Bias in AI Systems

Module 8 of a course on *Ethical Issues in AI*

Prepared by

John Hooker

Emeritus Professor, Carnegie Mellon University

Chautauqua, June 2024

The ethical problem

- Do AI-based decisions **treat groups equally** in a morally relevant sense?
    - *Groups may be based on race, ethnic background, gender, economic status, etc.*
- How should we **measure** group parity for purposes of ethics?

Example: Mortgage decisions

- An application may be rejected, despite sound finances, because…
  - *The applicant belongs to a **minority group**.*
  - *The **default rate** is higher for the minority group.*

Example: Mortgage decisions

- An application may be rejected, despite sound finances, because…
  - *The applicant belongs to a **minority group**.*
  - *The **default rate** is higher for the minority group.*
- **Remove** race/ethnic group from data?

Example: Mortgage decisions

- An application may be rejected, despite sound finances, because…
  - *The applicant belongs to a **minority group**.*
  - *The **default rate** is higher for the minority group.*
- **Remove** race/ethnic group from data?
  - *That may not work.*
  - *There may be **latent bias** even in sanitized data.*

Example: Mortgage decisions

- Why latent bias?
  - *The applicant may be **rejected** due to having an address in a **low-income neighborhood**, where people have a higher default rate.*
  - *Members of minority group are more likely to live in a low-income neighborhood due to **historical discrimination**.*
  - *Their address nonetheless **correlates** a higher default rate.*

Other examples

- Parole (minimize recidivism risk)
- Interviewing and hiring
- College admissions

What to do about it?

- **Option 1**: Get rid of AI.
  - *Even though this **reduces prediction accuracy**.*
  - *Fails **utilitarian** principle, unless using AI is **not generalizable**.*
    - There is arguably an **implicit agreement** with applicant to use only financial criteria.
    - Violating this agreement is **not generalizable**.

What to do about it?

- **Option 2**: Improve AI to satisfy the implicit agreement.
  - *Apply **statistical bias** metrics.*
  - ***Adjust** AI predictions to **get rid** of bias. This requires **explicitly considering** minority status in the decision.*
  - *A popular approach, incentivized by equal opportunity laws.*
    - Scheduled classes (India)
    - Bumiputera quotas (Malaysia)
    - Fair Housing Act (US), e.g.

EQUAL HOUSING
OPPORTUNITY

Bias metrics

- Bias metrics are ways of measuring whether two groups are treated equally.
  - *For short, we refer to these groups as the **majority** and **minority** (= protected group).*
- Most popular metrics:
  - ***Demographic parity***
  - ***Equalized odds***
    - We focus on equality of opportunity
  - ***Predictive rate parity***

Bias metrics

- Bias metrics are ways of measuring whether two groups are treated equally.
  - *For short, we refer to these groups as the **majority** and **minority** (= protected group).*
- Most popular metrics:
  - ***Demographic parity***
  - ***Equalized odds***
    - We focus on equality of opportunity
  - ***Predictive rate parity***
- These are usually **incompatible**.
  - *Must choose one or none!*

Assessing bias metrics

- **“Fair”** treatment of groups seems an intuitively compelling idea.
- But there are **problems**
  - *“Group parity” has **dozens** of mathematical definitions.*
  - *Fairness is itself a **notoriously vague** concept.*
    - What seems **fair** to one person seems **unfair** to another.

Assessing bias metrics

- **“Fair”** treatment of groups seems an intuitively compelling idea.
- But there are **problems**
  - *“Group parity” has **dozens** of mathematical definitions.*
  - *Fairness is itself a **notoriously vague** concept.*
    - What seems **fair** to one person seems **unfair** to another.
- We assess parity metrics directly with **ethical principles**.
  - *Rather than trying to guess which one measures “fairness.”*

Assessing bias metrics

- **Demographic parity**.
  - *Definition:*

    % of majority group **accepted** = % of minority group **accepted**

    - Probability of **accepting** a given person (e.g., for loan) is the same for the two groups**.**
  - *Characteristics:*
    - Compensates for **historical discrimination** that makes a minority person less likely to be qualified.
    - But rules out selecting a greater fraction of minority persons when they are **more qualified** than average (as in Malaysia).

