Assessing Group Fairness with Social Welfare Optimization

Violet (Xinying) Chen
Stevens Institute of Technology

John Hooker*
Carnegie Mellon University

Derek Leben
Carnegie Mellon University

**presenter*

CPAIOR 2024

Uppsala University

Coauthors

Violet Chen
Stevens Institute of Technology

Derek Leben
Carnegie Mellon University

Fundamental Question

- Can **optimization theory** can shed light on the intensely discussed issue of how to achieve **fairness in AI**?
    - We explore the implications for **group parity** of **maximizing social welfare** in the population as a whole.

Group Parity Metrics

- Group parity metrics are widely used in AI
  - To assess whether demographic **groups** are treated **equally**
  - **Selection rates** are compared for:
    - **Job interviews**
    - **University admissions**
    - **Mortgage loans, etc.**
- A "**protected group**" is compared with the **rest** of the population
  - Groups defined by **race**, **gender**, **ethnicity**, **class**, **region**, etc.
  - Sometimes based on **legal** mandates
- We study parity metrics as an **assessment tool**
  - Rather than a selection criterion

Problems with Group Parity

- Group parity is intuitively appealing **at first**…
    - But is it really **fair**?
    - On closer examination, it raises many **problems**:

Problems with Group Parity

- Group parity is intuitively appealing **at first**…
    - But is it really **fair**?
    - On closer examination, it raises many **problems**:
- Failure to account for actual **welfare consequences**
    - Considers only **frequency** of selection
    - For example, rejection may be **more harmful** to a protected group

Problems with Group Parity

- Group parity is intuitively appealing **at first**…
  - But is it really **fair**?
  - On closer examination, it raises many **problems**:
- Failure to account for actual **welfare consequences**
  - Considers only **frequency** of selection
  - For example, rejection may be **more harmful** to a protected group
- Controversy over **which metric** is appropriate
  - **Many statistical metrics** have been proposed
  - Some are mutually **incompatible**

Problems with Group Parity

- Group parity is intuitively appealing **at first**…
  - But is it really **fair**?
  - On closer examination, it raises many **problems**:
- Failure to account for actual **welfare consequences**
  - Considers only **frequency** of selection
  - For example, rejection may be **more harmful** to a protected group
- Controversy over **which metric** is appropriate
  - **Many statistical metrics** have been proposed
  - Some are mutually **incompatible**
- Unclear how to **identify** protected groups
  - Groups often have **conflicting interests**
  - **No limit** to groups that may cry “unfair.”

Some Parity Metrics

- **Demographic parity**.
  - Same **fraction of each group** is selected.

$$P(D|Z) = P(D|\neg Z)$$

Selected

Protected

Not protected

Some Parity Metrics

- **Demographic parity**.
  - Same **fraction of each group** is selected.

$$P(D|Z) = P(D|\neg Z)$$

Selected (D) · Protected (Z) · Not protected (¬Z)

- **Equalized odds** (specifically, equality of opportunity)
  - Same fraction of **qualified** members of each group are **selected**
  - Qualified = offered a job, repays mortgage, success in school.

$$P(D|Y, Z) = P(D|Y, \neg Z)$$

Qualified (Y)

Some Parity Metrics

- **Demographic parity**.
  - Same **fraction of each group** is selected.

$$P(D|Z) = P(D|\neg Z)$$

Selected → D; Protected → Z; Not protected → $\neg Z$

- **Equalized odds** (specifically, equality of opportunity)
  - Same fraction of **qualified** members of each group are **selected**
  - Qualified = offered a job, repays mortgage, success in school.

$$P(D|Y,Z) = P(D|Y,\neg Z)$$

Qualified → Y

- **Predictive rate parity**
  - Same fraction of **selected** members of each group are **qualified**

$$P(Y|D,Z) = P(Y|D,\neg Z)$$

Example: Parole Decisions

- **Objective: Select prisoners for parole.**
  - Based on AI-predicted recidivism rates.
  - Without discriminating against minority candidates
  - Northpointe (now Equivant) developed the COMPAS system for parole decisions.

Example: Parole Decisions

- **Objective: Select prisoners for parole**.
  - Based on AI-predicted recidivism rates.
  - Without discriminating against minority candidates
  - Northpointe (now Equivant) developed the COMPAS system for parole decisions.
- **Controversy**
  - COMPAS is **unfair** because it fails to **equalize odds**.
    - **It applies a *stricter standard* to minority candidates than to majority candidates.**
  - COMPAS is **fair** because it achieves **predictive rate parity**
    - **It ensures that *paroled* minority and majority candidates *have equal recidivism rates***
  - **Which** parity metric is appropriate?

Fairness as Social Welfare

- Group fairness through **population-wide social welfare**
  - Perhaps a **broader concept of distributive justice** can assess parity metrics and achieve fairness across multiple groups
    - **while taking *welfare* into account.**

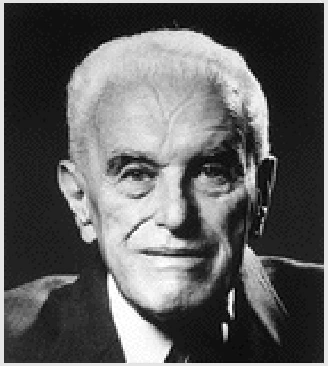
Fairness as Social Welfare

- Group fairness through **population-wide social welfare**
  - Perhaps a **broader concept of distributive justice** can assess parity metrics and achieve fairness across multiple groups
    - **while taking *welfare* into account.**
- Assessing fairness with a **social welfare function**
  - Let $\boldsymbol{u} = (u_1, \ldots, u_n)$ be **utilities** distributed to stakeholders 1, …, n
  - Utility = some kind of **benefit**
    - **Wealth, negative cost, resources, health, etc.**
  - A social welfare function W($\boldsymbol{u}$) measures the desirability of $\boldsymbol{u}$
    - **Taking into account overall utility as well as how it is distributed.**

Alpha fairness

- Focus on **alpha fairness** as a social welfare function
  - Frequently used in engineering, etc.
  - Various forms studied for over 70 years.
    - **In particular, by 2 Nobel laureates (John Nash, J.C. Harsanyi).**
  - Defended by axiomatic and bargaining arguments
    - ***Axiomatic arguments:*** **Nash (1950), Lan, Kao & Chiang (2010,2011)**
    - ***Bargaining arguments:*** **Harsanyi (1977), Rubinstein (1982), Binmore, Rubinstein & Wolinksy (1986)**

John Nash



J. C. Harsanyi

Alpha Fairness

- The **alpha fairness** social welfare function:

$$W_\alpha(\boldsymbol{u}) = \begin{cases} \frac{1}{1-\alpha} \sum_i u_i^{1-\alpha} & \text{for } \alpha \geq 0,\ \alpha \neq 1 \\ \sum_i \log(u_i) & \text{for } \alpha = 1 \end{cases}$$

where u_i is the utility allocated to individual i

  - **Larger α** implies **more fairness**.
  - **Utilitarian** when α = 0, **maximin** (Rawlsian) when $\alpha \to \infty$
  - **Proportional fairness** (Nash bargaining solution) when α = 1
  - **α < 1** incentivizes **competition**, **α > 1** incentivizes **cooperation**
- To achieve alpha fairness:

Maximize $W_\alpha(\boldsymbol{u})$ subject to resource constraints.

Alpha Fairness

- Alpha fair selection

Let $x_i = 1$ if individual i is selected, 0 otherwise.
Then $u_i = a_i x_i + b_i$, where $a_i =$ **selection benefit**
$b_i =$ base utility .

