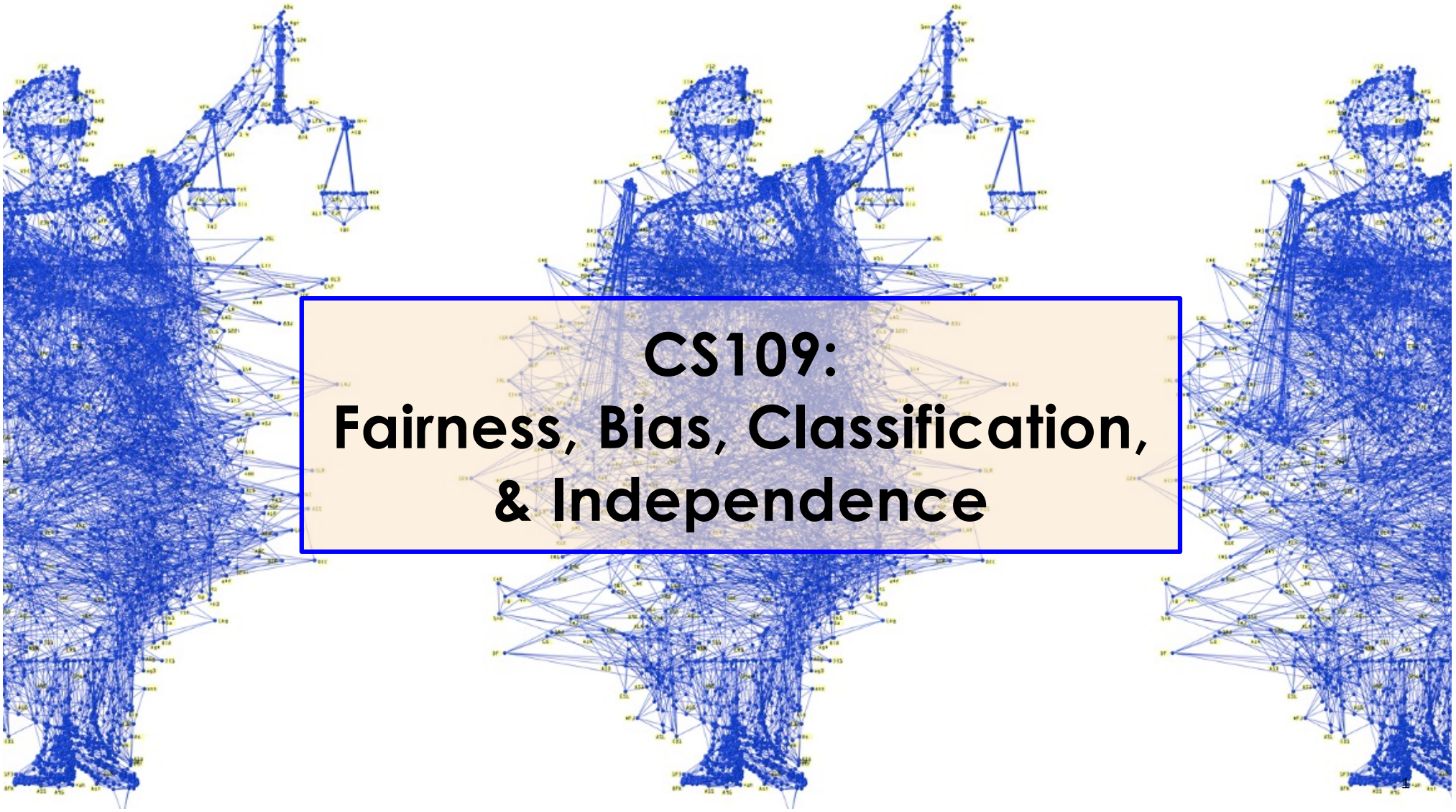


CS109:
Fairness, Bias, Classification, & Independence

Working towards fairness and equality is our ethical responsibility as computer scientists

ACM Code of Ethics and Professional Conduct

ACM Code of Ethics and Professional Conduct

Preamble

Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good. The ACM Code of Ethics and Professional Conduct ("the Code") expresses the conscience of the profession.