Assessing bias metrics

- **Demographic parity**.
  - *Definition:*

    % of majority group **accepted** = % of minority group **accepted**

    - Probability of **accepting** a given person (e.g., for loan) is the same for the two groups**.**
  - *Ethical assessment:*
    - May violate **generalizability** by overriding evident qualifications.
    - May maximize long-term **utility** by providing equal opportunity to marginalized groups.
    - May reduce long-term **utility** if there is backlash from the majority.

Assessing bias metrics

- **Equalized odds**.
  - *Definition:*

% of **qualified** majority accepted	=	% of **qualified** minority accepted

    - Probability of accepting a **qualified** person (e.g., for loan) is the same for the two groups**.**
  - *Characteristics:*
    - Can allow **few** minority persons to be accepted if **relatively few are qualified** due to social and historical factors.
    - But allows selecting a greater fraction of minority persons when they are **more qualified** than average.

Assessing bias metrics

- **Equalized odds.**
  - *Definition:*

    % of **qualified** majority accepted = % of **qualified** minority accepted

    - Probability of accepting a **qualified** person (e.g., for loan) is the same for the two groups**.**
  - *Ethical assessment:*
    - Consistent with any implied **agreement** to consider only evident qualifications.
    - May maximize long-term **utility** by avoiding backlash.
    - May reduce long-term **utility** by failing to address chronic discrimination.

Assessing bias metrics

- **Predictive rate parity**.
  - *Definition:*

% of **accepted** majority persons who are **qualified**	=	% of **accepted** minority persons who are **qualified**

    - Probability that an **accepted** person is **qualified** (e.g., for loan) is the same for the two groups**.**
  - *Characteristics:*
    - Avoids appearance that **acceptance standards** are different for the two groups.
    - Can allow **few** minority persons to be selected if they are as **qualified as accepted majority persons**.

Assessing bias metrics

- **Predictive rate parity**.
  - *Definition:*

% of **accepted** majority persons who are **qualified**	=	% of **accepted** minority persons who are **qualified**

    - Probability that an **accepted** person is **qualified** (e.g., for loan) is the same for the two groups**.**
  - *Ethical assessment:*
    - May violate **generalizability** by overriding evident qualifications.
    - May maximize long-term **utility** by avoiding backlash.
    - May reduce long-term **utility** by failing to address chronic discrimination.

Assessing bias metrics

- **Highly publicized example: Parole**
  - *COMPAS predictions achieve **predictive rate parity**.*
    - Minority parolees have **same recidivism rate** as majority parolees.
  - *But they **do not equalize odds**.*
    - Apparently qualified minority candidates are about **40% less likely to be paroled** than qualified majority candidates.

From: *Pro Publica*, 23 May 2016

Assessing bias metrics

- **Highly publicized example: Parole**
  - *COMPAS predictions achieve **predictive rate parity**.*
    - Minority parolees have **same recidivism rate** as majority parolees.
  - *But they **do not equalize odds**.*
    - Apparently qualified minority candidates are about **40% less likely to be paroled** than qualified majority candidates.
  - Debate **still unresolved.**

From: *Pro Publica*, 23 May 2016

Assessing bias metrics

- **Counterfactual fairness.**
  - *Definition:*

% of minority persons accepted in the **actual world**	=	% of minority persons accepted in an **alternate world** where they belong to the **majority**

