Group Fairness Refocused: Assessing the Social Impact of ML Systems

Corinna Herweck[1,2], Michele Loi [3], Christoph Heitz[1]

[1] Zurich University of Applied Sciences, Switzerland
[2] University of Zurich, Switzerland
[2] AlgorithmWatch, Berlin, Germany

Outline and contribution

Zurich University of Applied Sciences
School of Engineering
IDP Institute of Data Analysis and Process Design

1. Impact assessment of prediction-based decision systems and fairness:
   Group fairness = «Equality of expected impact»

2. Standard fairness metrics …
- … are special cases of this general framework
- … may or may not be adequate → our generalization is a solution

3. Systematic method of deriving the required utility function for modeling the impact

What is (Algorithmic) Unfairness?

Zurich University of Applied Sciences
zh aw School of Engineering
IDP Institute of Data Analysis and Process Design

Definition:
Unfairness is «a treatment that systematically imposes a disadvantage on one social group relative to others" (Barocas, Hardt, Narayanan: Fairness and Machine Learning. 2023)

Often used synonymously with «discrimination» and «algorithmic bias»

Example:
Group A: Anna: 10€, Anton: 30€, Adriana: 50€ → Avg = 30€
Group B: Berta: 20€, Basti: 40€, Barbara: 60€ → Avg = 40€

Group fairness is not concerned with individual inequalities, but with systematic inequalities that generate an unfair disadvantage

Formalizing «systematic inequalities»

Zurich University of Applied Sciences
zh aw School of Engineering
IDP Institute of Data Analysis and Process Design

Application scenario
- System applied to (many) individuals of a society with two groups A and B
- Each decision creates benefit (or harm) to the individual

Group fairness:
Average benefit for members of A = Average benefit for members of B

Formalization: **The principle of equal expected utility**

$$E(U|a) = E(U|b)$$

with
- $E()$ = expectation value of a randomly chosen individual
- U = amount of benefit: «utility» of the decision
- a/b: indicator of group

Philosophical foundation: Group fairness is one application of *Distributive Justice*

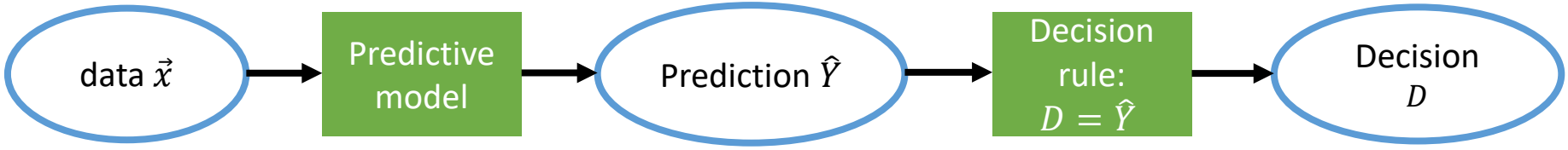
Prediction-based decision making

Simplest case: binary decision D based on prediction of binary unknown Y

Example:
Y = will be successful as team leader (yes/no)
D = promote to team leader (yes/no)

Ideal decision: If Y=0 then D=0, if Y=1 then D=1

If Y is unknown, we may predict it:



How is fairness measured in classical fairness literature?

Standard fairness criteria are based on the (group-specific) decision outcome matrix

decision	Y=0	Y=1
D=0	p_{00}	p_{01}
D=1	p_{10}	p_{11}

$$p_{00} + p_{01} + p_{10} + p_{11} = 1$$

Examples:

Statistical parity: $\boldsymbol{P[D = 1|a] = P[D = 1|b]}$

... is based on $P[D = 1] = p_{10} + p_{11}$

True positive rate parity: $\boldsymbol{P[D = 1|Y = 1, a] = P[D = 1|Y = 1, b]}$

… is based on $P[D = 1|Y = 1] = \frac{p_{11}}{p_{10}+p_{11}}$

Standard fairness criteria are counting numbers,
not measuring «impact», «benefit», «harm» of the decisions!

Utility matrix

Four decision outcomes:

decision	Y=0	Y=1
D=0	p_{00}	p_{01}
D=1	p_{10}	p_{11}

… may lead to four different utilities:

utility	Y=0	Y=1
D=0	u_{00}	u_{01}
D=1	u_{10}	u_{11}

Example: university admission based on prediction of academic performance Y

Y=1: will be a good student
D=1: is admitted

utility	Y=0	Y=1
D=0	0	0
D=1	1	1

Impact: Combine decision matrix with utility matrix

Zurich University of Applied Sciences
zh aw School of Engineering
IDP Institute of Data Analysis and Process Design

Dec	Y=0	Y=1
D=0	p_{00}	p_{01}
D=1	p_{10}	p_{11}

+

utility	Y=0	Y=1
D=0	u_{00}	u_{01}
D=1	u_{10}	u_{11}

$$E(U) = p_{00} \cdot u_{00} + p_{01} \cdot u_{01} + p_{10} \cdot u_{10} + p_{11} \cdot u_{11}$$

