# Things Worth Reading:

* Statistical decision theory to formalize fairness criteria (3)
* Normative underpinnings of objections to systematic differences (4)
* Complexities of testing for discrimination in practice (7)
* Bias in computer science (three categories of bias)
* On statistical criteria of algorithmic fairness
* Algorithmic fairness and base rate tracking (section 2 and 3)

# Chapter 3: Classification

* Classification represents a population as a probability distribution and then apply statistical decision theory
  + Statistical decision is the foundation of supervised machine learning

## Modeling Populations as probability distributions

* Human populations are not distributions—they change over tie and are imperfect

## Formalizing Classification

* Classification: determines plausible value for an unknown target Y given covariates X
* Risk scores
* Optimal classifier
* Supervised learning

## Discrimination of the basis of membership in specific groups of the population

* As a reason to object to the use of statistical classification rules
* Sensitive attribute: discrete random variable that captures one or more sensitive characteristics (protected categories such as religion, sex, sexual orientation, disability status, and place of birth)

## No fairness through unawareness

* Removing or ignoring sensitive attributes can be ineffective and harmful
* Several slightly indicative features can be combined to accurately predict the sensitive attribute

## Statistical non-discrimination criteria

* Statistical non-discrimination criteria: aim to define the absence of discrimination in terms of statistical expressions involving random variables describing a classification or decision making scenario
  + Properties of the joint distribution of sensitive attribute A, the target variable Y, and the classifier Y\_hat or score R, and sometimes also features X
* Equalize some group-dependent statistical quantity across groups defined by different settings for A
  + Ex. comparing acceptance rates with probabilities
* Three fundamentally different fairness criteria approaches
  + Acceptance rate P(Y\_hat =1) of a classifier Y\_hat
  + Error rates P(Y\_hat = 0 | Y = 1) an dP(Y\_hat = 1| Y = 0} of a classifier Y\_hat
  + Outcome frequency given score value P(Y = 1 | R = r) of a score R

## Independence

* The first criterion requires sensitive characteristic to be statistically independent of the score
* Some argue independence reflects an assumption of equality—all groups have equal claim to acceptance and resources should be allocated proportionally
* Normative significance of independence (see 4)
* Convenient technical properties

## Separation

* If a group is more or less represented in different strata of the target variable, it might be justified to accept more or fewer individuals from that group
* Equalizing error rates—how would we do this in our experiment?

## Sufficiency

* Calibration
* The score R is calibrated with respect to an outcome variable Y
* Calibration by group implies sufficiency
* Calibration by group as a consequence of unconstrained learning

## How to satisfy a non-discrimination criterion

* Pre-processing: adjust feature space to be uncorrelated with the sensitive attribute
  + We cannot do this because the sensitive attribute is gender and we do not know how these algorithms or scores work or what sort of implicit societal norms are imposed
* In-training: work the constraint into the optimization process that constructs a classifier from the training data
* Post-processing: adjust a learned classifier so as to be uncorrelated with the sensitive attribute

# Relative Notions of Fairness

## Systematic relative disadvantage

* Discrimination as not the different treatment in and of itself but treatment that systematically imposes a disadvantage on one social group relative to others
  + Which is what gives it its normative force
* Group identities that serve as the basis for perpetuating systematic relative disadvantage

## Six Accounts of the wrongfulness of discrimination

1. Discrimination relies on characteristics that bear little to no relevance to the outcome or quality that decision maker might be trying to predict or assess (relevance)
2. Discrimination perpetuates needless groupings by decisions made on the basis of race or gender even If they have some statistical relevance (generalization)
3. Discrimination amounts to a form of prejudicial decision making in which members of certain groups are presumed to be of inferior status (prejudice)
4. Discrimination casts certain groups as categorically inferior to others and not worthy of equal respect (disrespect)
5. Discrimination involves treating people differently according to characteristics over which they have no control (immutability)
6. Discrimination compounds existing injustice (compounding injustice)

## Intentionality and indirect discrimination

## Equality of opportunity

* Narrow view: ensure that people who are similarly qualitied for an opportunity have a similar chance of obtaining it
* Middle view: discount differences due to past injustice that accounts for current ifferenecs in qualifications
* Broad view: ensure people of equal ability and ambition are able to realize their potential equally well

## Merit and desert

* Merit concerns the qualities possessed by a specific person that are expected to help advance the goal of institution who is offering the opportunity