The Code is designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. Additionally, the Code serves as a basis for remediation when violations occur. The Code includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. Each principle is supplemented by guidelines, which provide explanations to assist computing professionals in understanding and applying the principle.

Section 1 outlines fundamental ethical principles that form the basis for the remainder of the Code. Section 2 addresses additional, more specific considerations of professional responsibility. Section 3 guides individuals who have a leadership role, whether in the workplace or in a volunteer professional capacity. Commitment to ethical conduct is required of every ACM member, and principles involving compliance with the Code are given in Section 4.

The Code as a whole is concerned with how fundamental ethical principles apply to a computing professional's conduct. The Code is not an algorithm for solving ethical problems; rather it serves as a

On This Page

Preamble

1. GENERAL ETHICAL PRINCIPLES.

1.1 Contribute to society and to human well-being, acknowledging that all peop are stakeholders in computing.

1.2 Avoid harm.

1.3 Be honest and trustworthy.

1.4 Be fair and take action not to discriminate.

1.5 Respect the work required to produ new ideas, inventions, creative works, and computing artifacts.

1.6 Respect privacy.

1.7 Honor confidentiality.

2. PROFESSIONAL RESPONSIBILITIES.

2.1 Strive to achieve high quality in bot the processes and products of

1.2 Avoid harm.

In this document, "harm" means negative consequences, especially when those consequences are significant and unjust. Examples of harm include unjustified physical or mental injury, unjustified destruction or disclosure of information, and unjustified damage to property, reputation, and the environment. This list is not exhaustive.

Well-intended actions, including those that accomplish assigned duties, may lead to harm. When that harm is unintended, those responsible are obliged to undo or mitigate the harm as much as possible. Avoiding harm begins with careful consideration of potential impacts on all those affected by decisions. When harm is an intentional part of the system, those responsible are obligated to ensure that the harm is ethically justified. In either case, ensure that all harm is minimized.

To minimize the possibility of indirectly or unintentionally harming others, computing professionals should follow generally accepted best practices unless there is a compelling ethical reason to do otherwise. Additionally, the consequences of data aggregation and emergent properties of systems should be carefully analyzed. Those involved with pervasive or infrastructure systems should also consider Principle 3.7.

A computing professional has an additional obligation to report any signs of system risks that might result in harm. If leaders do not act to curtail or mitigate such risks, it may be necessary to "blow the whistle" to reduce potential harm. However, capricious or misguided reporting of risks can itself be harmful. Before reporting risks, a computing professional should carefully assess relevant aspects of the situation.

1.2 Avoid harm.

In this document, "harm" means negative consequences, especially when those consequences are significant and unjust. Examples of harm include unjustified physical or mental injury, unjustified destruction or disclosure of information, and unjustified damage to property, reputation, and the environment. This list is not exhaustive.

Well-intended actions, including those that accomplish assigned duties, may lead to harm. When that harm is unintended, those responsible are obliged to undo or mitigate the harm as much as possible. Avoiding harm begins with careful consideration of potential impacts on all those affected by decisions. When harm is an intentional part of the system, those responsible are obligated to ensure that the harm is ethically justified. In either case, ensure that all harm is minimized.

To minimize the possibility of indirectly or unintentionally harming others, computing professionals should follow generally accepted best practices unless there is a compelling ethical reason to do otherwise. Additionally, the consequences of data aggregation and emergent properties of systems should be carefully analyzed. Those involved with pervasive or infrastructure systems should also consider Principle 3.7.

A computing professional has an additional obligation to report any signs of system risks that might result in harm. If leaders do not act to curtail or mitigate such risks, it may be necessary to "blow the whistle" to reduce potential harm. However, capricious or misguided reporting of risks can itself be harmful. Before reporting risks, a computing professional should carefully assess relevant aspects of the situation.

