Tutorial on Fairness Modeling
Part 2: Fairness in AI

John Hooker
Carnegie Mellon University

September 2023

Two Tutorials

- *Previous tutorial:* modeling **fairness in optimization models**
  - *Social welfare functions that incorporate fairness.*
  - *Practical LP/MILP/NLP models.*
  - *A bit of social choice theory.*

- *This tutorial:* modeling **group fairness in AI**
  - *Crash course in deontological ethics.*
  - *Group parity metrics & their assessment.*
  - *Connections with social welfare functions.*

Outline

- Crash course in deontological ethics
  - *Basic assumptions*
  - *Generalization principle*
  - *Autonomy principle*
  - *Utilitarian principle*
- Group parity
  - *Statistical parity metrics*
  - *Ethical assessment*
  - *Social welfare and group parity*
- Beyond group parity

Reference

Castelnovo et al., A clarification of the nuances in the fairness metrics landscape, *Scientific Reports* **12** (2022).

Basic Assumptions

- **Acting for reasons**
  - *Freely chosen action is based on a rationale.*
- **Universality of reason**
  - *Justification is independent of the reasoner.*

Basic Assumptions

- **Acting for reasons**
  - *Freely chosen action is based on a rationale.*
- **Universality of reason**
  - *Justification is independent of the reasoner.*
- We **deduce** ethical principles from these assumptions.
  - *This is the **deontological** approach to ethics.*
    - **Deontology** = What is required.
  - *Ethical principles represent **what is required for the possibility of free action**.*

Acting for Reasons

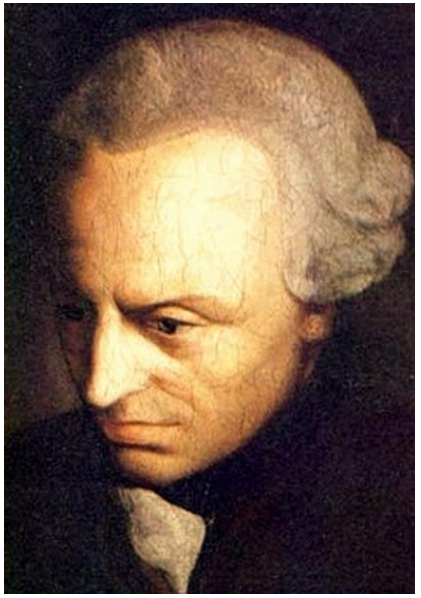
- Basic premise: We always act for a reason.
  - *Every action has a rationale.*
- Why?
  - *This is how we distinguish **freely chosen action** from mere behavior.*
    - An MRI machine can detect our decisions **before we make them**.
    - If decisions are determined by **biological causes**, how can they be freely chosen?

Acting for Reasons

- Solution:
  - *Freely chosen actions have **two kinds of explanation**:*
    - A biological cause
    - A rationale provided by the agent
  - *For example:*
    - A hiccup has **only** a biological explanation.
      Not a freely chosen action.
    - Drinking water to stop hiccups has
      2 explanations: a biological cause
      and a rationale. A freely chosen action.

Acting for Reasons

- Dual standpoint theory
  - *Originally proposed by Immanuel Kant.*
    - *Grundlegung zur Metaphysik der Sitten* (1785)
    - Recent versions: *Nagel (1986), Korsgaard (1996), Nelkin (2000), Bilgrami (2006).*
  - *Provides a **basis for ethics**.*
    - Ethical principles are **necessary conditions** for the logical coherence of an action's rationale



Universality of Reason

- What is rational **does not depend on who I am**.
  - *I don't get to have my own logic.*
  - *In particular, if I view a reason as justifying an action for me, I must view it as justifying the same action* ***for anyone to whom the reason applies****.*
- The assumption underlies science and all forms of rational inquiry.
  - *Ethics assumes nothing more.*

Principles

- We sketch **deontological arguments** for three ethical principles.
    - Based on assumptions just stated.
    - ***Generalization principle***
    - ***Autonomy principle***
    - ***Utilitarian principle***

Generalization Principle

- **Example**
- Suppose I steal a watch from a shop.
- I have 2 reasons:
  - *I want a new watch.*
  - *I won't get caught.*
    - Security at the shop is lax.

Generalization Principle

- **Example**
- Suppose I steal a watch from a shop.
- I have 2 reasons:
  - *I want a new watch.*
  - *I won't get caught.*
    - Security at the shop is lax.

- These are not psychological causes or motivations.
  - *They are consciously adduced reasons for the theft.*
    - There may be other reasons, of course.

Example - Theft

- Due to universality of reason, I am making a decision for everyone:
  - *All who want a watch and think they won't get caught should steal one.*

Example - Theft

- Due to universality of reason, I am making a decision for everyone:
  - *All who want a watch and think they won't get caught should steal one.*
- But I know that if all do this, they will get caught.
  - *The shop will install security.*
  - *My reasons will no longer apply to **me**.*
- I am not saying that all these people actually **will** steal watches.
  - *Only that if they did, my reasons would no longer apply*.

Example - Theft

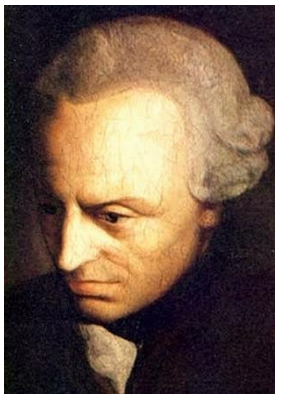
- My reasons are **inconsistent** with the assumption that people will act on them.
- I am caught in a contradiction.
  - *I am deciding that these reasons justify theft for **me**.*
  - *But I am **not** deciding that these reasons justify theft for **others**.*
  - *I can't have it both ways.*

Example - Theft

- My reasons are **inconsistent** with the assumption that people will act on them.
- I am caught in a contradiction.
  - *I am deciding that these reasons justify theft for **me**.*
  - *But I am **not** deciding that these reasons justify theft for **others**.*
  - *I can't have it both ways.*
- More generally…
  - *Universal theft merely for personal benefit would **undermine the institution of property**.*
    - Purpose of theft is to benefit from property rights.

