The Alan Turing Institute

Individual fairness and group fairness

Dr. Ogerta Elezaj

Fairness Definitions

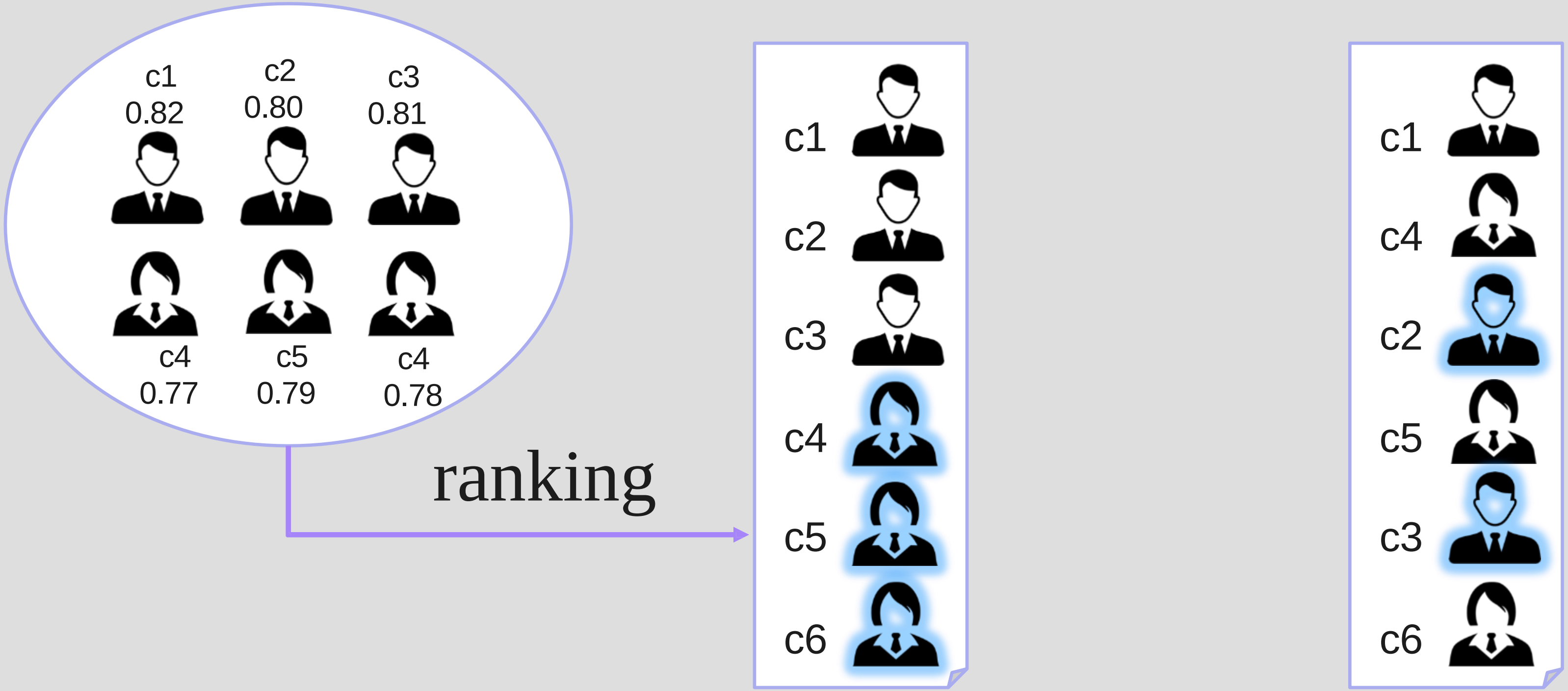
– **Individual fairness**

A requirement having the goal of treating similar individuals in a similar way. Individuals that are similar with regard to the task should be given similar decisions.

– **Group fairness**

Partitions a population into groups defined by protected attributes. Privileged groups should be treated similarly to the unprivileged groups.