1.4 Be fair and take action not to discriminate.

The values of equality, tolerance, respect for others, and justice govern this principle. Fairness requires that even careful decision processes provide some avenue for redress of grievances.

Computing professionals should foster fair participation of all people, including those of underrepresented groups. Prejudicial discrimination on the basis of age, color, disability, ethnicity, family status, gender identity, labor union membership, military status, nationality, race, religion or belief, sex, sexual orientation, or any other inappropriate factor is an explicit violation of the Code. Harassment, including sexual harassment, bullying, and other abuses of power and authority, is a form of discrimination that, amongst other harms, limits fair access to the virtual and physical spaces where such harassment takes place.

The use of information and technology may cause new, or enhance existing, inequities. Technologies and practices should be as inclusive and accessible as possible and computing professionals should take action to avoid creating systems or technologies that disenfranchise or oppress people. Failure to design for inclusiveness and accessibility may constitute unfair discrimination.

1.4 Be fair and take action not to discriminate.

The values of equality, tolerance, respect for others, and justice govern this principle. Fairness requires that even careful decision processes provide some avenue for redress of grievances.

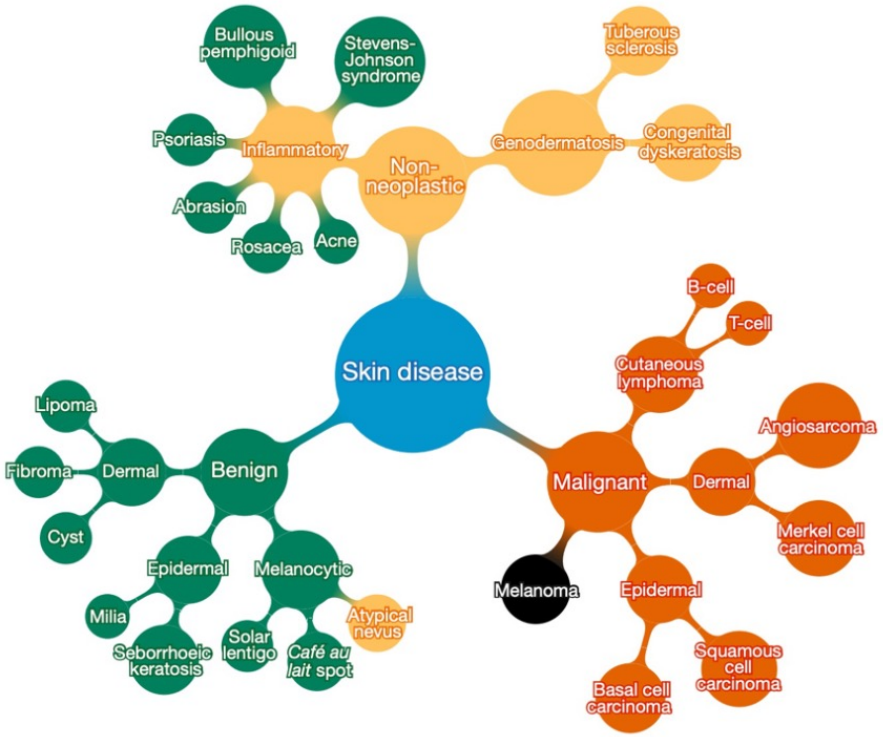
Computing professionals should foster fair participation of all people, including those of underrepresented groups. Prejudicial discrimination on the basis of age, color, disability, ethnicity, family status, gender identity, labor union membership, military status, nationality, race, religion or belief, sex, sexual orientation, or any other inappropriate factor is an explicit violation of the Code. Harassment, including sexual harassment, bullying, and other abuses of power and authority, is a form of discrimination that, amongst other harms, limits fair access to the virtual and physical spaces where such harassment takes place.

The use of information and technology may cause new, or enhance existing, inequities. Technologies and practices should be as inclusive and accessible as possible and computing professionals should take action to avoid creating systems or technologies that disenfranchise or oppress people. Failure to design for inclusiveness and accessibility may constitute unfair discrimination.