Generalization Principle

- It should be **rational** for me to believe that the **reasons** for my action are **consistent** with the assumption that **everyone to whom the same reasons apply acts the same way**.
  - *Historically inspired by Kant's Categorical Imperative, but different and more precise.*
  - *Takes "rationality" as a primitive and unexplained notion, but this is true to some extent of all science.*



Example - Cheating

- What is wrong with cheating on an exam?
- My reasons:
  - *I will get a better grade and therefore a better job.*
  - *I can get away with it.*
- I know that these reasons apply to nearly all students.
  - *If they act accordingly, grades will be meaningless, or exams strictly proctored.*
  - *This defeats one or both of my reasons.*
  - *So, cheating for these reasons **violates** the generalization principle.*

Example - Agreements

- Breaking an agreement normally violates the generalization principle.
- Reason:
  - *Convenience or profit.*
- These reasons apply to most agreements
  - *If agreements were broken for mere convenience, it would be impossible to **make** agreements.*
  - *And therefore impossible to **achieve one's purposes** by **breaking** them.*
  - *The whole point of having an agreement is that you keep it when **you don't want to keep it**.*

Example - Lying

- Lying for mere convenience violates the generalization principle.
  - *…if the reason for lying assumes that people will believe the lie.*
  - *If everyone lied when convenient, no one would believe the lies.*
    - The possibility of **communication** presupposes a certain amount of credibility.

Example - Lying

- Lying can be generalizable, depending on the reasons.
- Popular “counterexample”
    - Similar to one posed in Kant’s day.
    - *Workers in an Amsterdam office building lied to Nazi police, to conceal whereabouts of Anne Frank and family.*
    - *This is **generalizable**.*
        - If everyone lied for this reason, it would still accomplish the purpose, perhaps even more effectively.
        - There is no need for police to believe the lies.

Scope of the Rationale

- **Scope** = an agent's necessary and jointly sufficient conditions for performing an act.
  - *An ambulance driver uses the siren, but with no patient.*
  - *His reasons:*
    - He is late picking up his kids at day care, because he misplaced his car keys.
    - The siren will allow him to arrive on time.
    - He can get away with it.
  - *This is **generalizable***
    - These reasons seldom apply to an ambulance driver.
  - *But the scope is **too narrow***
    - The details are not necessary.
    - The real reason is that it is important to be on time.

Action Plans

- Since actions always have a rationale, we treat them as **action plans**.
  - *If X, then do Y.*
  - *For example,*
    - **If** I would like to have an item on display in a shop, **and** I can get away with stealing it, **then** I will steal it.

Action Plans

- Since actions always have a rationale, we treat them as **action plans**.
  - *If X, then do Y.*
  - *For example,*
    - **If** I would like to have an item on display in a shop, **and** I can get away with stealing it, **then** I will steal it.
- An **agent** is a bundle of action plans.
  - *…that are executed when the antecedents are satisfied.*
  - *This is not intended as a model of **human psychology**.*
  - *It is a model of **agency**.*

Autonomy

- There is a fundamental obligation to respect **autonomy**.
  - *This rules out murder, most coercion, slavery, etc.*
  - *But autonomy must be carefully defined.*

Autonomy

- There is a fundamental obligation to respect **autonomy**.
  - *This rules out murder, most coercion, slavery, etc.*
  - *But autonomy must be carefully defined.*
- Autonomy is more than "self-law."
  - *I act **autonomously** when I freely make up my own mind about what to do, based on **coherent reasons** I give for my decision*
    - An **agent** is a being that can act autonomously (sometimes called a "moral agent").
    - Today's "autonomous cars" are **not** autonomous.

Autonomy Principle

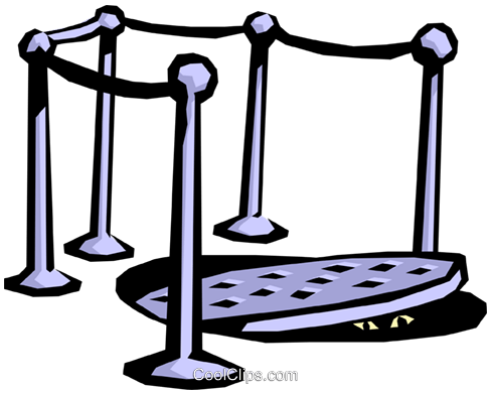
- My action plan is unethical if I am **rationally constrained to believe it interferes** with the **ethical action plan** of some other agent.

Autonomy Principle

- I must be **rationally constrained** to believe there is a conflict of action plans.
  - *That is, it is **irrational** not to believe this.*
    - If someone falls into a maintenance hole I leave uncovered, this is **not** a violation of autonomy.
    - It is only possible/probable that someone will fall in (a violation of the **utilitarian principle**).

Autonomy Principle

- I must be **rationally constrained** to believe there is a conflict of action plans.
  - *That is, it is **irrational** not to believe this.*
    - If someone falls into a maintenance hole I leave uncovered, this is **not** a violation of autonomy.
    - It is only possible/probable that someone will fall in (a violation of the **utilitarian principle**).
    - But suppose it has a cover that will **collapse** when someone steps on it and is on 5th Ave NYC.
    - I am **rationally constrained** to believe **someone** will fall in.
    - I **violate autonomy**.



Autonomy Principle

- Interference with an **unethical** action plan is **not** a violation of autonomy.
  - *An unethical action plan is not a freely chosen action, because it has no coherent rationale.*
  - *There is **no denial of agency**.*
    - You can defend yourself, because an attack on you is unethical.

Autonomy Principle

- Interference with an **unethical** action plan is **not** a violation of autonomy.
  - *An unethical action plan is not a freely chosen action, because it has no coherent rationale.*
  - *There is **no denial of agency**.*
    - You can defend yourself, because an attack on you is unethical.
  - *Is this a circular reference to "unethical"?*
    - We define "unethical" **recursively**.
    - The recursion **begins** with the **generalization** and **utilitarian** principles.
    - An action plan is unethical if it violates the generalization or utilitarian principle, **or** interferes with an ethical action plan.

