and genetic test results: a mutation carrier's fight to achieve full cover. *Medical Journal of Australia*, 199(5):363–366, September 2013.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016.

Serge-Christophe Kolm. *Modern Theories of Justice*. MIT Press, 1996.

Michael Kosfeld, Markus Heinrichs, Paul J. Zak, Urs Fischbacher, and Ernst Fehr. Oxytocin increases trust in humans. *Nature*, 435:673–676, 2 June 2005.

Gerald H. Kramer. The ecological fallacy revisited: Aggregate- versus individual-level findings on economics and elections, and sociotropic voting. *The American Political Science Review*, 77(1):92–111, 1983.

Alfred L. Kroeber. *Anthropology: Race, Language, Culture, Psychology, Prehistory*. Harcourt, Brace and Company, revised edition, 1948.

A. Faik Kurtulmus. Uncertainty behind the veil of ignorance. *Utilitas*, 24(1):41–62, 2012.

George Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. The University of Chicago Press, 1987.

Yueh-Ting Lee, Lee J. Jussim, and Clark R. McCauley, editors. *Stereotype Accuracy: Toward Appreciating Group Differences*. American Psychological Association, 1995.

Claude Lévi-Strauss. *The Elementary Structures of Kinship*. Beacon Press, 1969.

Kasper Lippert-Rasmussen. "We are all different": Statistical discrimination and the right to be treated as an individual. *The Journal of Ethics*, 15(1-2):47–59, 2011.

Mark J. Machina. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, 27(4):1622–1668, 1989.

John Maddox. The case for the human genome. *Nature*, 352(6330):11, 1991.

Karl Marx. *Economic and Philosphical Manuscripts of 1844*. Progress Publishers, Moscow, 1959.

Craig McGarty. *Categorization in social psychology*. Sage, 1999.

Craig McGarty, S. Alexander Haslam, Karen J. Hutchinson, and Diana M. Grace. Determinants of perceived consistency: The relationship between group entitativity and the meaningfulness of categories. *British Journal of Social Psychology*, 34:237–256, 1995.

Craig McGarty, Vincent Y. Yzerbyt, and Russell Spears, editors. *Stereotypes as Explanations: The Formation of Meaningful Beliefs about Social Groups*. Cambridge University Press, 2002.

Peter D. Medawar. *The Uniqueness of the Individual*. Basic Books, 1957.

Robert K. Merton. *On the Shoulders of Giants: The Post-Italianate Edition*. University of Chicago Press, 1965.

Thomas Nagel. Moral luck. In *Mortal Questions*. Cambridge University Press, 1979.

Donald W. Pfaff. *The Neuroscience of Fair Play: Why we (Usually) Follow the Golden Rule*. Dana Press, 2007.

Steven Piantadosi, David P. Byar, and Sylvan B. Green. The ecological fallacy. *American Journal of Epidemiology*, 127 (5):893–904, 1988.

Steven Pinker. *The Blank Slate: The Modern Denial of Human Nature*. Penguin, 2003.

Karl R. Popper. *The Open Society and its Enemies, Volume II: The High Tide of Prophecy: Hegel, Marx and the Aftermath*. Rutledge and Kegan Paul, 1945.

Donatella Porrini. Risk classification efficiency and the insurance market regulation. *Risks*, 3(4):445–454, 2015.

Alfred R. Radcliffe-Brown. *Structure and Function in Primitive Society*. The Free Press, 1952.

Tage Shakti Rai and Alan Page Fiske. Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, 118(1):57, 2011.

John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

Ray Rees and Achim Wambach. The microeconomics of insurance. *Foundations and Trends in Microeconomics*, 4(1-2): 1–163, 2008.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, December 2010.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, Mar 2011.

William S. Robinson. Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15(3): 351–357, 1950.

William A. Rottschaefer. Naturalizing ethics: The biology and psychology of moral agency. *Zygon*, 35(2):253–286, 2000.

Michael Ruse. *Evolutionary Naturalism*. Routledge, London and New York, 1995.

Michael Ruse. The biological sciences can act as a ground for ethics. In Francisco Ayala and Robert Arp, editors, *Contemporary Debates in Philosophy of Biology*. Wiley-Blackwell, 2009.

Michael Ruse and Edward O. Wilson. Moral philosophy as applied science. *Philosophy*, 61(236):173–192, April 1986.

Louise B. Russell. Do we really value identified lives more highly than statistical lives? *Medical Decision Making*, pages 556–559, July 2014.