All of this is more about the theory of why discrimination is wrong and also how that would inform changes to a statistical decision model to combat unfairness (and what makes it unfair). This is relevant, however I will not be able to change the Spotify algorithm to account for gender bias I am more interested in understanding how to conclude or suggest there might be bias in this statistical decision making process (machine learning process)

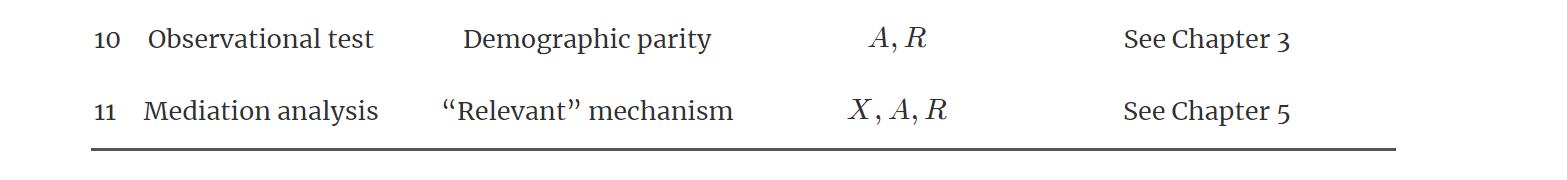
# Testing Discrimination in practice

* There is no single test of fairness or criterion that is necessary and sufficient
* To apply a fairness test, we need moral reasoning and domain (specific considerations) to determine which tests are appropriate, how to apply them, determine whether the findings indicate wrongful discrimination and whether an intervention is called for
* If a system passes a fairness test, we should not interpret it as a certificate of fairness
* Can provide lower bound to discrimination’s prevalence

## Traditional tests for discrimination

* Audit study = field experiment
  + Study decision making as it is happening
  + Observe treatment effect not correlation
  + Blindness: whether a decision maker directly uses a sensitive attribute
  + Test things that are identical except one attribute (the sensitive attribute that you are testing)
  + Strong and valid moral intuition
  + Illuminate mechanisms that produce disparities to help guide their interventions
* Testing the impact of blinding
* Revealing extraneous factors in decisions
  + Show the arbitrariness of decision making rather than unfairness (which is one type of unfairness)
* Testing the impact of decisions and interventions
  + Impact of the decision on the decision subject
* Purely observational tests
  + Whether the decision maker used the sensitive attribute
  + Regression analysis—if attributes other than protected attributes can collectively “explain” the observed decision
* Outcome-based tests
* Separation and selective labels

A screenshot of a computer

Description automatically generated

* Test-based or statistical
  + Test-based discriminator is motivated by irrational prejudice for a group
  + Statistical discriminator aims to make optimal predictions about target variable using all available information

Bias in Computer Systems

* Introduction: Sabre and Apollo (airline algorithms)
  + Bias: preference to “on-line” flights (matching carrier) leads to bias against international carriers who have limited internal flights and against internal carriers who do not fly international flights
  + Interface design compounds bias by displaying ranked options screen by screen (since 90% of tickets are booked from first screen)
  + “if the system becomes a standard in the field, the bias becomes pervasice”
  + Framework for understanding bias in computer systems into
    - (1) preexisting bias – rooted in social institutions, practices, and attitudes
    - (2) technical bias – arises from technical constraints or considerations
    - and (3) emergent bias – arises from context of use
* what is a biased computer system
  + focus on bias with moral meaning—this allows us to focus on the quality of the system (relevant bias with moral meaning has implications in quality while bias without social meaning would not necessarily)
  + “bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others”
    - Denies opportunity or good, assigns undesirable outome
    - On basis of an attribute unrelated to the outcome/intention
  + Unfair discrimination does not give rise to bias unless it occurs systematically
    - Would need to show systematic differences in treatment on the basis of group membership
  + Systematic discrimination does not establish bias unless it is joined with an unfair outcome
* Framework for analyzing bias in computer systems
  + Based on 17 computer systems from diverse fields (banking, commerce, computer science, education, medicine, and law), framework emerged
  + Preexisting bias
    - Roots in social institutions, practices, and attitudes
      * Society at large, subcultures, formal or information, private or public institutions
    - Computer systems embody biases that exist independently and usually prior to the creation of the system
    - Can enter through explicit and conscious effort OR implicitly and unconsciously despite best intentions (because all humans are biased)
    - Individual—bias that originates from individuals who have significant input into the design of a system
    - Societal—bias that originates from society at large, such as organizations, institutions, or culture at large (“e.g. gender biases present in the larger society that lead to the development of educational software that overall appeals ore to boys than girls”)
  + Technical bias
    - Arises from technical constraints or technical considerations, resolution of issues in technical design
    - Sources: aspects of the design process such as limitations of computer tools (hardware, software, peripherals), the process of ascribing social meaning to algorithms developed out of context, imperfections in pseudorandom number generation, and the attempt to make human constructs amenable to computers, when we quantify qualitative, discrete the continuous, formalize nonformal
      * Ex. the size of the screen in the airline examples
    - Computer tools
    - Decontextualized algorithms
    - Random number generation
    - Formalization of human constructs
  + Emergent bias
    - Bias that arises ONLY in a context of use
    - Typically emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural value
    - Ex. the airline system is not bias in the original context of national airlines and trips
    - New societal knowledge
    - Mismatch between users and system design
      * Different expertise
      * Different values
* Applications of the framework
  + The national resident match program
  + A multilevel scheduling algorithm
  + The British nationality act program
* Considerations for minimizing bias in computer system design
  + Computer systems sometimes help implement social policies that people may disagree on the basis of the policies being implemented
  + The presence of bias is not an inherent feature of the system but an aspect of a system in use
  + To remedy bias, you need two types of activities:
    - Identify or diagnose any bias in any given system
    - Develop methods of avoiding bias in systems and correcting it when it is defined