Autonomy Principle

- Coercion with **informed consent** is **not** a violation of autonomy.
  - *An auto manufacturer is **rationally constrained to believe** that some people will be killed or seriously injured in its cars.*
    - This is coercion: it **compels** some customers to be dead or incapacitated.
  - *It is **no violation of autonomy***
    - Drivers and passengers **give informed consent** to the risk.
    - Their action plan is actually, “If I want to travel to point X, and I am not the victim of an accident, then I will travel there by car.”
    - We **do** have violation if there is a **hazardous defect** in the car known to the manufacturer but not the customer.

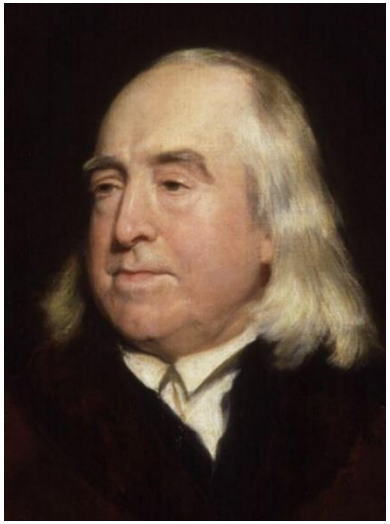
Autonomy Principle

- Coercion with **informed consent** is **not** a violation of autonomy.
  - *An auto manufacturer is **rationally constrained to believe** that some people will be killed or seriously injured in its cars.*
    - This is coercion: it **compels** some customers to be dead or incapacitated.
  - *It is **no violation of autonomy***
    - Drivers and passengers **give informed consent** to the risk.
    - Their action plan is actually, “If I want to travel to point X, and I am not the victim of an accident, then I will travel there by car.”
    - We **do** have violation if there is a **hazardous defect** in the car known to the manufacturer but not the customer.
    - How about pedestrians? Maybe they give informed consent to the **risk of walking on a street**.

Autonomy Principle

- Why a strong "rationally constrained" provision?
  - *It is a consequence of the **deontological argument** for the autonomy principle.*
    - Strictly speaking, I adopt an **entire action policy** rather than individual action plans.
    - If I am to be rational, the plans must be **mutually consistent** (same for beliefs in general that I adopt).
    - Inconsistency is a strong condition: I am **rationally constrained** to acknowledge it.
    - The **universality of reason** says that when adopting a policy, I adopt it for **everyone** (Kant says I "legislate").
    - So, the action plans I rationally attribute to **everyone** must be mutually consistent.
    - If I adopt a plan that **conflicts** with the plans I rationally attribute to others, I am **rationally constrained** to acknowledge the inconsistency.
    - My policy is **irrational** and therefore **unethical**.

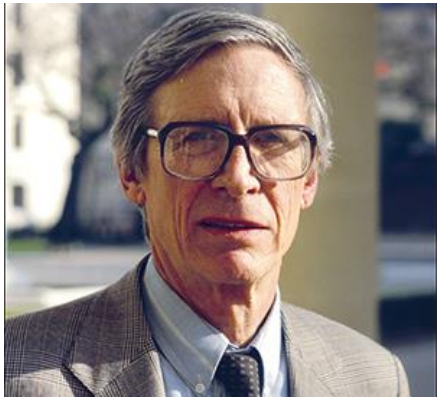
Utilitarian Principle



- This principle asks us to maximize total net expected "utility."
  - *As best we can estimate it.*
  - *"Greatest good for the greatest number," in Jeremy Bentham's formulation.*
  - *Utility = what the agent regards as* ***inherently valuable.***
    - That is, the end(s) to which one's actions are a means.
    - It was happiness/pleasure for classical utilitarians.
    - There must be an **ultimate end** to avoid infinite regress in the rationale for an act.

Utilitarian Principle

- Deontological argument – in brief.
  - *Due to **universality of reason**, if I regard an end as intrinsically valuable, I must regard it as valuable for **anyone**.*
    - It shouldn't matter who I am.
  - *My actions should take everyone else's utility as seriously as my own.*
    - This may not imply strict maximization of net expected utility, but we assume so for now.
    - For example, it may require some degree of distributive justice, as in the difference principle of John Rawls.



Utilitarian Principle

- What about **futility arguments**?
  - *My commanding officer orders me to torture a prisoner.*
    - The results are the same (or worse) if I refuse, as **someone else** will obey the order.
    - This shows that the torture passes the **utilitarian** test.

Abu Ghraib Prison, Iraq

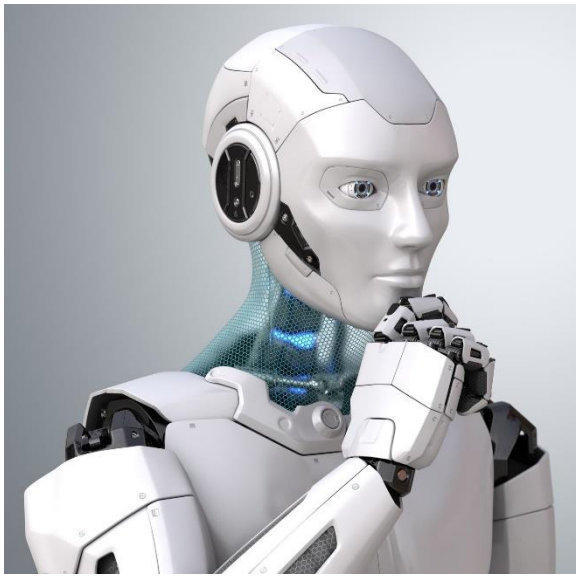
Utilitarian Principle

- What about **futility arguments**?
  - *My commanding officer orders me to torture a prisoner.*
    - The results are the same (or worse) if I refuse, as **someone else** will obey the order.
    - This shows that the torture passes the **utilitarian** test.
  - *Yet it violates the prisoner's **autonomy**.*
    - The willingness of others to do it is irrelevant.
    - What matters is the **incompatibility** of action plans.