Frederick Schauer. *Profiles, Probabilities and Stereotypes*. Harvard University Press, 2003.

Per O. Seglen. Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314:498–513, 15 February 1997.

Amartya K. Sen. Social choice theory. In K.J. Arrow and M.D. Intriligator, editors, *Handbook of Mathematical Economics*, volume 3, chapter 22, pages 1073–1181. Elsevier, 1986.

Amartya K. Sen. *Identity and Violence: The Illusion of Destiny*. W.W. Norton, 2006.

Amartya K. Sen. *The Idea of Justice*. Harvard University Press, 2009.

Shaul Shalvi and Carsten K.W. De Dreu. Oxytocin promotes group-serving dishonesty. *Proceedings of the National Academy of Sciences*, 111(15):5503–5507, 2014.

Feng Sheng, Yi Liu, Bin Zhou, Wen Zhou, and Shihui Han. Oxytocin modulates the racial bias in neural responses to others' suffering. *Biological Psychology*, 92:380–386, 2013.

Peter Singer. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press, 1981.

Walter Sinnott-Armstrong, editor. *Moral Psychology Volume 1: The Evolution of Morality: Adaptations and Innateness*. MIT Press, 2008a.

Walter Sinnott-Armstrong, editor. *Moral Psychology Volume 2: The Cognitive Science of Morality: Intuition and Diversity*. MIT Press, 2008b.

Brian Skyrms. *Evolution of the Social Contract*. Cambridge University Press, 1996.

Paul Starr. Social categories and claims in the liberal state. In Mary Douglas and David Hull, editors, *How Classification Works: Nelson Goodman among the Social Sciences*. Edinburgh University Press, 1992.

Joseph E. Stiglitz. Information and the change in the paradigm in economics. *The American Economic Review*, 92(3): 460–501, 2002.

Linda Stone. *Kinship and Gender: An Introduction*. Westview Press, 4th edition, 2010.

Alexander Tabarrok. Genetic testing: an economic and contractarian analysis. *Journal of Health Economics*, 13:75–91, 1994.

Margaret Thatcher. Atticus column. *The Sunday Times*, 10 July 1988.

Stefan T. Trautmann and Gijs van de Kuilen. Process fairness, outcome fairness, and dynamic consistency: Experimental evidence for risk and ambiguity. *Journal of Risk and Uncertainty*, 53(2-3):75–88, 2016.

Robert L. Trivers. The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1):35–57, 1971.

John C. Turner, Penelope J. Oakes, S. Alexander Haslam, and Craig McGarty. Self and collective: Cognition and social context. *Personality and Social Psychology Bulletin*, 20(5):454–463, 1994.

Vernon Van Dyke. Justice as fairness: For groups? *The American Political Science Review*, 69(2):607–614, 1975.

William Vickrey. Measuring marginal utility by reactions to risk. *Econometrica*, 13(319-333), 1945.

William Vickrey. On the prevention of gerrymandering. *Political Science Quarterly*, 76(1):105–110, 1961.

Alexandre Volokh. *n* guilty men. *University of Pennsylvania Law Review*, 146:173–216, 1997.

Peter Wade. *Race, Nature and Culture: An Anthropological Perspective*. Pluto Press, 2002.

R. Harrison Wagner. Impartiality and equity. *Theory and Decision*, 12:61–74, 1980.

Alex Walter. The anti-naturalistic fallacy: Evolutionary moral psychology and the insistence of brute facts. *Evolutionary Psychology*, 4:33–48, 2006.

Anna Wierzbicka. *English: Meaning and Culture*. Oxford University Press, 2006.

Wikipedia. Discrimination, 12 December 2017. URL https://en.wikipedia.org/wiki/Discrimination.

Bernard Williams. *Moral Luck: Philosophical Papers 1973–1980*. Cambridge University Press, 1981.

H. Peyton Young. Measuring the compactness of legislative districts. *Legislative Studies Quarterly*, 13(1):105–115, 1988.

H. Peyton Young. *Equity: In Theory and Practice*. Princeton University Press, 1994.

Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International World Wide Web Conference (WWW)*, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna Gummadi. Learning fair classifiers. *arXiv preprint arXiv:1507.05259*, 2016.

Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31 (1060–1089), 2017.

Rebecca Zwick. *Who Gets In? Strategies for Fair and Effective College Admissions*. Harvard University Press, 2017.