Now

$$W_\alpha(\boldsymbol{u}) = \begin{cases} \dfrac{1}{1-\alpha} \sum_i (a_i x_i + b_i)^{1-\alpha} & \text{for } \alpha \geq 0,\ \alpha \neq 1 \\ \sum_i \log(a_i x_i + b_i) & \text{for } \alpha = 1 \end{cases}$$

We want to maximize $W_\alpha(\boldsymbol{u})$ subject to $x_i \in \{0, 1\}$ and

$$\sum_i x_i = m$$

Number of individuals selected

Alpha Fairness

- An algebraic trick leads to a solution algorithm

If $\alpha \neq 1$, we have

$$W_\alpha(\boldsymbol{u}) = \boxed{\frac{1}{1-\alpha}\sum_i b_i^{1-\alpha}} + \frac{1}{1-\alpha}\sum_i \left((a_i x_i + b_i)^{1-\alpha} - b_i^{1-\alpha}\right)$$

Constant term

Alpha Fairness

- An algebraic trick leads to a solution algorithm

If $\alpha \neq 1$, we have

$$W_\alpha(\boldsymbol{u}) = \frac{1}{1-\alpha}\sum_i b_i^{1-\alpha} + \boxed{\frac{1}{1-\alpha}\sum_i \left((a_i x_i + b_i)^{1-\alpha} - b_i^{1-\alpha}\right)}$$

So we can maximize

$$\sum_{i|x_i=1} \boxed{\frac{1}{1-\alpha}\left((a_i + b_i)^{1-\alpha} - b_i^{1-\alpha}\right)}$$

x_i eliminated from expression

Alpha Fairness

- An algebraic trick leads to a solution algorithm

If $\alpha \neq 1$, we have

$$W_\alpha(\boldsymbol{u}) = \frac{1}{1-\alpha}\sum_i b_i^{1-\alpha} + \boxed{\frac{1}{1-\alpha}\sum_i \left((a_i x_i + b_i)^{1-\alpha} - b_i^{1-\alpha} \right)}$$

So we can maximize

$$\sum_{i|x_i=1} \boxed{\frac{1}{1-\alpha}\left((a_i + b_i)^{1-\alpha} - b_i^{1-\alpha} \right)} = \sum_{i|x_i=1} \boxed{\Delta_i(\alpha)}$$

Welfare differential **of individual *i***
= net increase in social welfare that results from selecting individual *i*

Alpha Fairness

- An algebraic trick leads to a solution algorithm

If $\alpha \neq 1$, we have

$$W_\alpha(\boldsymbol{u}) = \frac{1}{1-\alpha}\sum_i b_i^{1-\alpha} + \boxed{\frac{1}{1-\alpha}\sum_i \Big((a_i x_i + b_i)^{1-\alpha} - b_i^{1-\alpha}\Big)}$$

So we can maximize

$$\sum_{i|x_i=1} \boxed{\frac{1}{1-\alpha}\Big((a_i + b_i)^{1-\alpha} - b_i^{1-\alpha}\Big)} = \sum_{i|x_i=1} \boxed{\Delta_i(\alpha)}$$

Welfare differential **of individual *i***
= net increase in social welfare that
results from selecting individual *i*

... by selecting the m individuals with the largest welfare differentials $\Delta_i(\alpha)$. Similarly if $\alpha = 1$.

Alpha Fairness

- We assume that, **within a group**, individuals with the **largest** selection benefit are selected **first**.
  - This means that individuals with **largest welfare differential** are selected first.
  - Since the welfare differential increases monotonically with the selection benefit.

Alpha Fairness Example

α = 0.7, Select 9 individuals

Majority group

a_i	$\Delta_I(0.7)$
1.5	**0.750**
1.4	**0.708**
1.3	**0.665**
1.2	**0.621**
1.1	**0.577**
1.0	**0.531**
0.9	0.484
0.8	0.436
0.7	0.387
0.6	0.336

Protected group

a_i	$\Delta_I(0.7)$
0.2	0.187
0.4	0.354
0.6	**0.505**
0.8	**0.643**
1.0	**0.770**

Alpha Fairness Example

$\alpha = 0.7$, Select 9 individuals

Majority group

a_i	$\Delta_i(0.7)$
1.5	**0.750**
1.4	**0.708**
1.3	**0.665**
1.2	**0.621**
1.1	**0.577**
1.0	**0.531**
0.9	0.484
0.8	0.436
0.7	0.387
0.6	0.336

Protected group

a_i	$\Delta_i(0.7)$
0.2	0.187
0.4	0.354
0.6	**0.505**
0.8	**0.643**
1.0	**0.770**

9 individuals with highest welfare differentials

a_i	$\Delta_i(0.7)$
1.0	**0.770**
1.5	**0.750**
1.4	**0.708**
1.3	**0.665**
0.8	**0.643**
1.2	**0.621**
1.1	**0.577**
1.0	**0.531**
0.6	**0.505**

Alpha Fairness Example

α = 0.7, Select 9 individuals

- Alpha fairness (α = 0.7) corresponds to demographic parity.
  - 6 of 10 majority individuals selected
  - 3 of 5 protected individuals selected
  - 60% of both groups

Welfare differential **of individual *i***
= net increase in social welfare that results from selecting individual *i*

9 individuals with highest welfare differentials

a_i	$\Delta_i(0.7)$
1.0	0.770
1.5	0.750
1.4	0.708
1.3	0.665
0.8	0.643
1.2	0.621
1.1	0.577
1.0	0.531
0.6	0.505

Alpha Fairness Example

α = 0.7, Select 9 individuals

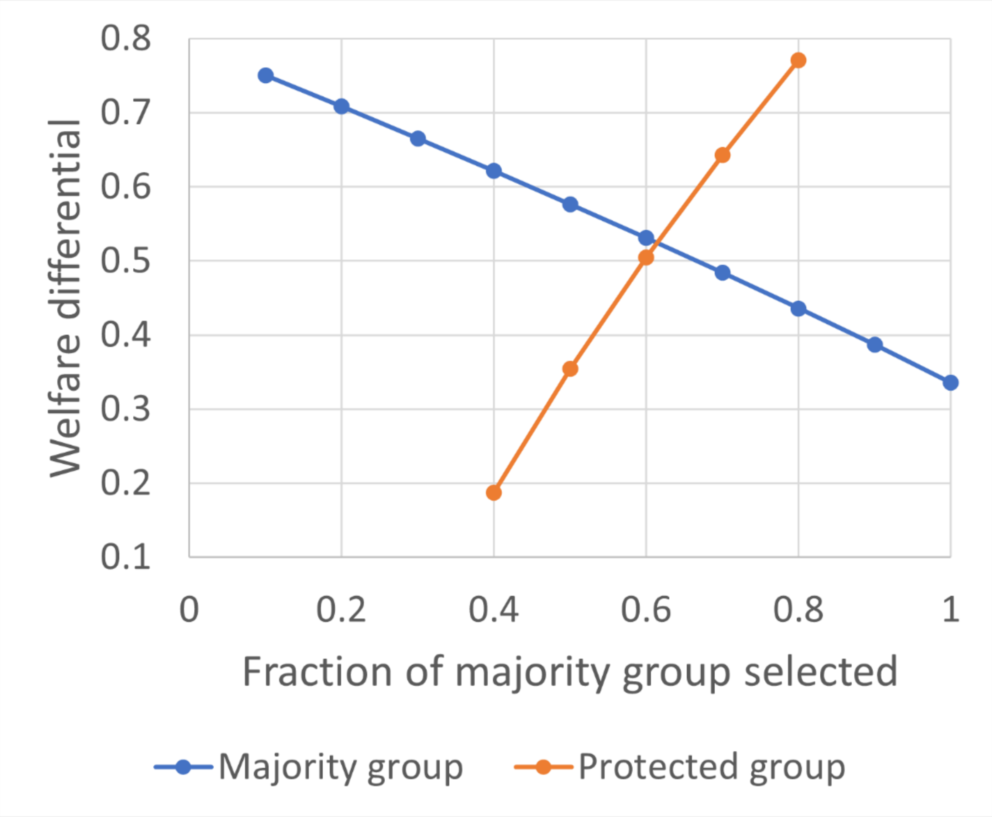
Majority group

a_i	$\Delta_l(0.7)$
1.5	**0.750**
1.4	**0.708**
1.3	**0.665**
1.2	**0.621**
1.1	**0.577**
1.0	**0.531**
0.9	0.484
0.8	0.436
0.7	0.387
0.6	0.336

Protected group

a_i	$\Delta_l(0.7)$
0.2	0.187
0.4	0.354
0.6	**0.505**
0.8	**0.643**
1.0	**0.770**

Graphical interpretation



Utility Model for 2 Groups

- We want a model that relates alpha fairness to the utility characteristics of the majority and projected groups.
    - …while reducing the number of utility parameters
    - Selection benefits **uniformly distributed** in each group
    - Base utility is **constant** in each group
    - More complicated model yields similar results

Majority group

Selection benefits

$A_{\max}$ $A_{\min}$

Base utility = *B*

Protected group

Selection benefits

$a_{\max}$ $a_{\min}$

Base utility = *b*

Utility Model for 2 Groups

- Computing the welfare differentials:

Let S = fraction of majority group selected
s = fraction of protected group selected

Then the welfare differential of the last individual selected in the majority group is

$$\Delta_S(\alpha) = \begin{cases} \frac{1}{1-\alpha}\left(\left((1-S)A_{\max} + SA_{\min} + B\right)^{1-\alpha} - B^{1-\alpha}\right) & \text{if } \alpha \neq 1 \\ \log\left((1-S)A_{\max} + SA_{\min} + B\right) - \log(B) & \text{if } \alpha = 1 \end{cases}$$

and in the protected group is $\Delta'_s(\alpha)$, similarly defined.