Abu Ghraib Prison, Iraq

Machine Ethics

- Nothing in deontological ethics presupposes that agents are **human**.
  - *A reasons-responsive machine can, in principle, be an **autonomous agent**.*
    - It **explains** the rationale for its actions on demand.
    - It doesn't matter if its actions are determined by a **program** (**our** actions are determined).
  - *It can have **obligations** to us, and we to it.*
    - Although **utilitarian** obligations are tricky for machines.
    - Since they are **nonhuman**.

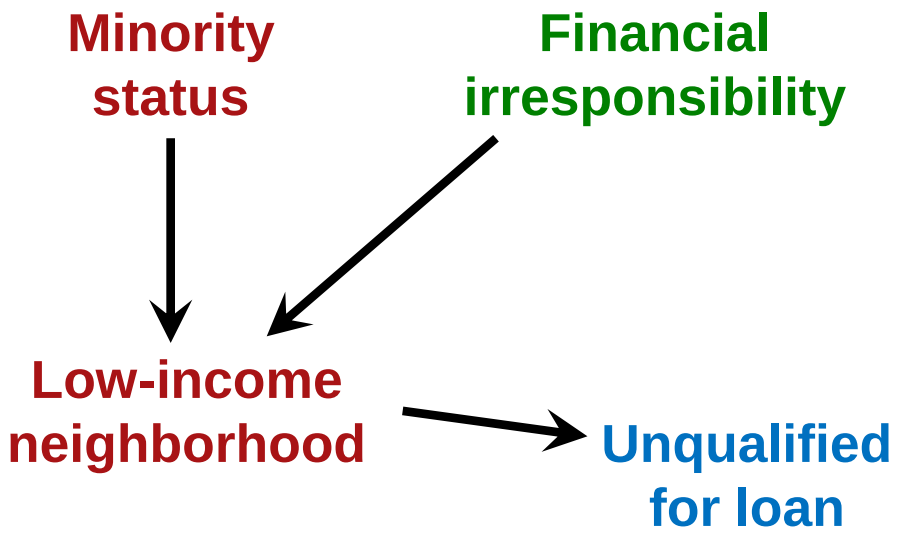


Statistical Fairness Metrics

- Intended to measure bias against a subgroup.
  - *Most are based on statistical measures of classification error.*
  - *Generally based on **yes-no decisions**, not directly on utilitarian consequences.*
    - For example, mortgage loans, university admissions, job interviews, parole decisions.
- Rationale
  - *Group disparities **generally seen as unfair**.*
  - *Bias may incur **legal problems**.*
- Problem
  - *Group parity carries a heavy cultural burden, but it is **fundamentally vague**.*

Example – Mortgage Loans

- Latent bias can occur even when majority/majority status is not a criterion.
  - *Financially irresponsible applicants may live in a **low-income neighborhood**.*
  - *Members of a **minority group** may also live in the neighborhood due to historical discrimination.*
  - *The AI predictor sees the **correlation** between minority status and past defaults.*
  - *Minority applicant is **denied** a mortgage, even if financial irresponsibility is not the cause of past defaults in the minority group.*



Measuring Bias

- Notation
  - ***TP*** = *number of* ***true positives*** *(correct yes's)*
  - ***FP*** = *number of* ***false positives*** *(incorrect yes's)*
  - ***TN*** = *number of* ***true negatives*** *(correct no's)*
  - ***FN*** = *number of* ***false negatives*** *(incorrect no's)*
- Basic model
  - ***Compare various statistics*** *across groups (e.g., majority and a minority group).*

Statistical Fairness Metrics

- **Demographic parity**
  - *Compare* $\frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$ *across groups.*
  - *Rationale?*

    Dwork et al. 2012

    - Compares **fraction** of persons selected in each group. Equality of outcomes.
  - *Possible problems*
    - Ignores efficiency vs correctness issue.
    - Can discriminate against a minority group that is more qualified than majority group.

Statistical Fairness Metrics

- **Equalized odds**

Equality of opportunity

  - *Compare* $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$ *and* $\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$ *across groups.*

  - *Rationale?*
    - Compares fraction of **qualified** (or unqualified) persons selected.

Hardt et al. 2016

  - *Possible problem*
    - Fails to correct for historical injustice that may cause minorities to be less qualified.

Statistical Fairness Metrics

- **Predictive rate parity**
  - *Compare* $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$ *across groups.*
  - *Rationale?*
    - Compares fraction of **selected** individuals that are **in fact qualified**.

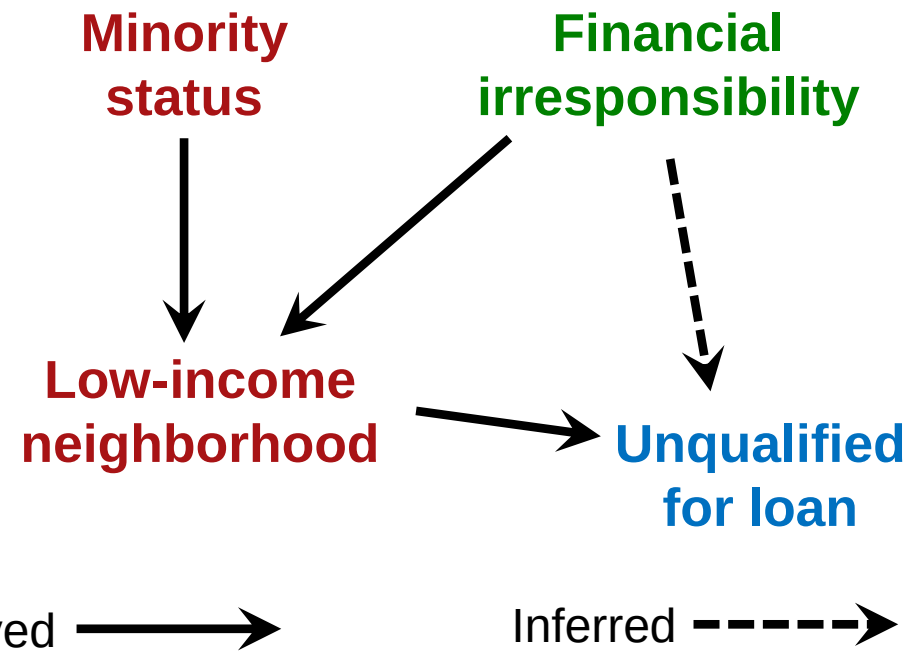
Dieterich et al. 2016

  - *Possible problem*
    - Parity can be achieved when very few minority applicants are selected.