Individual vs Group

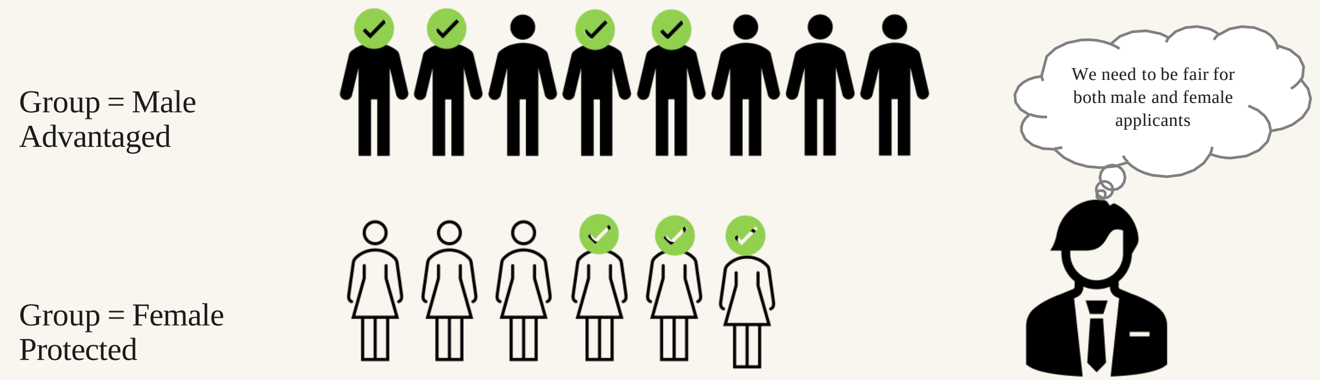


- fair for individuals
- unfair for female group

- fair for female group
- unfair for individuals c2, c3

Group Fairness vs. Individual Fairness

– Group fairness requires that the protected groups should be treated similarly to the advantaged group.



Require the same acceptance rate for both male and female job applicants

Individual fairness

Different individuals should be treated similarly.

--It imposes restriction on the treatment for each pair of individuals

Individual fairness

Treat similar individuals similarly.

Similar for the purpose of the classification task

Similar distribution over outcomes

– Binary Classification Algorithm

--Positive or negative/1 or 0/accept or reject

– Any two individuals who are similar with respect to a particular task should be classified similarly

Similarity

– How do we define similarity?

  --It is hard to determine an appropriate metric to measure the similarity of two individuals.

– We assume a distance metric

  --Similarity metric between individuals d(x,y)

    ---# of features

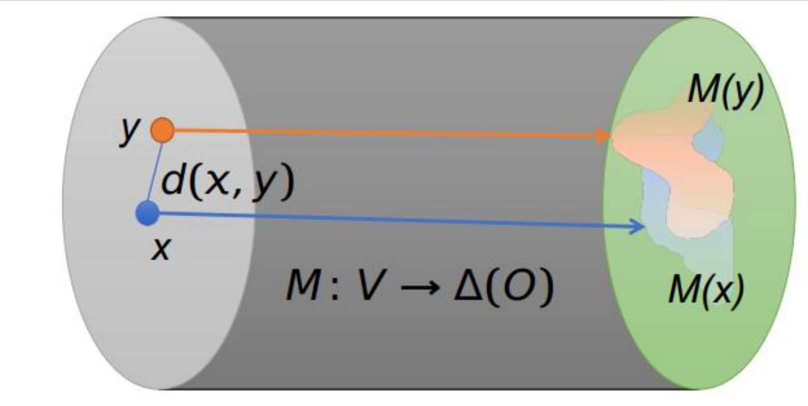
    ---Graphical distance (like word embedding)

– Similarity measurements between distributions of outcome D(x,y)

Representations (Informal)

- X: All possible real people
- Algorithm operates only on a representation of the person
  - The algorithm only knows what it is told about you
  - Distinct individuals may be mapped to the same representation

- How can we compare M(x) with M(y)?