Fairness in Classification: What are the Stakes?

Chihuahua or muffin?

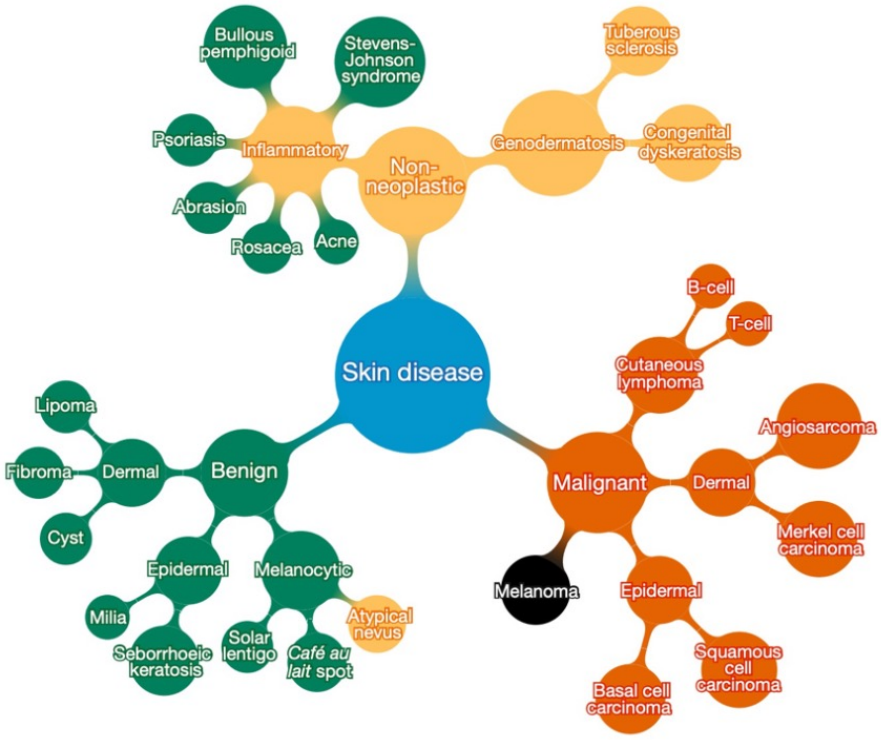
Fairness in Classification: What are the Stakes?



A machine learning algorithm performs **better than** the best dermatologists.

Developed in 2017 at Stanford.

Fairness in Classification



HEALTH

AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind

Machine learning has the potential to save thousands of people from skin cancer each year—while putting others at greater risk.

By Angela Lashbrook

> J Am Acad Dermatol. 2021 Jul 10;S0190-9622(21)02086-7. doi: 10.1016/j.jaad.2021.06.884. Online ahead of print.

Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection–A scoping review

Lisa N Guo [1], Michelle S Lee [1], Bina Kassamali [1], Carol Mita [2], Vinod E Nambudiri [3]