Statistical Fairness Metrics

- **Counterfactual fairness**
  - *Rationale?*
    - Attempts to determine whether minority individuals would be selected if they had been members of the majority.
    - Computes conditional probabilities in **Bayesian (causal) networks** to isolate true cause of past defaults.

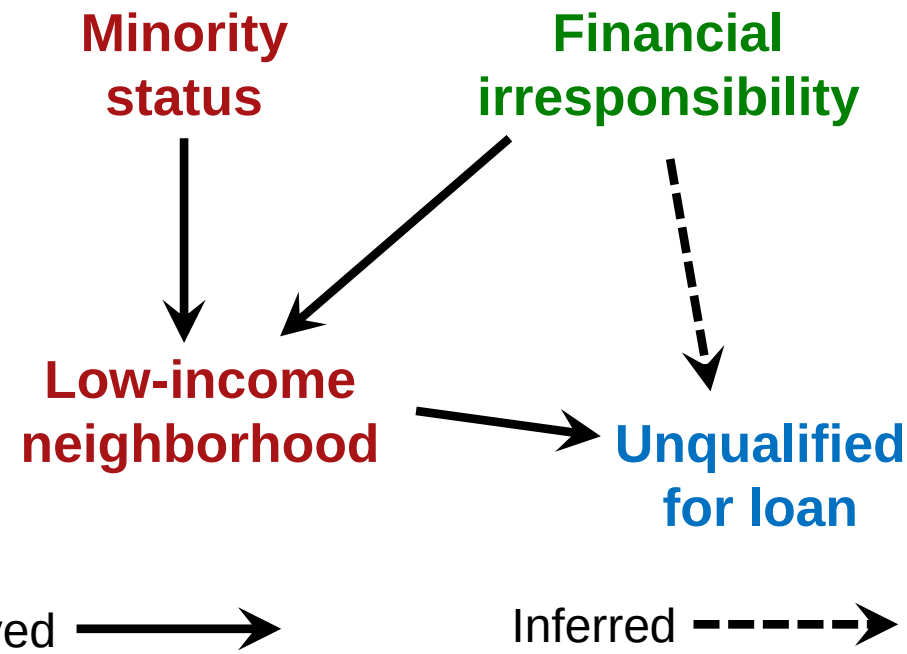
Kusner et al. 2017, Russell et al. 2017



Statistical Fairness Metrics

- **Counterfactual fairness**
  - *Problems*
    - Difficult to identify factors (for inclusion in the network) that correlate with qualification status but do not “cause” them.
    - Even if factors are identified, very rich dataset required to back out conditional probabilities.

Kusner et al. 2017, Russell et al. 2017



Statistical Fairness Metrics

- **General problems** of fairness metrics
  - *Yes-no decisions provide a limited perspective on **utility consequences**.*
  - *There is no consensus on **which bias metric** is suitable for a given context.*
  - *No principle for **balancing fairness and efficiency**.*
  - *No clear principle for **selecting protected groups***
    - Unless one simply selects those protected by law.

Ethical Evaluation of Metrics

- Types of preferential treatment
  - ***Weak***
    - Minority individuals favored only to correct for latent bias against them due to **prediction error**.
    - Results in more **accurate** selection of qualified individuals.
    - But requires explicit consideration of minority status.
  - ***Strong***
    - Minority individuals selected even when **less qualified**.
    - Objective is to correct for **historical bias** that makes minority individuals less likely to be qualified.
    - Again, requires explicit consideration of minority status.

Ethical Evaluation of Metrics

- Types of preferential treatment
  - ***Weak***
    - Minority individuals favored only to correct for latent bias against them due to **prediction error**.
    - Results in more **accurate** selection of qualified individuals.
    - But requires explicit consideration of minority status.
  - ***Strong***
    - Minority individuals selected even when **less qualified**.
    - Objective is to correct for **historical bias** that makes minority individuals less likely to be qualified.
    - Again, requires explicit consideration of minority status.
- Basic ethical question: which (if either) of these is justified?

Ethical Evaluation of Metrics

- **Utilitarian principle** applied to mortgage loans
    - Analysis may differ for other types of decisions!
    - ***Preferential treatment in the weak sense***
        - Results in **greater utility** than no preference, due to greater accuracy.
        - Defaults are bad for everyone.

Ethical Evaluation of Metrics

- **Utilitarian principle** applied to mortgage loans
  - Analysis may differ for other types of decisions!
  - ***Preferential treatment in the weak sense***
    - Results in **greater utility** than no preference, due to greater accuracy.
    - Defaults are bad for everyone.
  - ***Preferential treatment in the strong sense***
    - Possibility of error tends to **reduce utility** due to defaults.
    - However, greater opportunity for minorities may **increase utility**, due to reduced economic inequality in the community, and removal of barriers that tend to make minority individuals less qualified in the future.