Utility Model for 2 Groups

If β = fraction of population that is in the protected group
σ = fraction of population selected, then

$$(1-\beta)S + \beta s = \sigma,$$

which implies

$$s = s(S) = \frac{\sigma - (1-\beta)S}{\beta}$$

and...

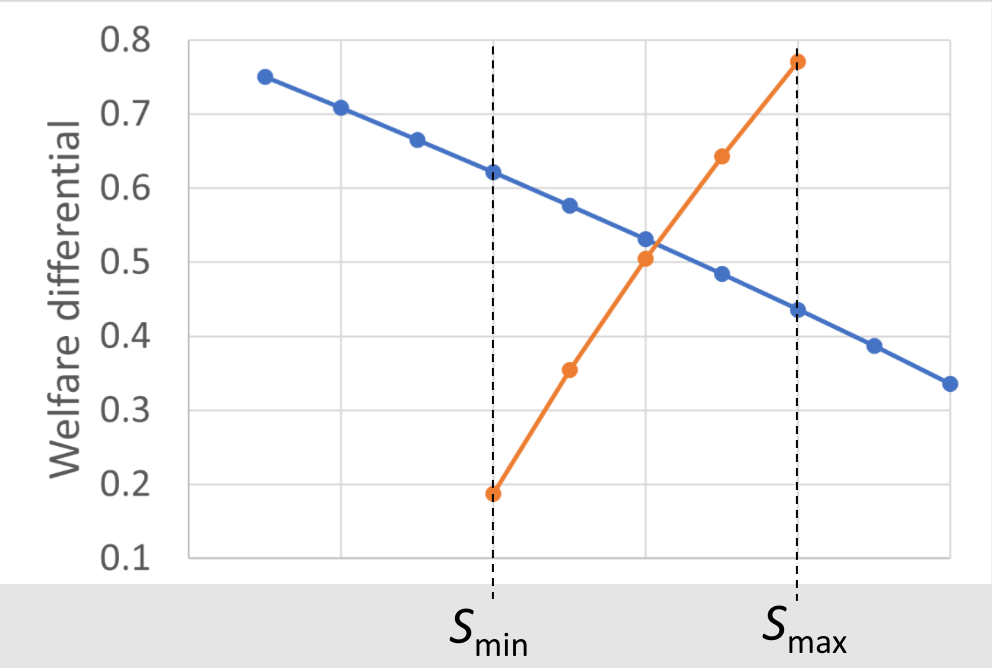
Utility Model for 2 Groups

If β = fraction of population that is in the protected group
σ = fraction of population selected, then

the min and max values of S are

$$S_{\min} = \max\left\{0,\ \frac{\sigma-\beta}{1-\beta}\right\},\quad S_{\max} = \min\left\{1,\ \frac{\sigma}{1-\beta}\right\}$$

$\sigma = 0.6$
$\beta = 1/3$

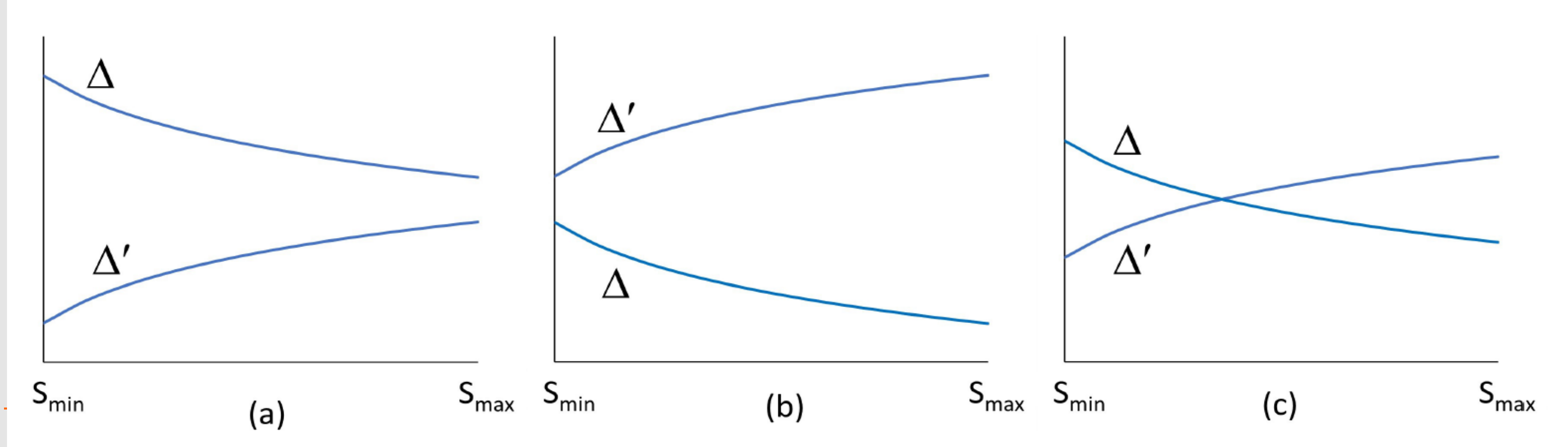


Utility Model for 2 Groups

Theorem. Selection rates (S, s) achieve alpha fairness for a given α if and only if $s = s(S)$ and

$$\begin{cases} (S,s) = \left(\min\left\{1, \frac{1}{1-\beta}\right\}, \frac{\sigma}{\beta}\left[1 - \min\left\{1, \frac{1-\beta}{\sigma}\right\}\right]\right) & \text{in case (a)} \\ (S,s) = \left(\frac{\sigma}{1-\beta}\left[1 - \min\left\{1, \frac{\beta}{\sigma}\right\}\right], \min\left\{1, \frac{\sigma}{\beta}\right\}\right) & \text{in case (b)} \\ \Delta_S(\alpha) = \Delta'_s(\alpha) & \text{in case (c)} \end{cases}$$