Affiliations + expand

PMID: 34252465 DOI: 10.1016/j.jaad.2021.06.884

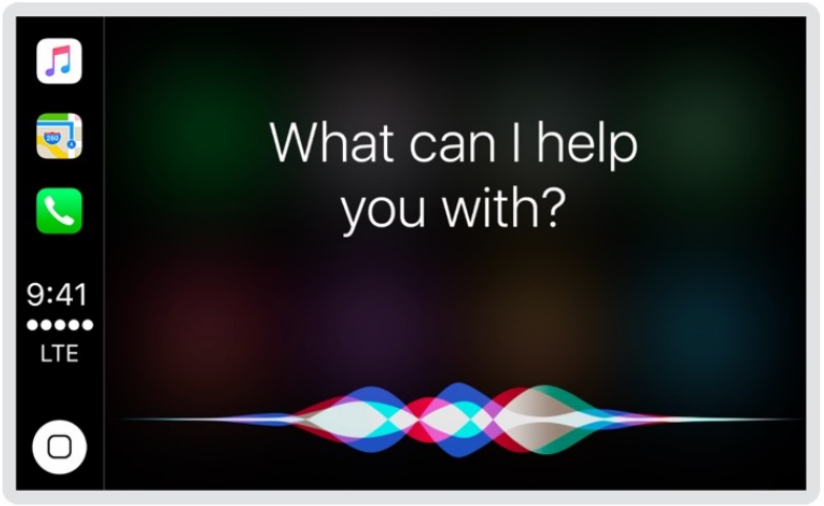
No abstract available

Keywords: artificial intelligence; machine learning; melanoma; racial diversity; skin cancer; skin of color.

“Quality of Service” Harms

“Quality-of-service harms can occur when a system does not work as well for one person as it does for another, even if no opportunities, resources, or information are extended or withheld” (Crawford)

Quality of Service Harms in Voice & Facial Recognition



Fairness … with conditional independence!

Definition of bias in computer systems

Nissenbaum: we will use "bias to refer to computer systems that **systematically and unfairly discriminate** against certain individuals or groups of individuals in favor of others.

A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate"

Allocation Harms

Allocation harms can occur when AI systems unequally extend or withhold opportunities, resources, or information.

Quality of service harms can lead to allocation harms, for example if failure to diagnose a condition leads to lack of medical treatment.

What is a just distribution of outcomes for:

- Hiring
- Lending
- School admissions

Three Formal Definitions of Fairness

Fairness through Unawareness
Fairness through Awareness: Independence
Fairness through Awareness: Separation

Fairness through Unawareness

Motivating idea: “The way to stop discrimination on the basis of race is to stop discriminating on the basis of race” – Chief Justice Roberts

Note: Fairness through unawareness of some federally “protected categories” (subset of sensitive features) is legally required in domains like lending.

How to do it:

1. Exclude the sensitive feature (race, gender, age, etc) from your dataset
2. (Recommended) Also exclude proxies for the sensitive feature (name, zip code)

Case Study: Facebook Ads & Job/Housing Recommendations

Facebook creates "Lookalike" feature for advertisers: upload a "source list" and find users with "common qualities" to target ads, including for housing and jobs

March 2018: National Fair Housing Alliance (NFHA) & other civil rights groups sue Facebook over violations of the Fair Housing Act

March 2019: As part of settlement, Facebook agrees not to use "age, gender, relationship status, religious views, school, political views, interested in, or zip code" in creating lookalike audience

New "Special Ad" Audiences Still Biased

Gender: Equally Biased

Age: Almost as Biased

Race: more difficult to measure given the tools provided but still somewhat biased