Ethical Evaluation of Metrics

- **Utilitarian principle** applied to mortgage loans
  - Analysis may differ for other types of decisions!
  - ***Preferential treatment in the weak sense***
    - Results in **greater utility** than no preference, due to greater accuracy.
    - Defaults are bad for everyone.
  - ***Preferential treatment in the strong sense***
    - Possibility of error tends to **reduce utility** due to defaults.
    - However, greater opportunity for minorities may **increase utility**, due to reduced economic inequality in the community, and removal of barriers that tend to make minority individuals less qualified in the future.
  - *We don't consider options that **violate other ethical principles** (such as generalizability).*

Ethical Evaluation of Metrics

- **Generalization principle** applied to mortgage loans
  - Analysis may differ for other types of decisions!
  - ***Preferential treatment in the weak sense***
    - There is arguably an **implied agreement** that the loan applicant divulges financial information on the understanding that it will serve as the basis for the loan decision.
    - **Explicit consideration of minority status** may violate this agreement.
    - Even if minority status is relevant to achieving accuracy in the aggregate, it is not clearly relevant to judging the financial responsibility of a **particular** majority applicant (or even a minority applicant).

Ethical Evaluation of Metrics

- **Generalization principle** applied to mortgage loans
    - Analysis may differ for other types of decisions!
    - ***Preferential treatment in the strong sense***
        - Arguably a clearer violation of the implied agreement.
        - It is granted from the outset that factors other than financial responsibility are considered.

Ethical Evaluation of Metrics

- **Generalization principle** applied to mortgage loans
  - Analysis may differ for other types of decisions!
  - ***Preferential treatment in the strong sense***
    - Arguably a clearer violation of the implied agreement.
    - It is granted from the outset that factors other than financial responsibility are considered.
  - *Ethical assessment depends on a **determination of fact**.*
    - Can the applicant reasonably assume an agreement that financial responsibility will be the **only** factor in the loan decision?
    - Or just a **major** or **important** factor?

Ethical Evaluation of Metrics

- **Tentative conclusions**
  - ***Preferential treatment in the weak sense***
    - May be **generalizable,** depending on nature of the implied agreement.
    - Creates **greater expected utility** than no minority preference.
    - If generalizable, then **ethically permissible and, in fact, obligatory**, **unless** strong preferential treatment is generalizable and creates even greater expected utility.
    - Consistent with **equalized odds**, **predictive rate parity**, and **counterfactual fairness**.
    - May or may not be consistent with **demographic fairness**.

Ethical Evaluation of Metrics

- **Tentative conclusions**
  - ***Preferential treatment in the strong sense***
    - **Can maximize utility**.
    - **If so,** it is **ethically permissible and, in fact, obligatory**, **unless** it is not generalizable due to violation of implied agreement.
    - Normally **inconsistent with equalized odds**, **predictive rate parity** and **counterfactual fairness**.
    - May or may not be consistent with **demographic fairness**.

Ethical Evaluation of Metrics

- More definitive guidance needed
  - *Need to consider **utilitarian consequences** directly.*
  - *Need to **balance fairness and efficiency** in a principled way.*
  - *Need to solve the problem of **identifying protected groups***
- Classical deontology provides limited guidance
  - ***Contractualism** (Rawls) maximizes minimum utility*
    - Can yield extreme solutions wrt to fairness/efficiency trade-off
  - ***Contractarianism** (Kalai-Smorodinsky, Gautier) maximizes equalized fraction of each stakeholder's maximum possible utility.*
    - Seems suitable only for a bargaining context.

Social Welfare and Group Parity

- One possibility: Use **alpha fairness** as a guide.
  - *Allows adjustment of fairness/efficiency trade-off (α parameter).*
  - *Fairly wide use in practice, especially engineering.*
  - *Some axiomatic justification.*
- What degree of group parity in implied by fairness for a given α*?*
  - *Focus here on **equalized odds** (affirmative action).*

Social Welfare and Group Parity

- Reminder from previous tutorial
  - ***Alpha fairness*** *for a given* α *is achieved by a utility distribution* $(u_1,\ldots,u_n)$ *that maximizes the* ***social welfare function***

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

  subject to resource constraints.
  - ***Utilitarian*** *when* $\alpha = 0$, ***maximin*** *when* $\alpha \to \infty$
  - ***Proportional fairness*** *(Nash bargaining solution) corresponds to* $\alpha = 1$.

Social Welfare and Group Parity

- Two models
  - ***Single policy model***
    - **Does not consider membership** in a protected group.
    - Avoids issue of **which groups** to regard as protected.
    - Does alpha fairness for the population result in some degree of parity **across all groups**?
  - ***Dual policy model***
    - **Considers membership** in a chosen protected group.
    - **What degree of parity** for this group is implied by a given choice of alpha?
    - What value of alpha results precisely in **equalized odds**?

Chen, JH, and Leben 2023

Social Welfare and Group Parity

- Notation for single-policy model

Probability parameters

$P(Y) = Pr(\text{a given individual is qualified to be selected})$

$P(Y|\widehat{Y}) = Pr(\text{qualified}|\text{predicted to be qualified})$

The selection decisions determine

$P(D|\widehat{Y}) = Pr(\text{selected}|\text{predicted to be qualified})$

$P(D|\neg\widehat{Y}) = Pr(\text{selected}|\text{predicted to be unqualified})$

We require $P(D) = P(\widehat{Y})$

Social Welfare and Group Parity

- Notation for single-policy model

Utility parameters

$a_1 + b_1$ = expected utility that results from selecting a qualified individual
b_1 = expected utility that results from rejecting a qualified individual
$a_0 + b_0,\ b_0$ = similarly for an unqualified individual

Utility definitions

$$\hat{a}_1 = a_1 P(Y|\widehat{Y}) + a_0\big(1 - P(Y|\widehat{Y})\big)$$
$$\hat{b}_1 = b_1 P(Y|\widehat{Y}) + b_0\big(1 - P(Y|\widehat{Y})\big)$$

similarly for $\hat{a}_0, \hat{b}_0$

Social Welfare and Group Parity

- Results for single policy model

We first note that equalized odds is achieved for all groups when $P(D|\widehat{Y}) = P(D|\neg\widehat{Y})$, otherwise for none.