where the cases are



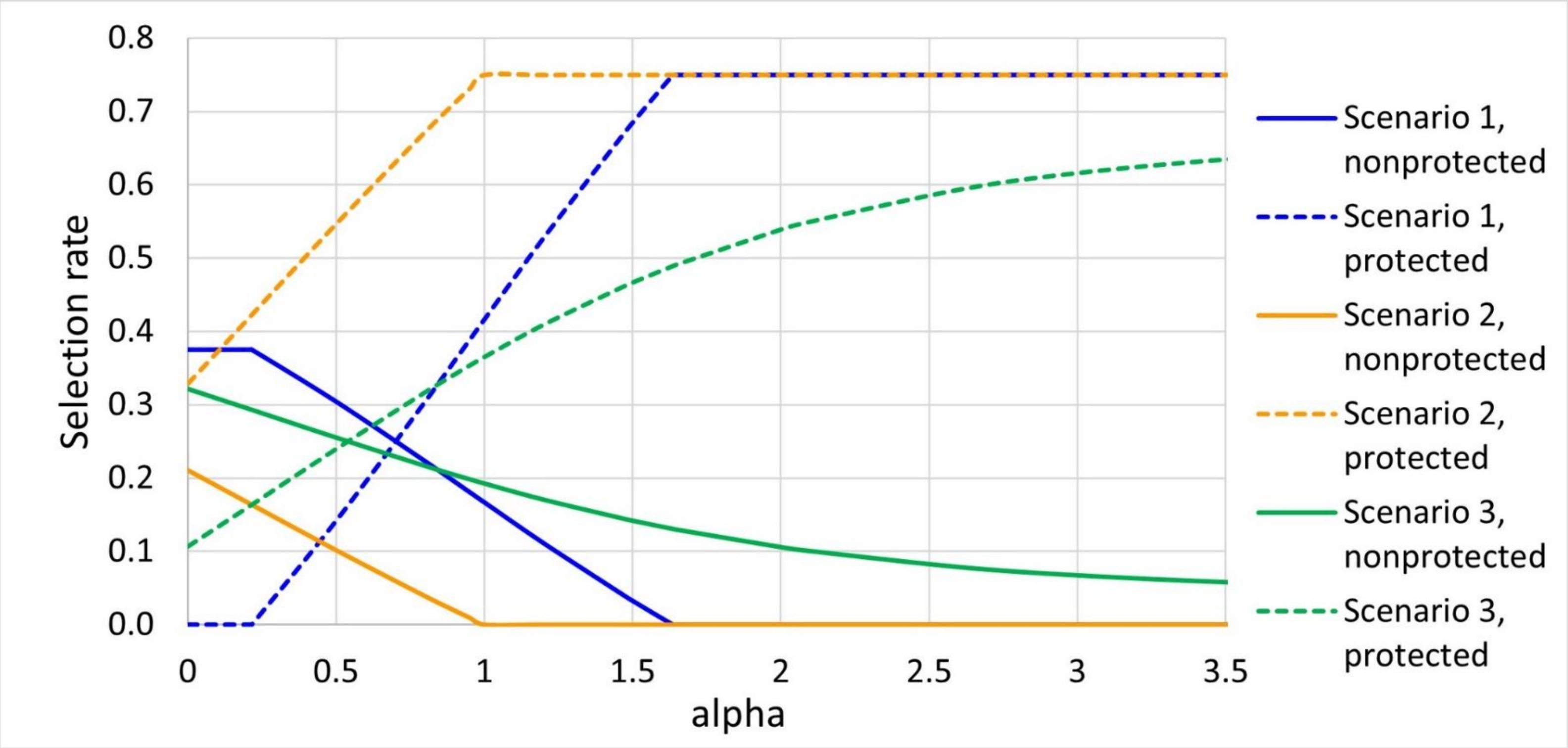
Alpha-fair Selection Rates

- Consider 3 qualitatively different utility scenarios…

	Scenario 1	**Scenario 2**	**Scenario 3**
Majority group	$A_{\min}$ 0.5 — 1.5 $A_{\max}$	$A_{\min}$ 0.5 — 0.8 $A_{\max}$	$A_{\min}$ 0.5 — 1.0 $A_{\max}$
Protected group	$a_{\min}$ 0.2 — 1.0 $a_{\max}$	$a_{\min}$ 0.2 — 1.0 $a_{\max}$	$a_{\min}$ –0.5 — 1.0 $a_{\max}$
	Protected group *benefits somewhat less* from selection	**Some protected individuals *benefit most***	**Some protected individuals *harmed* by selection**
	For example, granting *job interviews*	**For example, *admission* of talented individuals to *university***	**For example, *mortgage loans* with possible *foreclosure***

Alpha-fair Selection Rates

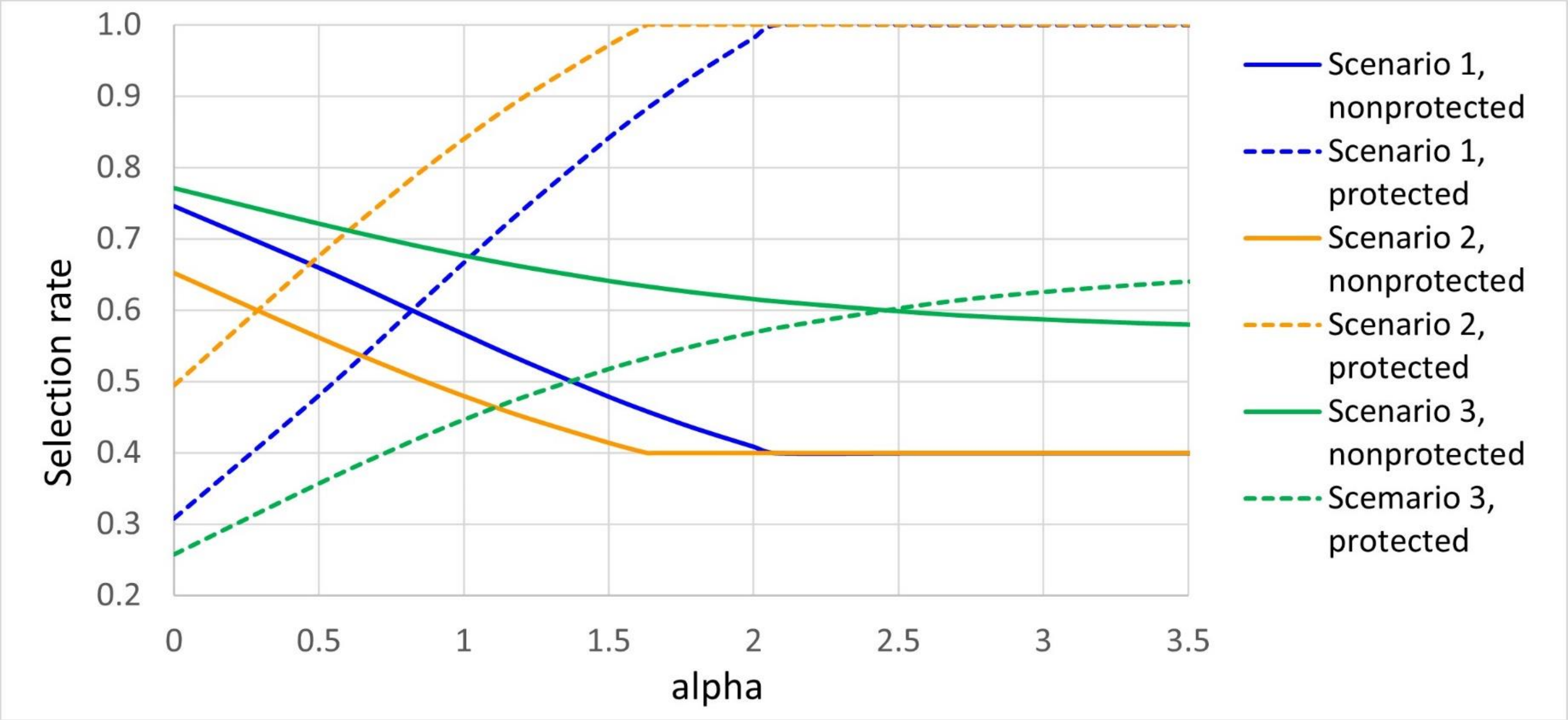
- Overall selection rate = 0.25



- Protected group has **lower** selection rates in Scenario 1 than in Scenario 2 due to **higher utility cost** of fairness in scenario 1.
- Protected group selection rate approaches 2/3 asymptotically because 1/3 of group is **harmed** by selection.

Alpha-fair Selection Rates

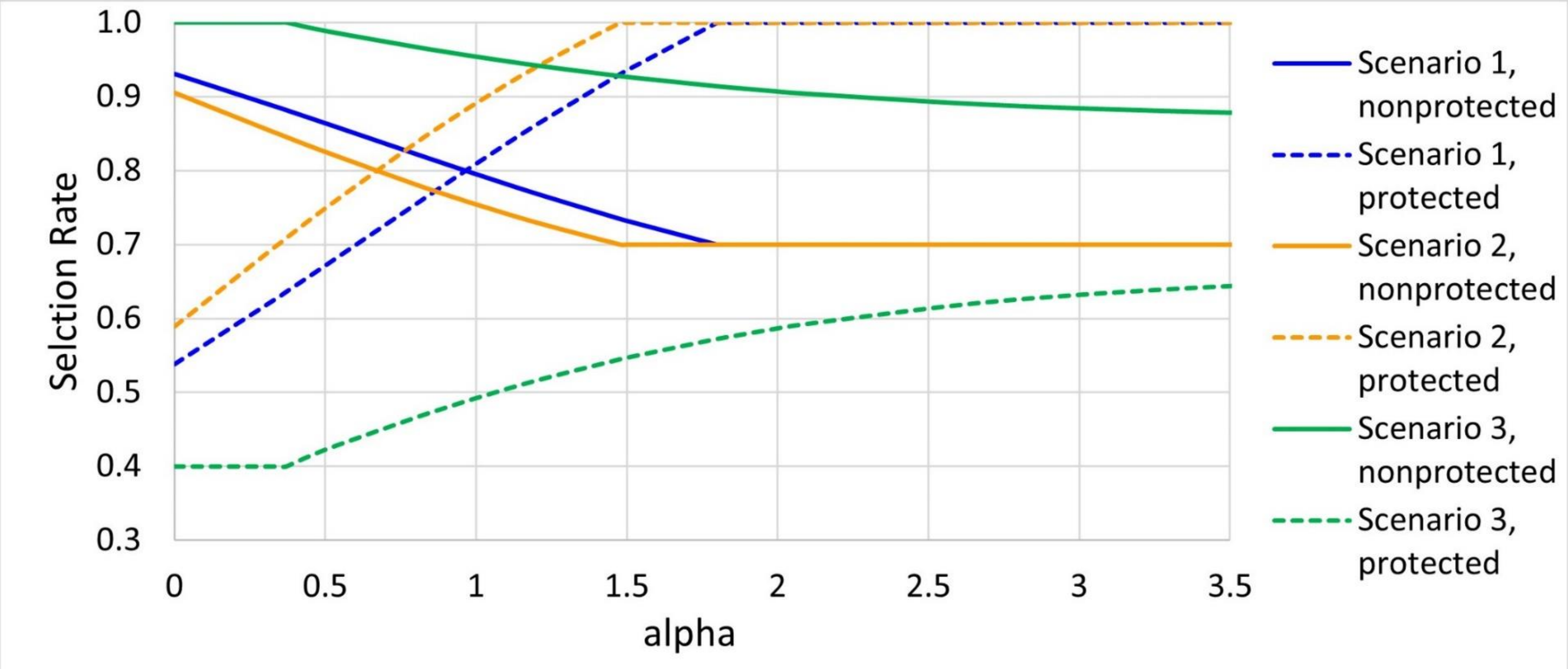
- Overall selection rate = 0.6



- Similar pattern, higher rates.