Political Views: Less Biased

Sapiezynski et. al 2019,

https://sapiezynski.com/papers/sapiezynski2019algorithms.pdf

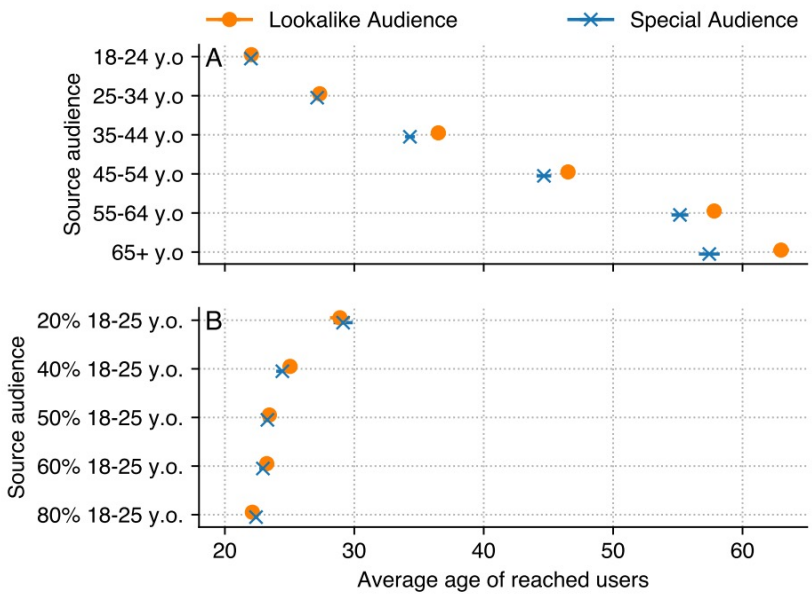
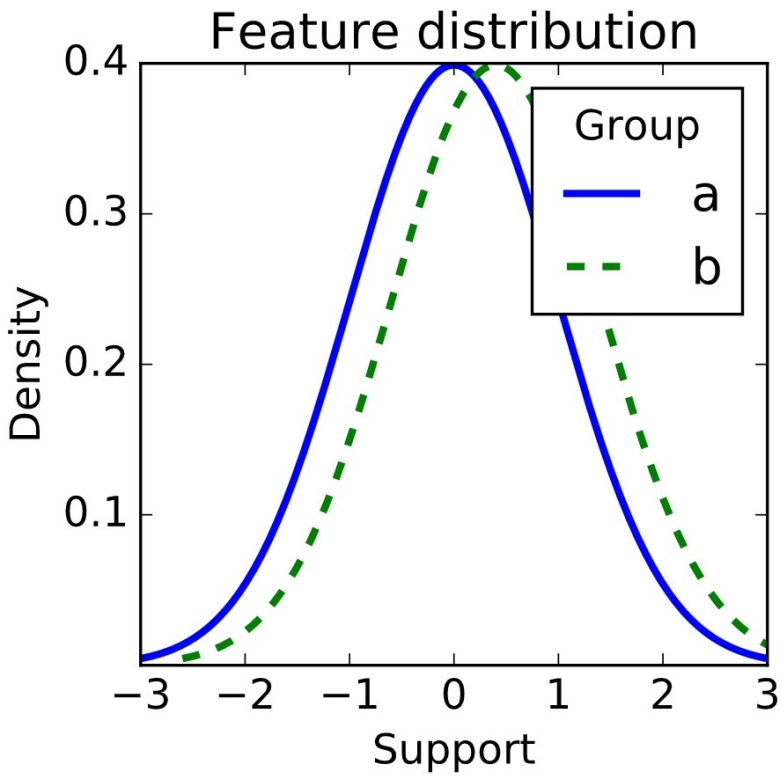
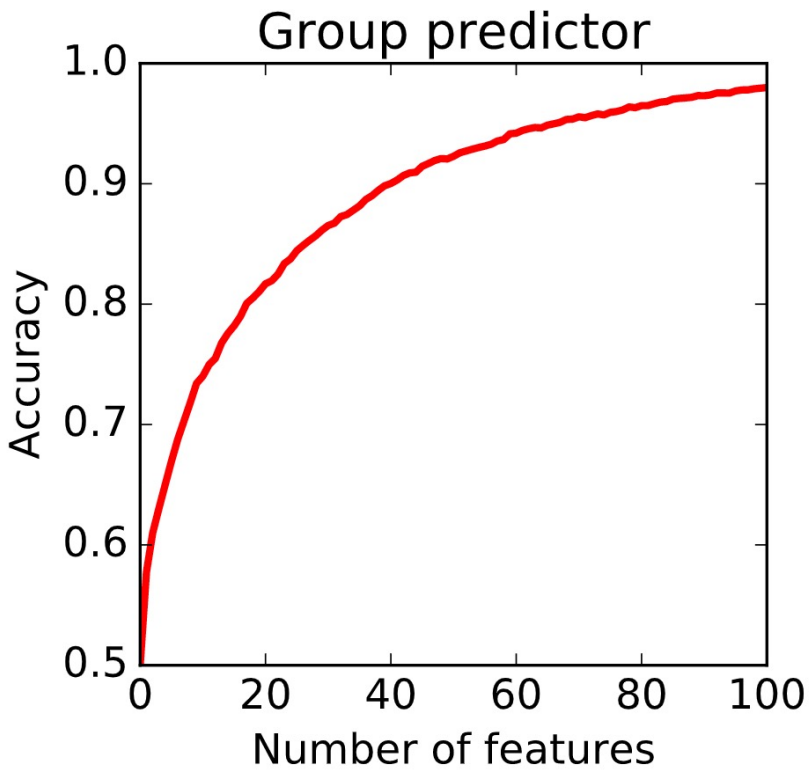