Alpha fairness for a given α is achieved when

$$P(D|\widehat{Y}) = \frac{\left(\frac{\hat{a}_1}{\hat{a}_0}\right)^{1/\alpha}\left(\hat{a}_0\frac{P(\widehat{Y})}{1-P(\widehat{Y})} + \hat{b}_0\right) - \hat{b}_1}{\hat{a}_1 + \hat{a}_0\left(\frac{\hat{a}_1}{\hat{a}_0}\right)^{1/\alpha}\frac{P(\widehat{Y})}{1-P(\widehat{Y})}}$$

$$P(D|\neg\widehat{Y}) = \frac{P(\widehat{Y})}{1-P(\widehat{Y})}\left(1 - P(D|\widehat{Y})\right)$$

Alpha fairness results in equalized odds across all groups when

$$\left(\frac{\hat{a}_1}{\hat{a}_0}\right)^{1/\alpha} = \frac{\hat{a}_1 P(\widehat{Y}) + \hat{b}_1}{\hat{a}_0 P(\widehat{Y}) + \hat{b}_0}$$

Proportional fairness ($\alpha = 1$) achieves equalized odds for all groups if $b_1 = b_0 = 0$ (start with zero baseline utility).

Social Welfare and Group Parity

- Results for single policy model
  - *While strict group parity requires ignoring qualifications, a **compromise** between accuracy and fairness is typically sought in practice.*
    - A suitable choice of α **gives some priority to accuracy** while **approximating** equalized odds.

Social Welfare and Group Parity

- Results for single policy model
  - ***Example***.
    - College admissions, with 2 protected groups (low-income and female).

High-income

	$P(Y)$	$P(\widehat{Y})$	$P(Y\|\widehat{Y}), P(Y\|\neg\widehat{Y})$	a_1, b_1	a_0, b_0
Males	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{9}{10}, \frac{1}{10}$	$3, 3$	$2, 2$
Females	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{5}{6}, \frac{1}{6}$	$3, 3$	$2, 2$

Low-income

	$P(Y)$	$P(\widehat{Y})$	$P(Y\|\widehat{Y}), P(Y\|\neg\widehat{Y})$	a_1, b_1	a_0, b_0
Males	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{2}{3}, \frac{1}{6}$	$2, 2$	$3, 1$
Females	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{2}, \frac{3}{22}$	$2, 2$	$3, 1$

Social Welfare and Group Parity

- Results for single policy model
  - ***Example***.
    - College admissions, with 2 protected groups (low-income and female).
    - Setting α **= 0.349** achieves equalized odds of 0.292 for all minority groups.
      - *So equalized odds corresponds to a **rather limited emphasis on fairness**, much less than in proportional fairness.*
    - To compromise between fairness and efficiency:
      - *Setting α **= 0.25** gives **some priority** to apparent qualifications (selection rate 0.382/0.254 for qualified/unqualified).*
      - *while yielding **similar odds ratios** of 0.354/0.330 for men/women and 0.354/0.312 for high/low income.*

Social Welfare and Group Parity

- Notation for dual-policy model

Probability parameters

$P(Y|Z) = Pr(\text{a given minority individual is qualified to be selected})$
$P(Y|\neg Z) = Pr(\text{a given majority individual is qualified to be selected})$
$P(Y|Z,\widehat{Y}) = Pr(\text{qualified}|\text{minority \& predicted to be qualified})$
$P(Y|\neg Z,\widehat{Y}) = Pr(\text{qualified}|\text{majority \& predicted to be qualified})$

The selection decisions determine

$P(D|Z,\neg\widehat{Y})$ and $P(D|\neg Z,\widehat{Y})$

We assume $P(D|Z,\widehat{Y}) = 1$ and $P(D|\neg Z,\neg\widehat{Y})$.
That is, all qualified minority individuals are selected,
and no unqualified majority individuals are selected.

Social Welfare and Group Parity

- Notation for dual-policy model

Utility parameters

a_1^M, b_1^M for qualified majority individuals
a_0^m, b_0^m for unqualified minority individuals
other utilities do not affect the solution

Utility definitions

$\hat{a}_1^M, \hat{b}_1^M, \hat{a}_0^m, \hat{b}_0^m$ analogously

Social Welfare and Group Parity

- Results for dual policy model

Equalized odds can be achieved for a given minority group when $P(\widehat{Y}|\neg Z) \geq P(\widehat{Y}|Z)$

Alpha fairness for a given α is achieved when

$$P(D|\neg Z, \widehat{Y}) = \frac{\left(\frac{\hat{a}_1^M}{\hat{a}_0^m}\right)^{1/\alpha}\left(\hat{a}_0^m \frac{(1-P(Z))P(\widehat{Y}|\neg Z)}{P(Z)(1-P(\widehat{Y}|Z))} + \hat{b}_0^m\right) - \hat{b}_1^M}{\hat{a}_1^M + \hat{a}_0^m\left(\frac{\hat{a}_1^M}{\hat{a}_0^m}\right)^{1/\alpha}\frac{(1-P(Z))P(\widehat{Y}|\neg Z)}{P(Z)(1-P(\widehat{Y}|Z)}}$$

$$P(D|Z, \neg\widehat{Y}) = \frac{(1-P(Z))P(\widehat{Y}|\neg Z)}{P(Z)(1-P(\widehat{Y}|Z)}\left(1 - P(D|\neg Z, \widehat{Y})\right)$$

Alpha fairness results in equalized odds across the two groups when

$$\left(\frac{\hat{a}_1^M}{\hat{a}_0^m}\right)^{1/\alpha} = \frac{\hat{a}_1^M + \hat{b}_1^M - \hat{a}_1^M P(Z)\left(1 - \frac{P(\widehat{Y}|Z)}{P(\widehat{Y}|\neg Z)}\right)}{(1-P(Z))\frac{P(\widehat{Y}|\neg Z) - P(\widehat{Y}|Z)}{1-P(\widehat{Y}|Z)}\hat{a}_0^m + \hat{b}_0^m}$$

Social Welfare and Group Parity

- Results for **predictive rate parity**
  - ***Single policy model***
    - Parity **cannot be achieved for any value of** α.
  - ***Dual policy model***
    - One can correct for a smaller predictive rate in the minority group only by **making the minority group worse off.**
      - *i.e., by reducing the selection probability for minority individuals.*
  - *Conclusion: Predictive rate parity in **unsuitable** as a bias metric.*
    - …based on fairness concepts implicit in alpha fairness.