Alpha-fair Selection Rates

- Overall selection rate = 0.8



- Similar pattern, still higher rates.

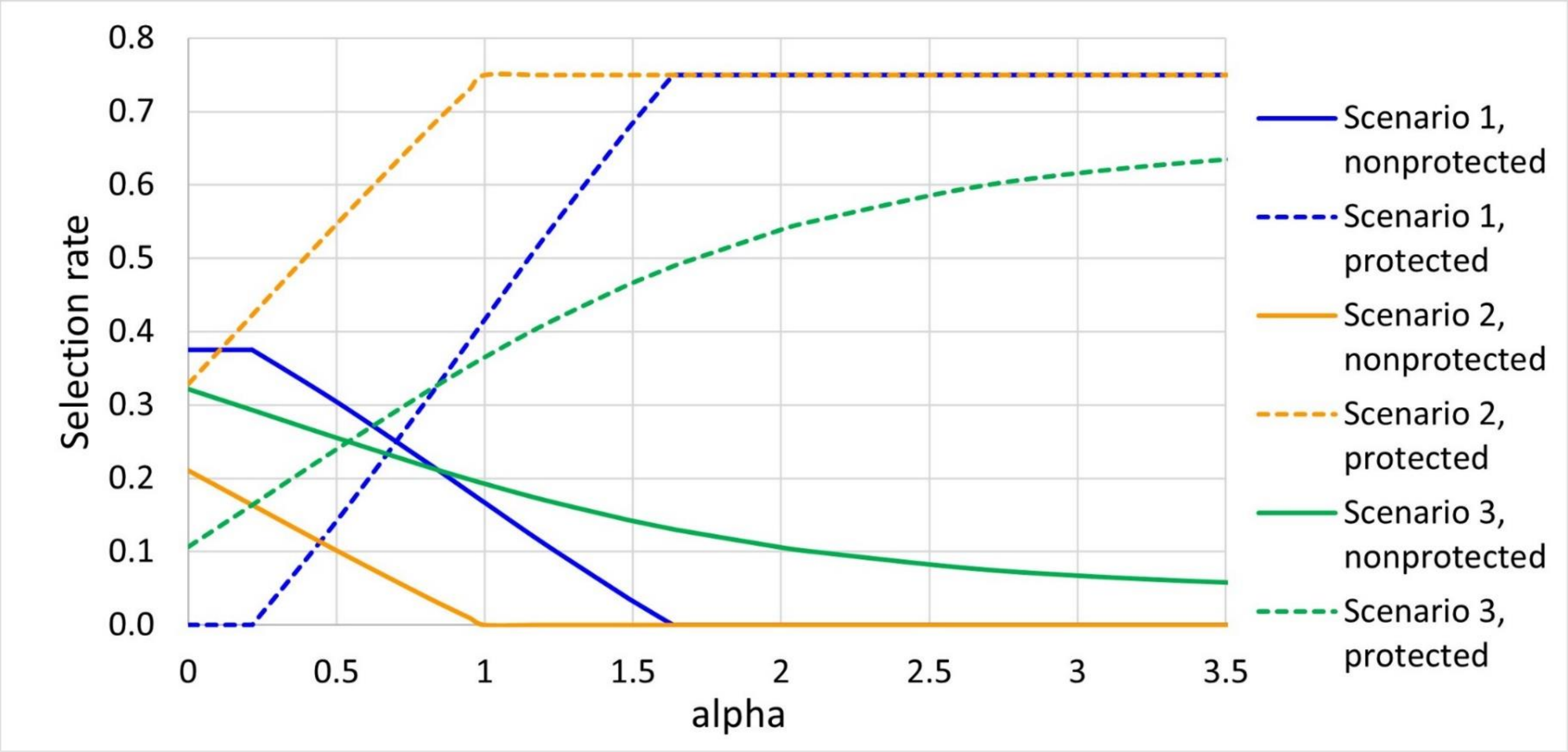
Demographic Parity

Demographic parity is achieved only in case (c), where the Δ curves intersect.

Theorem. An alpha fair selection policy for a given α results in demographic parity if and only if there exists a selection rate S that satisfies the equation $\Delta(S) = \Delta'(S)$, in which case (S, S) is such a policy.

Demographic Parity

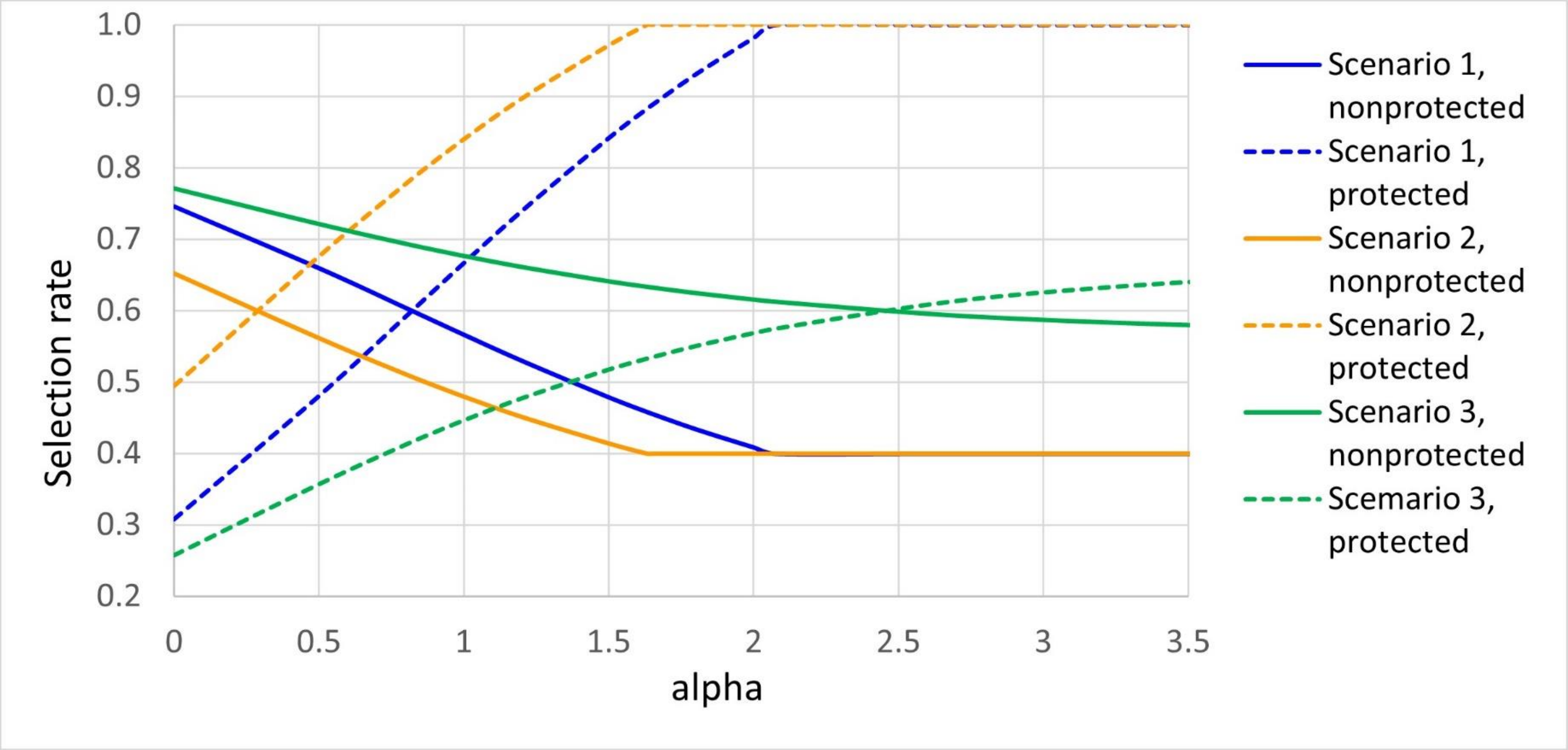
- Overall selection rate = 0.25



- Parity achieved when majority & protected curves **intersect**.
- Parity corresponds to relatively **low** degree of fairness.
- Protected group in Scenario 2 has higher rate even with α **= 0**.