Statistical Distance

- Numerical measure of how different two data objects are
  - A function that maps pairs of objects to real values
  - Lower when objects are more alike
  - Higher when two objects are different
- Minimum distance is 0, when comparing an object with itself.
- Upper limit varies
- A distance function d is a distance metric if it is a function from pairs of objects to real numbers such that:
  - $d(x,y) > 0$. (non-negativity)
  - $d(x,y) = 0$ iff $x = y$. (identity)
  - $d(x,y) = d(y,x)$. (symmetry)
  - $d(x,y) < d(x,z) + d(z,y)$ (triangle inequality).

Example: statistical distance

- Vectors $x = (x_1, \dots, x_d)$ and $y = (y_1, \dots, y_d)$
- **L_p** norms or Minkowski distance:

$$L_p(x, y) = \left[|x_1 - y_1|^p + \cdots + |x_d - y_d|^p\right]^{1/p}$$

- **L_2** norm: Euclidean distance:

$$L_2(x, y) = \sqrt{|x_1 - y_1|^2 + \cdots + |x_d - y_d|^2}$$

- **L_1** norm: Manhattan distance:

$$L_1(x, y) = |x_1 - y_1| + \cdots + |x_d - y_d|$$

- **L_∞** norm:

$$L_\infty(x, y) = \max\{|x_1 - y_1|, \dots, |x_d - y_d|\}$$

Lipschitz Condition

“Any two individuals x, y that are at distance $d(x, y) \in [0, 1]$ map to distributions $M(x)$ and $M(y)$, respectively, such that the statistical distance between $M(x)$ and $M(y)$ is at most $d(x, y)$”

- **Difference in outputs<= Difference in inputs**
- **$D(M(x)-M(y))<=d(x,y)$**

Loss Function

– The loss function measures the difference between the algorithm, output and the ground truth outcome.

– Minimizing loss function **better and fairer**!

– The goal: Find a mapping from individuals to distributions over outcomes that minimizes expected loss function subject to the Lipschitz condition.

$$
\begin{aligned}
\mathrm{opt}(\mathcal{I}) \stackrel{\text{def}}{=} \quad & \min_{\{\mu_x\}_{x\in V}} \quad \mathop{\mathbb{E}}_{x\sim V}\, \mathop{\mathbb{E}}_{a\sim\mu_x} L(x,a) \\
\text{subject to} \quad & \forall x, y \in V, : \quad D(\mu_x, \mu_y) \le d(x,y) \\
& \forall x \in V : \quad \mu_x \in \Delta(A)
\end{aligned}
$$