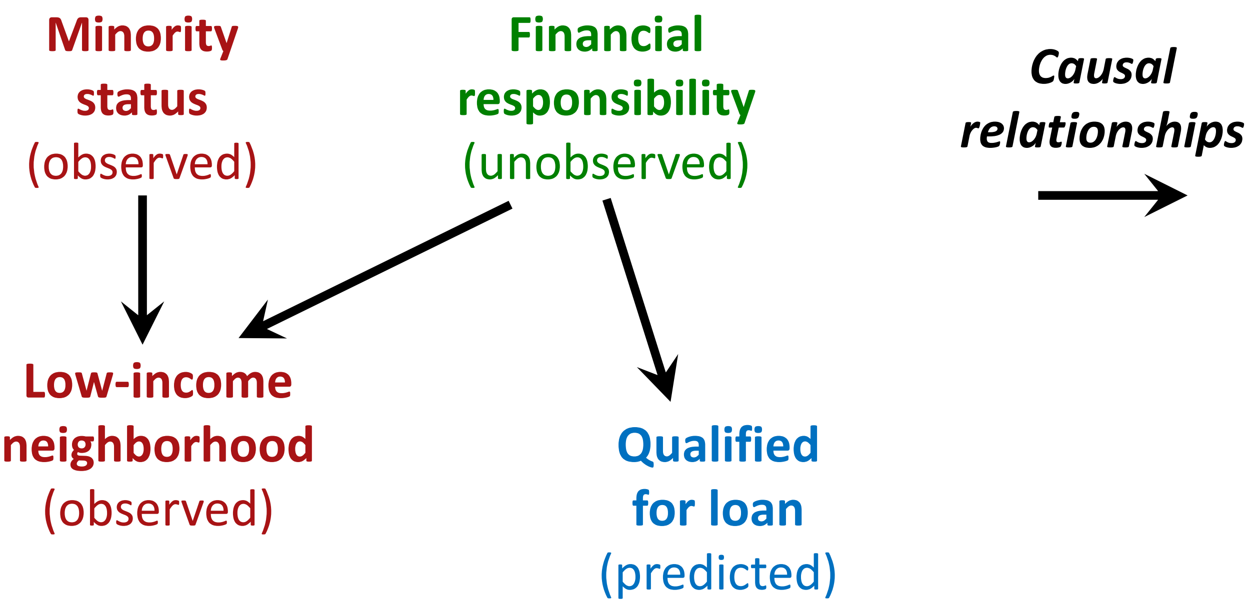
    - Acceptance probability of a given minority person **would have been the same** if that person belonged to the majority**.**
  - *Characteristics:*
    - Sounds **great**.
    - But how to **assess** this?

Assessing bias metrics

- **Counterfactual fairness**.
  - In the case of mortgage loans:
    - *Relevant factor is **financial responsibility**, but only minority status and address can be **observed**.*
    - *Acceptance decision must be the same if it were based **only on financial responsibility**.*
  - Represent this situation with a **causal network:**