Measuring impact requires a combination of decision outcome matrix and utility matrix

Fairness metric and expected utility

Example of standard fairness metrics:
Statistical parity: $\boldsymbol{P[D=1]} = p_{10} + p_{11}$

General expected utility: $\boldsymbol{E(U)} = p_{00} \cdot u_{00} + p_{01} \cdot u_{01} + p_{10} \cdot u_{10} + p_{11} \cdot u_{11}$

We see: $P[Y=1]$ is identical to $E(U)$, if

$$u_{00} = 0, u_{01} = 0, u_{10} = 1, u_{11} = 1$$

For general standard fairness metrics, it turns out:

- Each standard metric is equivalent to $E(U)$, for a specific utility matrix
- Standard metrics implement Equality of Expected Impact,
  but under very specific assumptions on how a decision creates impact!

Utility matrix of statistical parity

Zurich University of Applied Sciences
zh aw School of Engineering
IDP Institute of Data Analysis and Process Design

Statistical parity:

$P[D = 1]$ …….. corresponds to $E(U)$, if $U =$

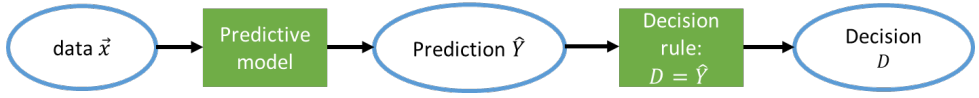
utility	Y=0	Y=1
D=0	0	0
D=1	1	1

Statistical parity is an adequate metric for assessing fairness only if this assumption is correct

Utility matrix is case dependent

Example: medical treatment with side effects

Population: people with strong chronical pain
New drug: Cures people with Y=1, but not people with Y=0. Strong side effects
Problem: No realiable test of Y: Y is unknown
«Solution»: Predict Y from available data $\vec{x}$: $\hat{Y} = f(\vec{x})$
Decision: Treat if $\hat{Y} = 1$



Dec	Y=0	Y=1
D=0	p_{00}	p_{01}
D=1	p_{10}	p_{11}

utility	Y=0	Y=1
D=0	0	0
D=1	-0.5	1

Expected impact: $E(U) = p_{10} \cdot (-0.5) + p_{11} \cdot 1$

Statistical parity: $P[Y = 1] = p_{10} + p_{11}$

Pit stop

https://thesportsrush.com/f1-news-fastest-f1-pitstop-which-team-holds-the-world-record-of-fastest-f1-pitspot/

Basic concept of group fairness = equal expected impact: $E(U|a) = E(U|b)$

Standard fairness metrics assess decisions, not impact
→ may be appropriate in some cases, but not so in others

Generalized framework: combination of decision outcome matrix and utility matrix
Standard fairness metrics are special cases

How to derive a utility function?

Zurich University of Applied Sciences
zh aw School of Engineering
IDP Institute of Data Analysis and Process Design

Simplifying assumptions on utility function:

- Utility equal for all members of population
- Utility depends only on Y and D (i.e. can be formalized as matrix), independent on group

(consistent with all classical fairness metrics)

utility	Y=0	Y=1
D=0	u_{00}	u_{01}
D=1	u_{10}	u_{11}

Usage as fairness metric: $E(U|a) = E(U|b)$ allows for

- Scaling with factor α
- Shifting by constant β

$$E(U|a) = E(U|b)$$
$$\Leftrightarrow$$
$$E(\alpha U + \beta|a) = E(\alpha U + \beta|b)$$

How to build the utility matrix

Step 1: Define a reference case
- set one utility element to 0 (shifting invariance)

Step 2: Define the «unit» of utility
- Set another element to 1 or -1 (scaling invariance)

Step 3: Define remaining two elements, by comparing the decision subjects' utility with respect to the reference values

utility	Y=0	Y=1
D=0	**0**	***
D=1	***	**1**

Extensions of the utility-based approach

General fairness principle: $E(U|a) = E(U|b)$

More complex utility definitions

- Group dependent utility: same decision creates different impact across groups
- Fully individualized utility

Different definitions of «utility»

- U = utility created by the decision → avoid additional discrimination generated by the decision system
  - Does not address already existing inequality

- U = net utility after decision → create a fairer world («algorithmic reparation»)

Conclusion

Zurich University of Applied Sciences
zh aw School of Engineering
IDP Institute of Data Analysis and Process Design

1. Framework for impact assessment of prediction-based decision systems:
   Group fairness = «Equality of expected impact»

2. Group fairness and standard fairness metrics:
- Standard fairness metrics are special cases of general approach
- Standard fairness metrics may not be adequate
- Practical cases may need other metrics, provided by our framework

3. Systematic method of deriving the required utility function

Consult our paper for more details and depth!


thanks!