Demographic Parity

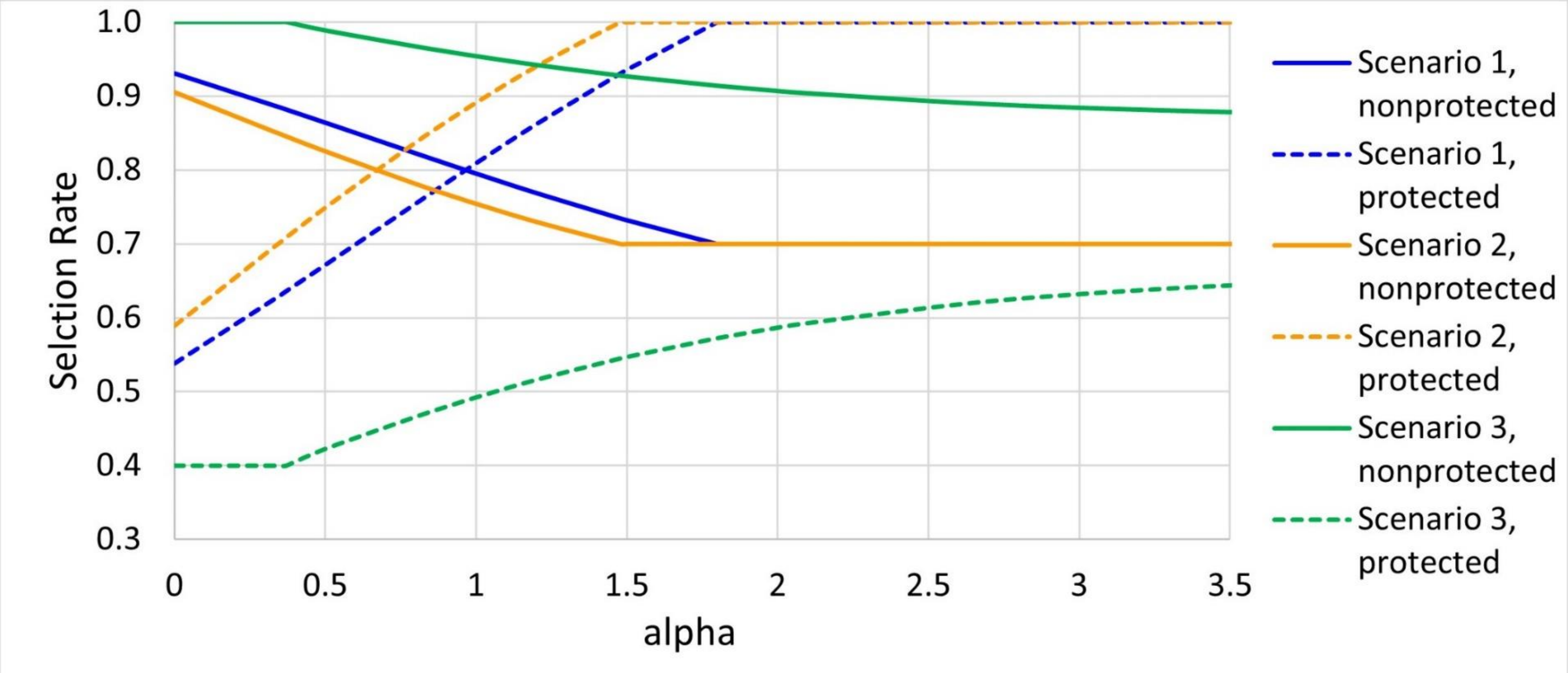
- Overall selection rate = 0.6



- Parity in Scenario 2 now requires a **slight** degree of fairness.
- Scenario 3 parity requires **large** α due to high cost of fairness.

Demographic Parity

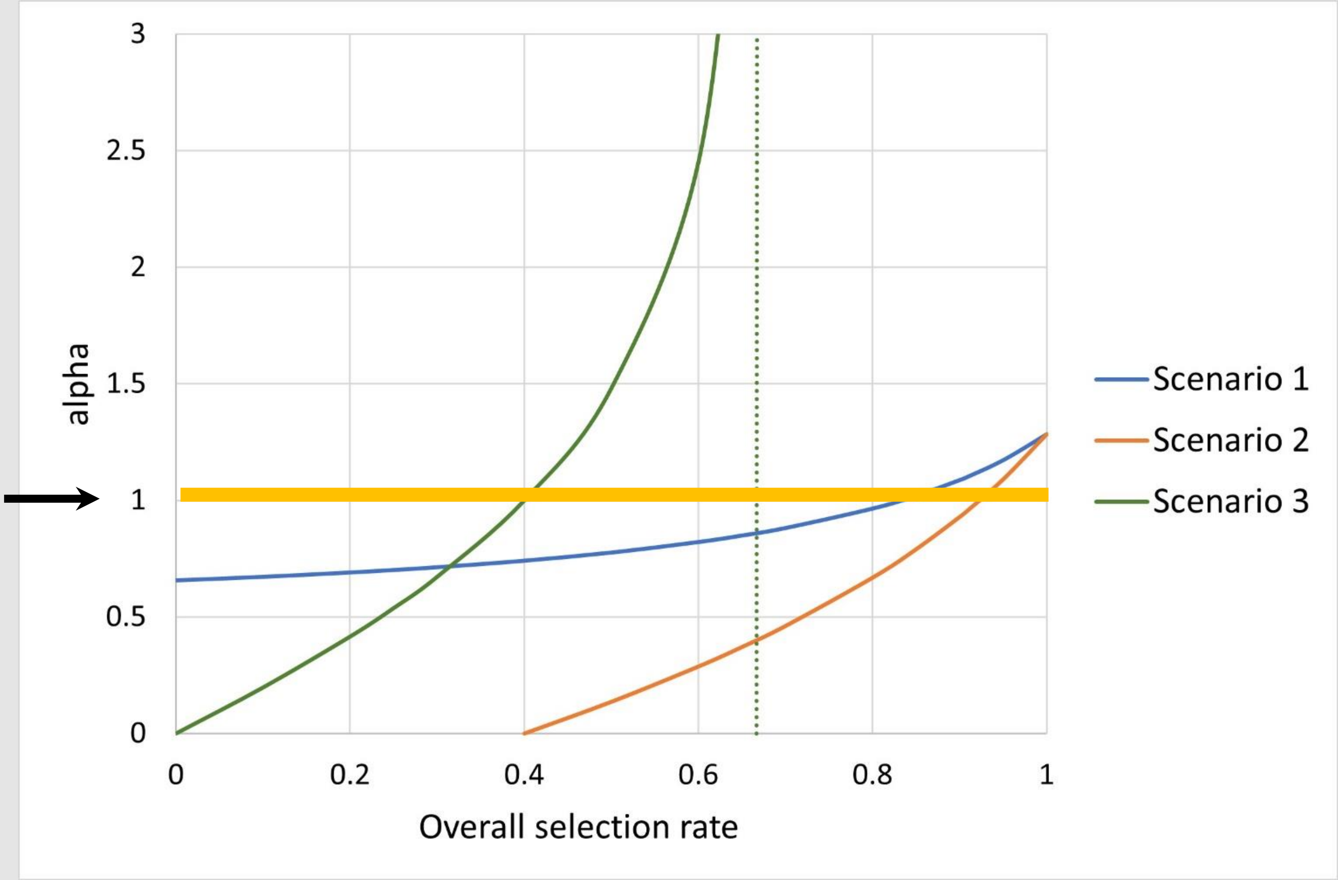
- Overall selection rate = 0.8



- Parity **impossible** in Scenario 3 because alpha fairness never calls for harmful selections.

Demographic Parity

Proportional fairness



- Alpha values that **achieve parity**.
- Parity generally corresponds to **less than proportional fairness**.

Equalized Odds

Suppose a fraction Q of the nonprotected group and a fraction q of the protected group are qualified.