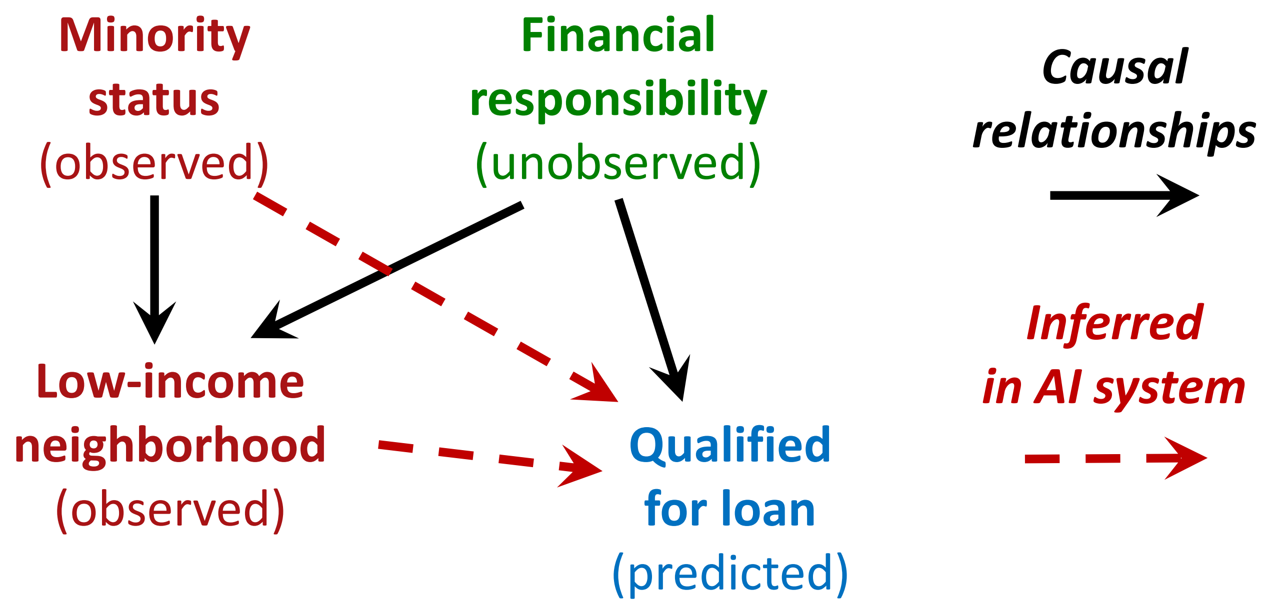
Bias metrics

- **Counterfactual fairness**



Bias metrics

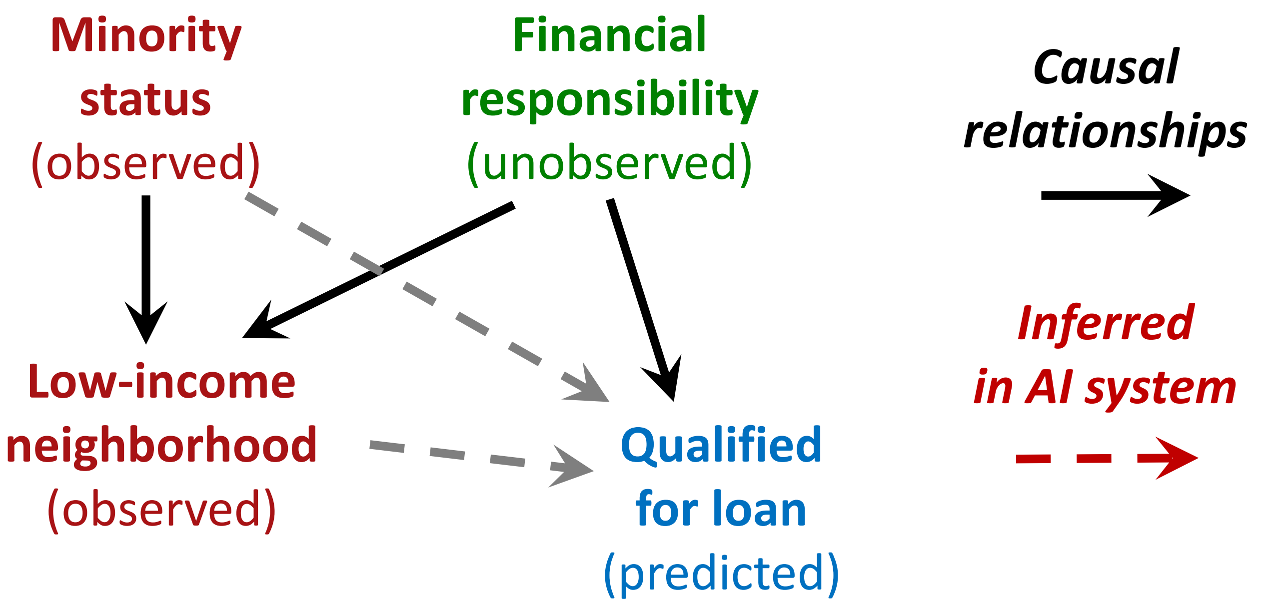
- **Counterfactual fairness**



Bias metrics

- **Counterfactual fairness**

Must use **Bayesian inference** to deduce financial responsibility



Bias metrics

- **Counterfactual fairness.**
  - *Problems:*
    - There may be **many** confounding factors in the network.
    - Bayesian inference requires a **rich data set**, usually unavailable.
    - The desired Bayesian calculations are possible only in networks with a **certain kind of structure**.
    - Still a **research area**.

Assessing bias metrics

- **Counterfactual fairness**.
  - *Ethical problem:*
    - Even if counterfactual inference is possible, do we want decisions to rely **solely on financial responsibility**?
    - Ethical arguments are similar to those surrounding **demographic parity**.