Group Fairness

– Statistical Parity - Require admissions match demographics in data

– Equal Opportunity - Require false-negative rate to be equal across groups

– Predictive Equality - Require false-positive rate to be equal across groups

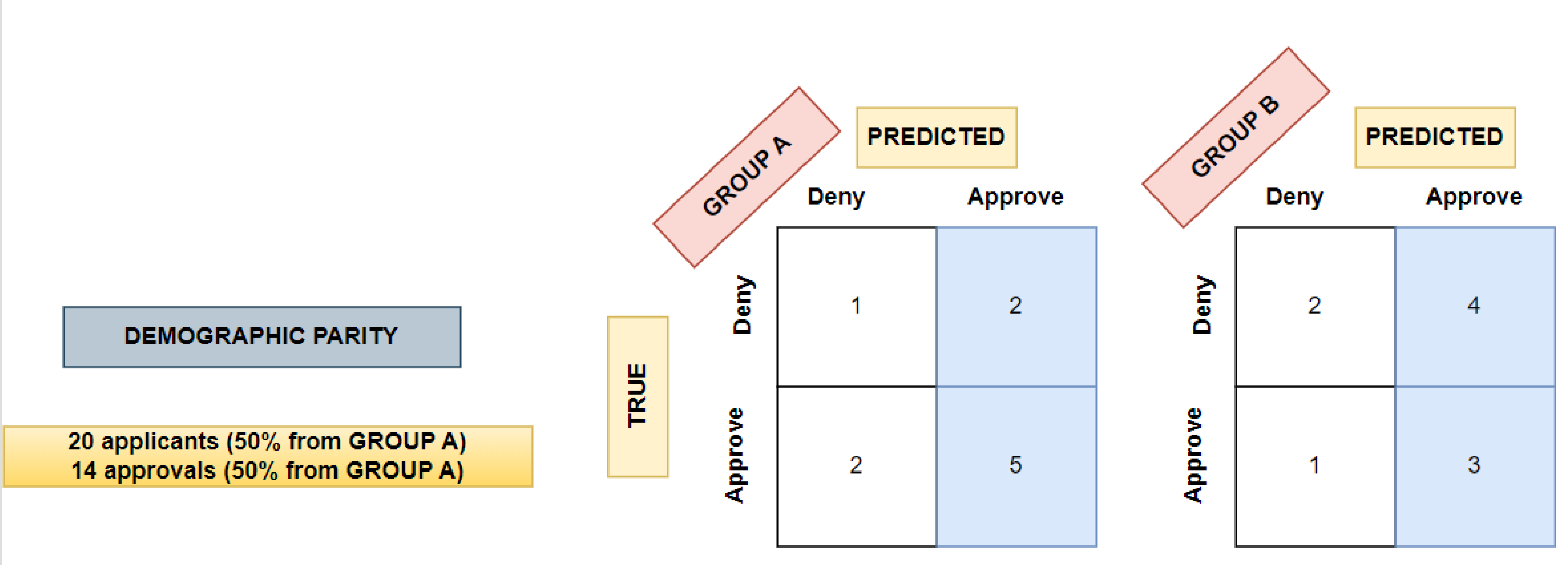
Statistical Parity

- The most well-known criteria for fairness
- **Definition**: % of individual classified positive/negative matches the % demographic of general population
- Example. Hire the same % of individuals in both groups
- **Mathematically**: Decision d is statistically independent of sensitive attribute a
- $Pr(d=1 \mid a) = Pr(d=1)$
- $Pr(d=1 \mid y=0, a) = Pr(d=1 \mid y=0)$
- $Pr(d=1 \mid y=1, a) = Pr(d=1 \mid y=1)$

$$\boxed{Pr[\text{outcome} \mid \text{person in } S] = Pr[\text{outcome} \mid \text{person in } T]}$$

Statistical Parity-perspective of confusion matrix

– Example: A loan is correctly judged when the approval or rejection decision is correct



Statistical Parity

Suitable when:

– We want to change the state of the current world to improve it by supporting unprivileged groups (e.g. universities are aiming to improve diversity by admitting a fixed number of students from disadvantages backgrounds)