Figure 3: Age breakdown of ad delivery to Lookalike and Special Ad audiences created from the same source audience, using the same ad creative. We can observe extremely similar levels of bias, despite the lack of age as an input to Special Ad audiences. Panel A shows the results for source audiences consisting only of users in one age bracket. Panel B shows the results of mixing the youngest and the oldest users in different proportions.





Many Features = Accurate Group Prediction

Sensitive attributes are often "redundantly encoded" in the dataset

Many of the features or datapoints are correlated with the sensitive attribute

In what way is Fairness through Unawareness Fair?

Procedural Fairness:

Focuses on the decision-making or classification *process*, ensures that the algorithm does not rely on unfair features.

Distributive Fairness:

Focuses on the decision-making or classification *outcome,* ensures that the distribution of good and bad outcomes is equitable.

In what way is "Fairness through Unawareness" Fair?

Procedural Fairness:

Focuses on the decision-making or classification *process*, ensures that the algorithm does not rely on unfair features.

> In our case, Facebook increases procedural fairness by removing "age, gender, relationship status, religious views, school, political views, interested in, zip code" from algorithm that creates Lookalike/SpecialAd audiences.

Distributive Fairness:

Focuses on the decision-making or classification *outcome,* ensures that the distribution of good and bad outcomes is equitable.

> In our case, little increase in distributive fairness because the outcome does not change very much.

Let's Try Fairness Through Awareness!

Awareness of what?

Independence & Demographic Parity

Sensitive Attribute = A

Two groups = a or b

Classifier Outcome or Score = R

The random variables (A, R) satisfy independence for binary classification (which we will study more later!) if:

- $P(R=1|A=a) = P(R=1|A=b)$
- E.g. acceptance rate should be the same for all groups

Relaxed Independence Condition

Another US legal standard is "disparate impact," also known as the 80% rule.

- Imagine people from group A and group B apply to a job.
- The percentage accepted from group B must be at least 80% of the percentage from group A accepted.

$$\frac{\mathbb{P}\{R = 1 \mid A = a\}}{\mathbb{P}\{R = 1 \mid A = b\}} \geq 1 - \epsilon \,.$$ where ε = 0.2.

False Positives and False Negatives

	Condition y = 1	Condition y = 0
Event $\hat{y} = 1$	True positive	False positive
Event $\hat{y} = 0$	False Positive	False Negative

This table is sometimes called a "confusion matrix"

False Positives and False Negatives

	Condition y = 1	Condition y = 0
Event $\hat{y} = 1$	True positive	False positive
Event $\hat{y} = 0$	False Positive	False Negative

= CAT! (True positive)

False Positives and False Negatives

	Condition y = 1	Condition y = 0
Event $\hat{y} = 1$	True positive	False positive
Event $\hat{y} = 0$	False Positive	False Negative

= CAT! (False Positive)

Fairness through Separation

Motivating idea: in some cases, a sensitive attribute is correlated with the target. Separation criterion allows correlation between the score and the sensitive attribute to the *extent that it is justified* by the target variable.

Definition: Random variables (R,A,Y) satisfy separation if R⊥A|Y (⊥ is conditionally independence)

Separation means that the true positive and false positive rates for both groups will be equal.

$$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$$
$$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}$$

Thank you!

Office Hours: https://calendly.com/kathleencreel
Email: kcreel@stanford.edu