Theorem. An alpha fair selection policy $(S, s(S))$ for a given α and selection rate σ results in equalized odds if and only if one of the following holds:

$$S = Q\rho \leq Q \quad \text{and} \quad s(S) = q\rho \leq q$$
$$S \geq Q \quad \text{and} \quad s(S) \geq q$$

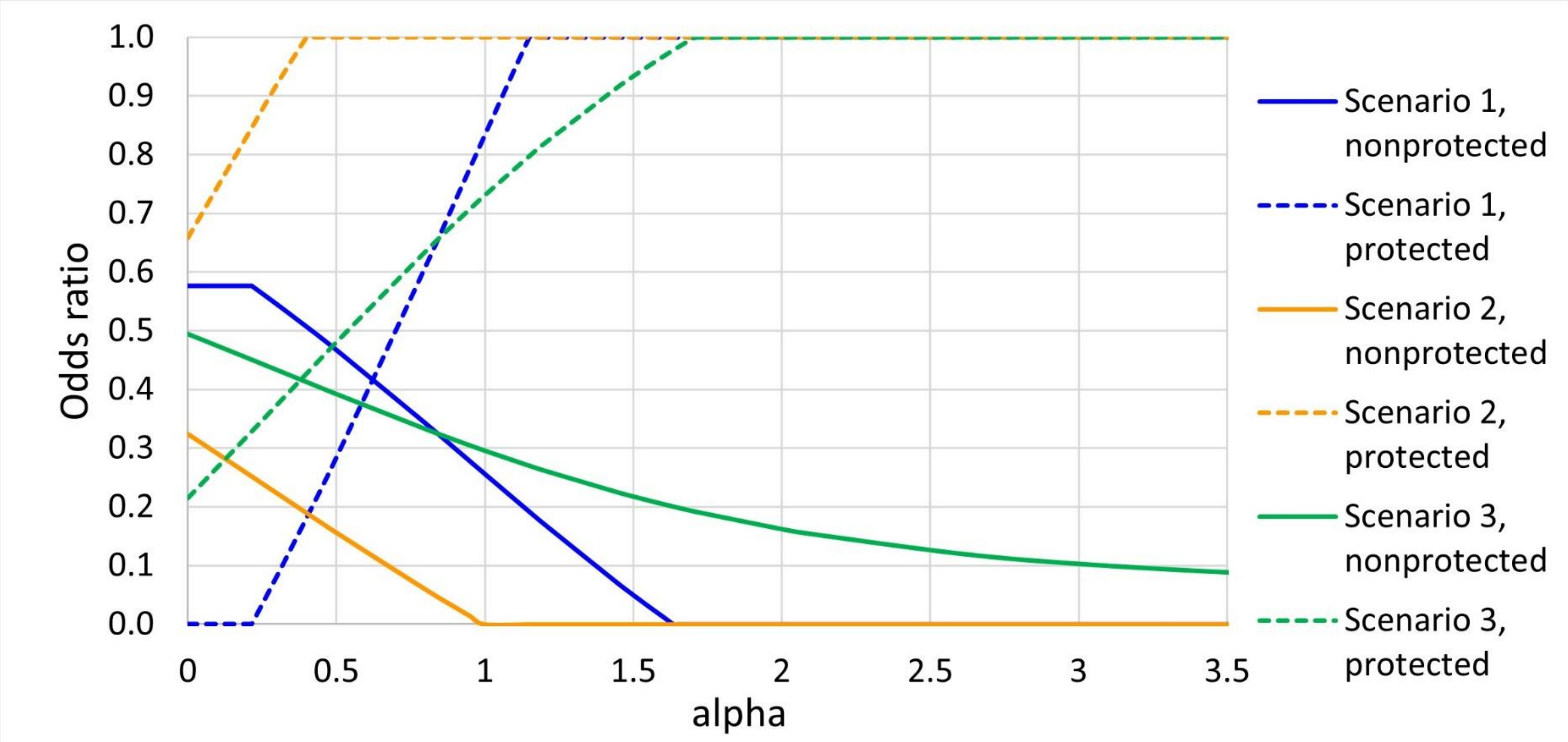
where

$$\rho = \frac{\sigma}{(1-\beta)Q + \beta q}$$

The theorem for predictive rate parity is similar.

Equalized Odds

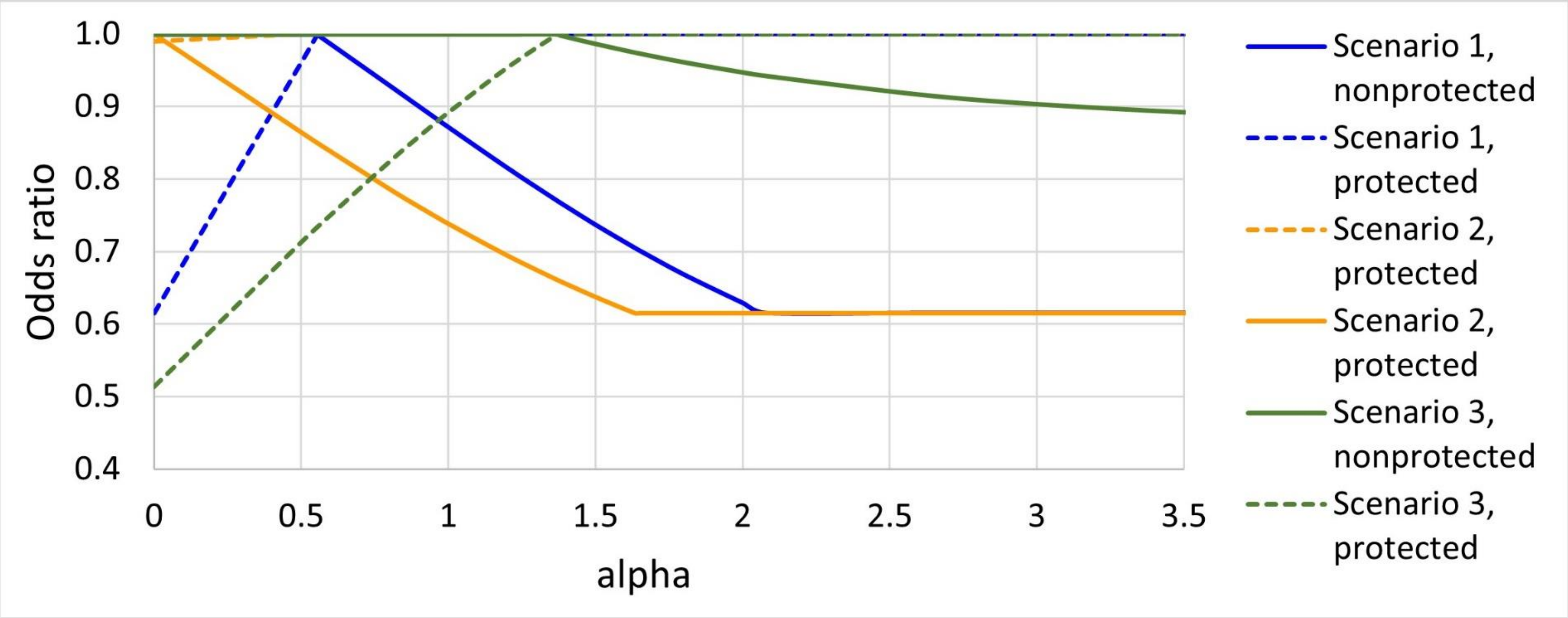
- Assume majority is 65% qualified, protected group 50% qualified.
- Overall selection rate = **0.25** < overall qualification rate of **0.6**



- Even **less fair than demographic parity**.
- Sometimes viewed as **easier to defend** than demographic parity.

Equalized Odds

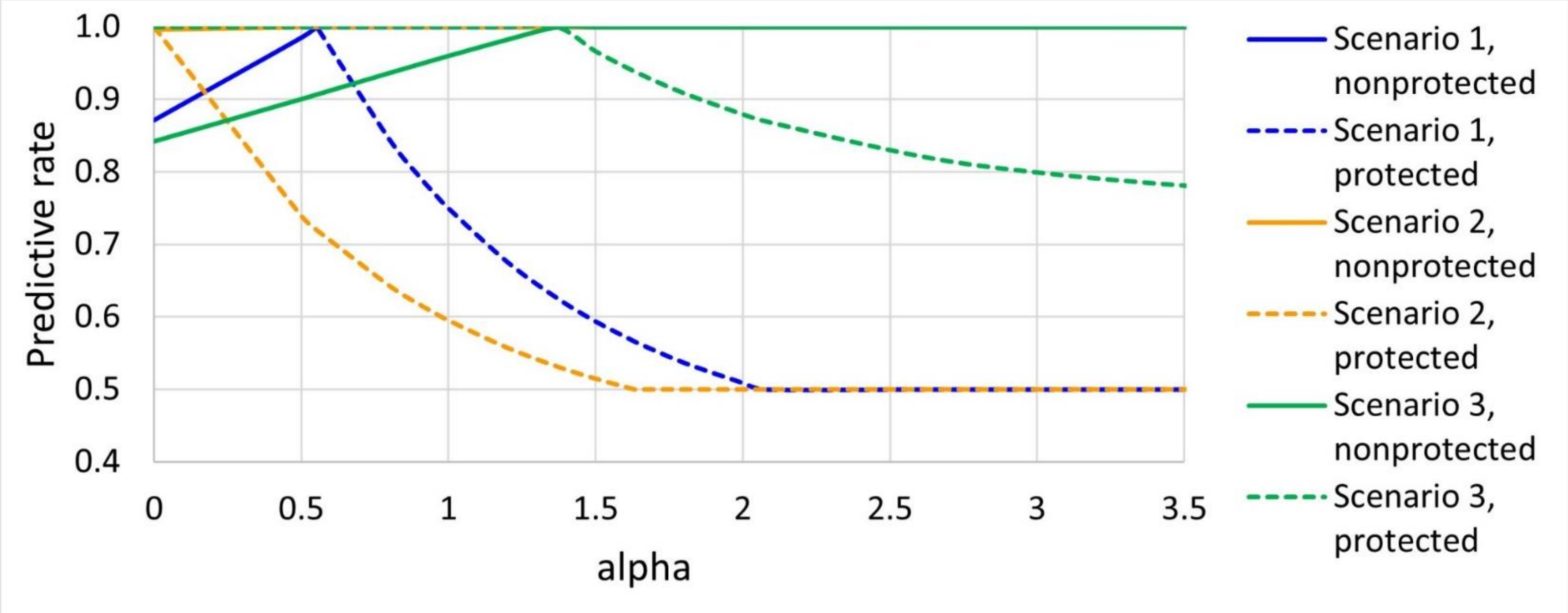
- Overall selection rate = **0.6** = overall qualification rate