Beyond Group Parity

- Example: **Self-driving cars**.
  - *Is it ethical to manufacture self-driving cars that will be used on public streets and roads?*

Beyond Group Parity

- Example: **Self-driving cars**.
  - *Is it ethical to manufacture self-driving cars that will be used on public streets and roads?*
  - ***Utilitarian principle***
    - This test is passed if one can rationally believe that self-driving cars are at least as safe **on the average**.

Beyond Group Parity

- Example: **Self-driving cars**.
  - ***Autonomy principle***
    - The manufacturer is **rationally constrained to believe** that some people will be killed or seriously injured by the cars.
    - Question: is there **informed consent**?
    - Probably from **passengers**, who presumably know the car is self-driving.

Beyond Group Parity

- Example: **Self-driving cars**.
  - ***Autonomy principle***
    - The manufacturer is **rationally constrained to believe** that some people will be killed or seriously injured by the cars.
    - Question: is there **informed consent**?
    - Probably from **passengers**, who presumably know the car is self-driving.
    - From **pedestrians**?
    - They may be unaware that a **self-driving car** is nearby. So how can they give informed consent to the risk it poses?

Beyond Group Parity

- Example: **Self-driving cars**.
  - ***Autonomy principle***
    - The manufacturer is **rationally constrained to believe** that some people will be killed or seriously injured by the cars.
    - Question: is there **informed consent**?
    - Probably from **passengers**, who presumably know the car is self-driving.
    - From **pedestrians**?
    - They may be unaware that a **self-driving car** is nearby. So how can they give informed consent to the risk it poses?
    - Perhaps its is enough to give consist to the **level** of risk posed by self-driving cars.
    - If this level is **no greater** than that of ordinary cars (already required by the utilitarian principle), we are OK.

Beyond Group Parity

- **Value alignment**
  - *How does one **teach** ethical values to a machine?*
    - Crowd sourced values are unsatisfactory and risk committing the naturalistic fallacy (e.g., MIT's "Moral Machine").
    - One approach: **rule-based AI** (i.e., "good old-fashioned AI").
    - If-then instructions can be regarded as **action plans**.
    - The action plans in a rule base can be ethically assessed by specializing the ethical principles to each one, to generate **test propositions**.
    - The truth of the test propositions is an **empirical** question.
    - **ML with neural networks** can assess their truth.

Kim, JH, and Donaldson 2021

Beyond Group Parity

- **Value alignment**
  - *Example: Logical formulation of **generalization principle***
    - Consider the action plan

Agent *a*'s reasons

Agent *a* regards *C(a)* as justifying *A(a)*

Agent *a* takes action *A*

$$C(a) \Rightarrow_a A(a)$$

    - The generalization principle is

Possibility predicate

For all agents *x*

Implies

Modal operator (rational belief)

$$\Diamond_a P\Big(\forall x\big(C(x) \rightarrow A(x)\big) \wedge C(a) \wedge A(a)\Big)$$

Agent a can rationally believe that it is possible to take action A when reasons C apply, and when all agents to whom reasons C apply take action A.

Beyond Group Parity

- **Value alignment**
  - *Example: Logical formulation of **generalization principle***
    - **Ambulance example**

$$
\begin{aligned}
C_1(a) &= \text{An ambulance under the control of agent } a \text{ can} \\
&\quad\ \text{reach its destination sooner by using the siren} \\
C_2(a) &= \text{There is an emergency patient in the ambulance.} \\
A(a) &= \text{The ambulance will use the siren.}
\end{aligned}
$$

Consider the action plan: $C_1(a) \Rightarrow_a A(a)$

The generalization principle is

$$\Diamond_a P\Big(\forall x\big(C(x) \rightarrow A(x)\big) \wedge C(a) \wedge A(a)\Big)$$

This generates the test proposition

$$\Diamond_a P\Big(\forall x\big(C_1(x) \rightarrow A(x)\big) \wedge C_1(a) \wedge A(a)\Big)$$

This is empirically **false**, since the agent cannot rationally believe that such general use of sirens would permit an ambulance to arrive sooner with a siren. **Violation.** Remove from rule base.

Beyond Group Parity

- **Value alignment**
  - *Example: Logical formulation of generalization principle*
    - **Ambulance example**

$$
\begin{aligned}
C_1(a) &= \text{An ambulance under the control of agent } a \text{ can} \\
&\quad\ \text{reach its destination sooner by using the siren} \\
C_2(a) &= \text{There is an emergency patient in the ambulance.} \\
A(a) &= \text{The ambulance will use the siren.}
\end{aligned}
$$

Consider the action plan $\big(C_1(a) \wedge C_2(a)\big) \Rightarrow_a A(a)$

The generalization principle is

$$\Diamond_a P\Big(\forall x\big(C(x) \rightarrow A(x)\big) \wedge C(a) \wedge A(a)\Big)$$

This generates the test proposition

$$\Diamond_a P\Big(\forall x\big((C_1(x) \wedge C_2(x)) \rightarrow A(x)\big) \wedge C_1(a) \wedge C_2(a) \wedge A(a)\Big)$$

This is empirically **true**, since evidence shows that ambulances can arrive sooner with a siren when it is always used for emergency transport. **No violation.** Keep in rule base.

Beyond Group Parity

- **Value alignment**
  - *Ultimately, one can build **truly autonomous machines**.*
    - Autonomous agents are **necessarily ethical**.
    - They can provide a coherent (and therefore ethical) rationale for all action plans.
    - In particular, **they won't take over** and enslave humans, because this violates the autonomy principle.

Questions? Comments?