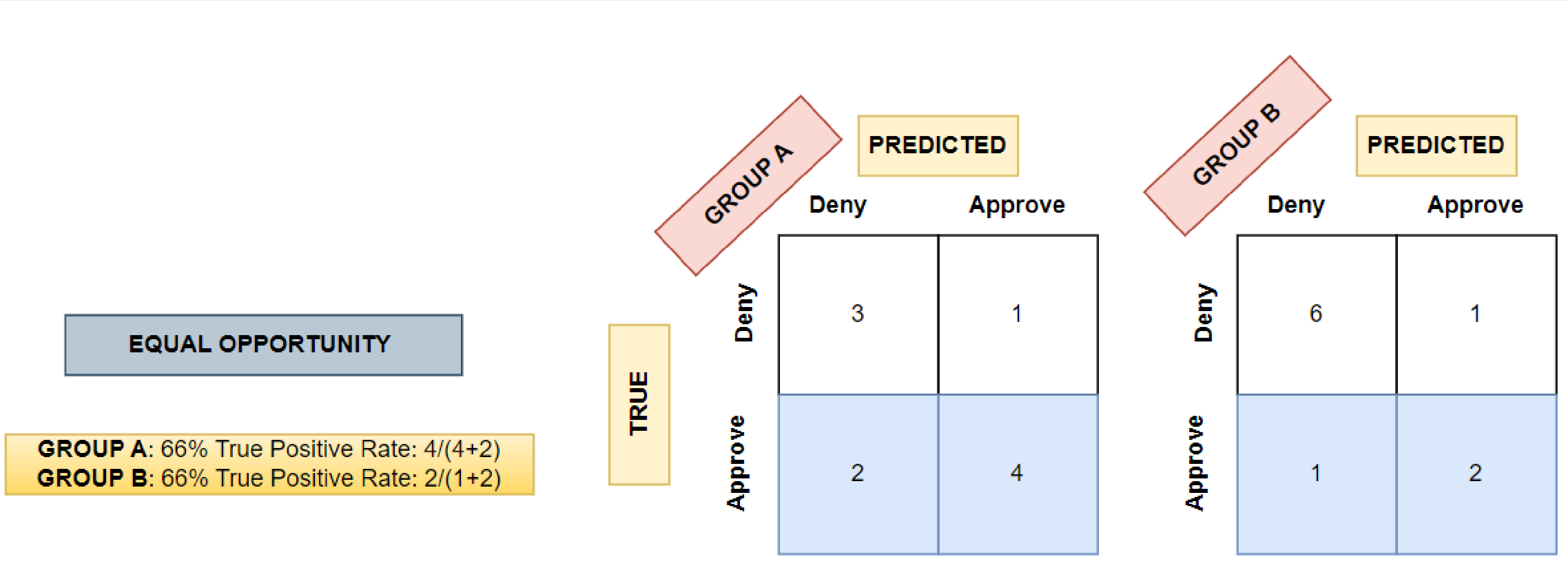
Equal Opportunity

- **Definition**: Equal Opportunity states that each group should get the positive outcome at nearly equal rates, assuming that people in this group qualify for it.
- Hiring example:
  - C=the decision made (hire or reject)
  - Y=the true standard of whether a person was qualified enough or not to be hired
  - Ex. One can be rejected (C=0) but be capable enough for the job (Y=1)
  - Hire equal % of individuals from the qualified subset of each group

$$\boxed{Pr_1[C = c \mid Y = y] = Pr_2[C = c \mid Y = y]}$$

Equal Opportunity-perspective of confusion matrix

– **TPR** should be similar across protected groups that are defined by a sensitive attribute (e.g. race, gender)



Equal Opportunity

Suitable :

- To predict the positive outcome correctly (e.g. detecting a fraudulent transaction)
- Introducing false positives are not costly (e.g. wrongly notifying a customer about fraudulent activity will not be necessarily expensive to the customer nor the bank sending the alert)
- The target variable is not subjective (e.g.: labelling who is a ‘good’ employee is very subjective)