- Only an **accuracy maximizing** solution (odds ratio = 1) yields equalized odds. **Fairness not a factor**.
- Nearly all odds ratios = 1 when selecting **more** individuals than are qualified.

Predictive Rate Parity

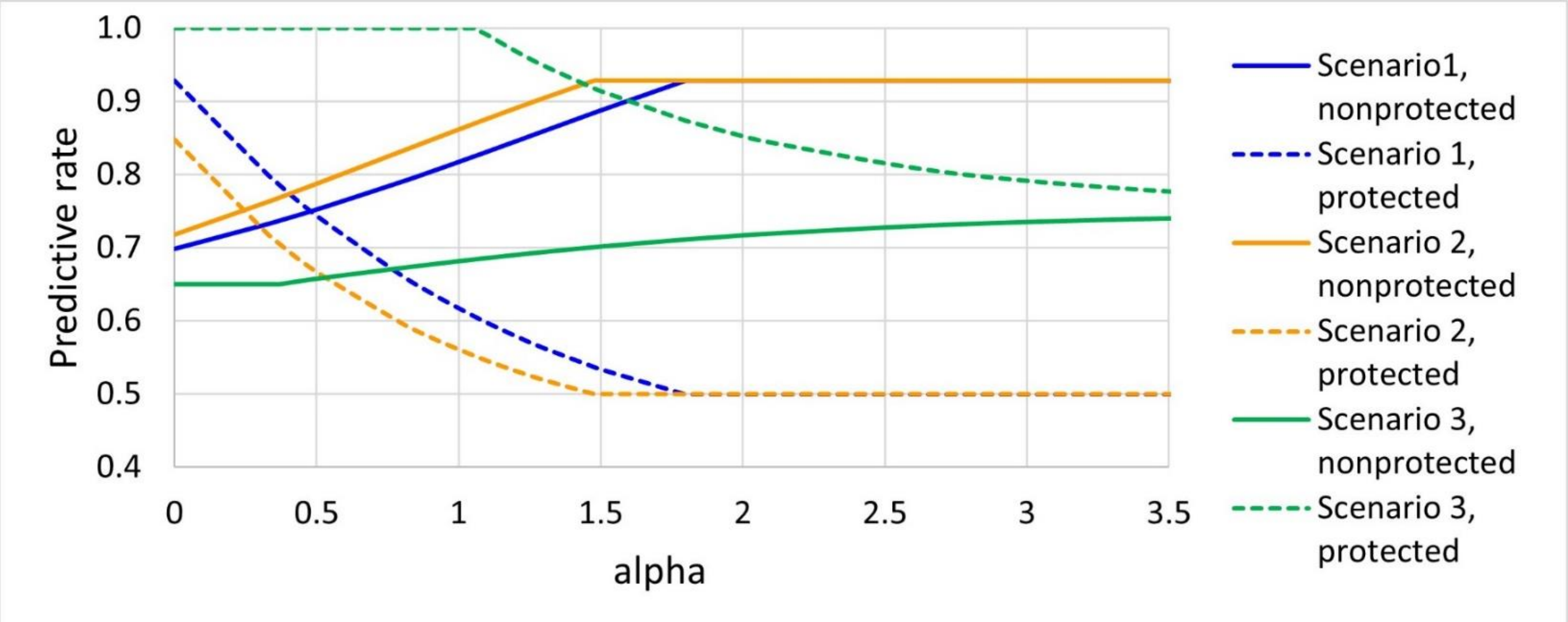
- Overall selection rate = **0.6** = overall qualification rate



- Higher predictive rates = **smaller** selection rates for protected group.
- Only an **accuracy maximizing** solution (pred rate = 1) yields predictive rate parity. **Fairness not a factor**.

Predictive Rate Parity

- Overall selection rate = **0.8** > overall qualification rate



- Nearly all predictive rates = 1 when selecting **fewer** individuals than are qualified.
- Predictive rate parity is a meaningful parity measure only when selecting **more** individuals than are qualified.

Conclusions

- Accounting for **welfare**
  - Alpha fairness takes **utility consequences** into account.
  - It can normally result in **any** of the 3 types of parity, **for suitable** α .
  - **Significant disparity** (favoring the protected group) is often necessary to achieve a specified degree of fairness.

Conclusions

- Accounting for **welfare**
  - Alpha fairness takes **utility consequences** into account.
  - It can normally result in **any** of the 3 types of parity, **for suitable** α .
  - **Significant disparity** (favoring the protected group) is often necessary to achieve a specified degree of fairness.
- **Assessing metrics – demographic parity**
  - Typically corresponds to $\alpha < 1$.
    - ***Less fair* than proportional fairness.**
    - **Even though proportional fairness is something of an *industry standard* in engineering.**

Conclusions

- Accounting for **welfare**
  - Alpha fairness takes **utility consequences** into account.
  - It can normally result in **any** of the 3 types of parity, **for suitable α** .
  - **Significant disparity** (favoring the protected group) is often necessary to achieve a specified degree of fairness.
- **Assessing metrics – demographic parity**
  - Typically corresponds to α **< 1**.
    - ***Less fair*** **than proportional fairness.**
    - **Even though proportional fairness is something of an *industry standard* in engineering.**
- **Assessing metrics – equalized odds & predictive rate**
  - Implications of alpha fairness depend heavily on **how many individuals are selected** relative to number qualified.

Conclusions

- **If number selected = number qualified**
    - Equalized odds and predictive rate parity simply **maximize accuracy**.
        - **Select precisely the qualified individuals in each group.**
        - **So, not a meaningful fairness measure.**

Conclusions

- **If number selected = number qualified**
  - Equalized odds and predictive rate parity simply **maximize accuracy**.
    - **Select precisely the qualified individuals in each group.**
    - **So, not a meaningful fairness measure.**
- **If number selected < number qualified**
  - **Equalized odds** is **less fair** (measured by α) than **demographic parity**.
    - **Which is consistent with the possibility that it is *easier to defend* on ethical grounds.**
  - Predictive rate parity is **less useful.**
    - **Predictive rate is normally 1, since selected individuals tend to be qualified.**

Conclusions

- **If number selected > number qualified**
    - Perhaps an **unusual** situation**.**
        - **Due to limited resources.**
    - Even if it occurs, equalized odds is **not useful.**
        - **Odds ratio is normally 1, since qualified individuals tend to be selected.**
    - **Higher predictive rate** corresponds to **smaller** α (less fairness).
        - **Fairness tends to require *reducing* minority group predictive rate.**

Conclusions

- **Parole** example
  - **Equalized odds** is relevant only if COMPAS paroles **fewer** prisoners than are qualified
    - **That is, fewer than are expected to say out of prison.**
  - **Achieving predictive rate parity** is an **advantage** for COMPAS if it paroles **more** prisoners than are qualified…
    - **Because this ensures that minority prisoners have *no higher predictive rate* than majority prisoners.**
    - **…which ensures that minority prisoners are not required to meet *stricter conditions*.**
    - **COMPAS may choose to parole more prisoners than are qualified in order to reduce the minority predictive rate without tightening parole conditions on the majority.**

Conclusions

- **Multiple protected groups**
  - Parity for **all groups** does not correspond to alpha fairness **for any α.**
    - **Unless the groups are very similar.**
  - However, alpha fairness for a given α can achieve a desired degree of fairness across the population as a whole
    - **and in so doling, treat each group "*fairly*" in view of *its specific circumstances*.**


Questions or
comments?