A few highlights from:

Please refer to the original paper for more information

A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity

now: Carnegie Mellon University →

Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

Michele Loi
University of Zürich
michele.loi@uzh.ch

Saarbrücken →

Krishna P. Gummadi
MPI-SWS
gummadi@mpi-sws.org

Andreas Krause
ETH Zürich
krausea@ethz.ch

person

(C, e)

⚠ partition assumed given

1. Economic Model

following Lefranc et al. (2009)

C: **circumstance** – not legitimate sources of discrimination
e: **"effort"** – all factors an individual can be held morally accountable for
ϕ: **policy**
π: **π^{th} quantile** of effort distribution

distributive justice

Definition 1 (Rawlsian Equality of Opportunity (R-EOP)) *A policy ϕ satisfies Rawlsian EOP if for all circumstances c, c' and all effort levels e,*

$$F^{\phi}(.|c, e) = F^{\phi}(.|c', e).$$

eg. male, eg. female, effort

Luck vs. conscious actions

Definition 2 (Luck Egalitarian Equality of Opportunity (e-EOP)) *A policy ϕ satisfies Luck Egalitarian EOP if for all $\pi \in [0, 1]$ and any two circumstances c, c':*

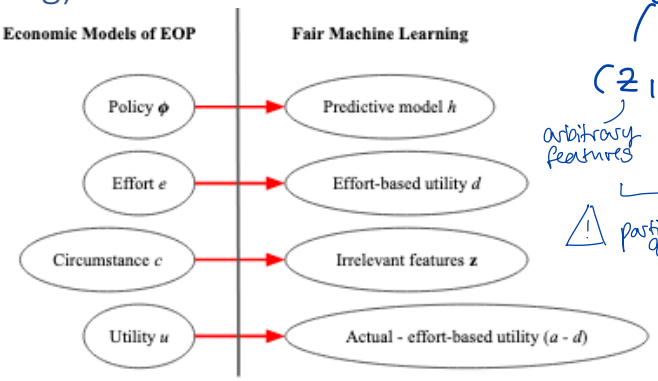
$$F^{\phi}(.|c, \pi) = F^{\phi}(.|c', \pi).$$

eg. male, eg. female, eg. to 25% of effort distribution

2. Machine Learning Model (Mapping)

$a \in [0,1]$: **actual utility**
observed
(actually received)

$d \in [0,1]$: **effort-based utility**
unobserved
(should receive based on w)



(z, w)

arbitrary features, accountability features

⚠ partition assumed given

Let u be the *advantage* or overall *utility* the individual earns as the result of being subject to predictive model h. For simplicity and unless otherwise specified, we assume u has the following simple form:

$$u = a - d. \tag{1}$$

That is, u captures the discrepancy between an individual's *actual* utility (a) and their *effort-based* utility d. With this formulation, an individual's utility is 0 when their actual and effort-based utilities coincide (i.e. $u = 0$ if $a = d$).

Definition 4 (Equality of Odds) *A predictive model h satisfies equality of odds if* $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall y, \hat{y} \in \mathcal{Y}$:

$$\mathbb{P}_{(\mathbf{X},Y)\sim F}[\hat{Y} = \hat{y} | \mathbf{Z} = \mathbf{z}, Y = y] = \mathbb{P}_{(\mathbf{X},Y)\sim F}[\hat{Y} = \hat{y} | \mathbf{Z} = \mathbf{z}', Y = y].$$

Separation
$\hat{Y} \perp Z | Y$

Definition 6 (Predictive Value Parity) *A predictive model h satisfies predictive value parity if* $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall y, \hat{y} \in \mathcal{Y}$:

$$\mathbb{P}_{(\mathbf{X},Y)\sim F}[Y = y | \mathbf{Z} = \mathbf{z}, \hat{Y} = \hat{y}] = \mathbb{P}_{(\mathbf{X},Y)\sim F}[Y = y | \mathbf{Z} = \mathbf{z}', \hat{Y} = \hat{y}].$$

Sufficiency
$Y \perp Z | \hat{Y}$

Rawls

Definition 7 (R-EOP for supervised learning) *Suppose $d = g(\mathbf{w}, y)$. Predictive model h satisfies Rawlsian EOP if for all $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$ and all $d \in [0, 1]$,*

utility distribution

$$F^h(.|\mathbf{Z} = \mathbf{z}, D = d) = F^h(.|\mathbf{Z} = \mathbf{z}', D = d).$$

assuming $U = A - D$
$D = Y$
$A = h(x) = \hat{Y}$

In the binary classification setting, if we assume the true label Y reflects an individual's effort-based utility D, Rawlsian EOP translates into equality of odds across protected groups:[3]

egalitarian

Definition 8 (e-EOP for supervised learning) *Suppose $d = f(\mathbf{x}, y, h)$. Predictive model h satisfies egalitarian EOP if for all $\pi \in [0, 1]$ and $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$,*

assuming $U = A - D$
$D = h(x) = \hat{Y}$
$A = Y$
binary

$$F^h(.|\mathbf{Z} = \mathbf{z}, \Pi = \pi) = F^h(.|\mathbf{Z} = \mathbf{z}', \Pi = \pi). \tag{2}$$

Next, we show that predictive value parity can be thought of as a special case of e-EOP, where the predicted label/risk $h(\mathbf{X})$ is assumed to reflect the individual's effort-based utility, and the true label Y reflects his/her actual utility.

assumes: same label $\hat{Y}$ people are equally accountable for their labels

	Notion of fairness	Effort-based utility D	Actual utility A	Notion of EOP
1	Equality of Odds	Y	$\hat{Y}$	Rawlsian
2	Predictive Value Parity	$\hat{Y}$	Y	egalitarian

Table 1: Interpretation of existing notions of algorithmic fairness for binary classification as special instances of EOP.

assumes: same pred. people are equally accountable for their predictions

Example 1: bank (see lecture)

Example 2: online-test (see paper)

3. Implications

On Recent Fairness Impossibility Results Several papers have recently shown that group-level notions of fairness, such as predictive value parity and equality of odds, are generally incompatible with one another and cannot hold simultaneously [Kleinberg *et al.*, 2017; Friedler *et al.*, 2016]. Our approach confers a moral meaning to these impossibility results: they can be interpreted as contradictions between fairness desiderata reflecting different and irreconcilable moral assumptions. For example predictive value parity and equality of odds make very different assumptions about the effort-based utility d: Equality of odds assumes all persons with similar true labels are equally accountable for their labels, whereas predictive value parity assumes all persons with the same predicted label/risk are equally accountable for their predictions. Note that depending on the context, usually only one (if any) of these assumptions is morally acceptable. We argue, therefore, that unless we are in the highly special case where $Y = h(\mathbf{X})$, it is often unnecessary—from a moral standpoint—to ask for both of these fairness criteria to be satisfied simultaneously.