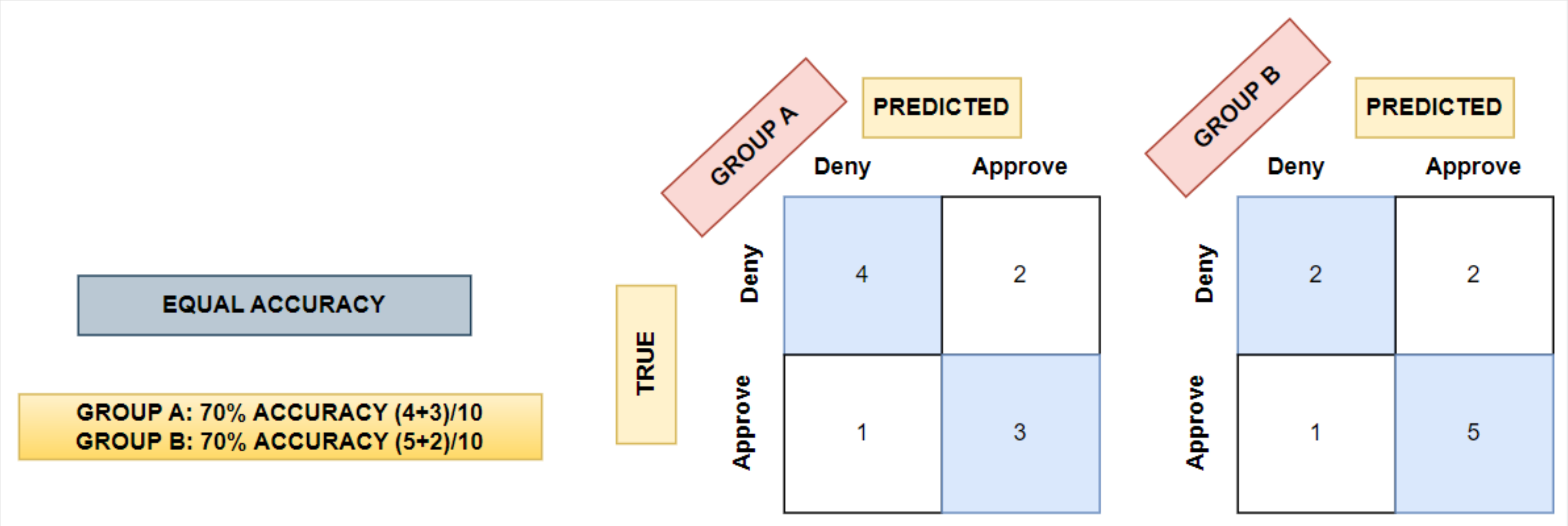
Predictive parity

- Both groups will have the same precision. i.e. for all the positives they predicted, they have the same proportions that the predictions are correct (true positive).
- Hiring example:
  - C=the decision made (hire or reject)
  - Y=the true standard of whether a person was qualified enough or not to be hired
  - Ex. One can be rejected (C=0) but be capable enough for the job (Y=1)

$$\boxed{Pr_1[Y = y \mid C = c] = Pr_2[Y = y \mid C = c]}$$

Predictive parity-perspective of confusion matrix

– Accuracy parity requires equal accuracy across groups



Group Fairness

– In practice, it is not possible to optimize a model for more than one type of fairness.

  -Further reading: https://arxiv.org/abs/2007.06024

– So which fairness criterion should you select, if you can only satisfy one?

  -As with most ethical questions, the correct answer is usually not straightforward, as real-world models typically cannot be expected to satisfy any fairness definition perfectly.

Group Fairness vs. Individual Fairness

- Group fairness does not guarantee individual fairness, or vice versa
- Individual fairness, under certain circumstances, can promote group fairness
- Since group fairness requires to satisfy conditions only on average among groups, it leaves room to bias discrimination inside the groups.

Conclusions

- So which fairness criterion should you select, if you can only satisfy one?

  - As with most ethical questions, the correct answer is usually not straightforward, as real-world models typically cannot be expected to satisfy any fairness definition perfectly.
  - The suitable notion of fairness must be chosen in the context of the specific use case and data at hand.
  - While in financial services group fairness can be adopted, it would not be appropriate in medical applications where gender and race can play an important role in understanding a patient’s symptoms.

Reading

– Castelnovo, A., Crupi, R., Greco, G. et al. A clarification of the nuances in the fairness metrics landscape. Sci Rep 12, 4209 (2022).

– Barocas, S., Hardt, M. & Narayanan, A. Fairness and Machine Learning (fairmlbook.org, 2019). http://www.fairmlbook.org.

– Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) 54, 1–35 (2021).