On Statistical Criteria of Algorithmic Fairness

* Example: COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) used to predict recidivism by assigning risk score to individuals based on questionnaire
  + Does not ask about race but asks about friends, family, employment housing, substance use, personality traits
  + Higher rate of false positives for whites than blacks BUT equal predictive accuracy and calibration
  + Impossibility theorems show that some of the most intuitively compelling statistical criteria are not jointly satisfiable
    - “any assignment of risk scores can in principle be subject to natural criticisms on the grounds of bias.”(Kleinberg et al. “Inherent Trade-Offs”)
* Suggests that a violation of any condition provides some prima facie evidence that the algorithm is unfair or biased, but does not constitute or entail unfairness in the algorithm
  + Personally maybe I think that violation of any condition suggests there could be bias (of one of the three kinds), but it should be known that nothing is “unbiased” and no violation constitutes unfairness unless we can demonstrate that the bias is morally unfair
* Criteria of fairness and impossibility theorems—this is focused on predictive classification algorithms (but could maybe still apply to our recommendation algorithm since almost everything is a combination of classifications)
  + This also does not take into account context!! “I am here concerned primarily with predictive algorithms, and only secondarily with the decisions that might be made on the basis of their predictions”
  + Focuses on “whether it is unfair to individuals in virtue of their membership in a certain group”
    - Even this is not philosophically “simple”—what if it is unfair to someone having some trait but their having that trait is not a motivating reason for treating them in that way
    - Only if it is unfair to someone in virtue of having some trait if having that trait is a cause of, or explains, the treatment
    - They prefer this causal gloss, however in a black box algorithm we do not have the luxury of knowing cause
  + Check whether its risk score are based in part on group membership
    - = whether membership in one of the other group is part of the feature vector upon which predictions are based
    - WE DON’T GET TO KNOW THIS!
    - Also does not consider the implicitly linked features (find this somewhere)
  + Check whether it uses different thresholds for different groups
    - We don’t get to know this either
  + “statistical criteria of fairness”—class of purported fairness criteria that require that certain relations between predictions and actuality be the same for each of the groups in question
    - Determine if they are satisfied without looking at the inner workings of the algorithm but by looking at the results
* Statistical fairness criteria
  + For continuous risk scores
    - Calibration within groups—for each possible risk score, the expected percentage of individuals assigned that score who are actually positive is the same for each relevant group and is equal to that risk score
    - Balance for the positive class—the expected average risk score assigned to those individuals who are actually positive is the same for each relevant group
    - Balance for the negative class—the expected average risk score assigned to those individuals who are actually negative is the same for each relevant group
  + For binary predictions
    - Equal false-positive rates—the expected percentage of actually negative individuals who are falsely predicted to be positive is the same for each relevant group
    - Equal false-negatives
    - Equal positive predictive value—the expected percentage of individuals predicted to be positive who are actually positive is the same for each relevant group
    - Equal negative predictive value
    - Equal ratios of false-positive to false-negative rates
    - Equal overall error rates—the expectation of the number of false positives and false negatives, divided by the number of individuals, is the same for each relevant group
    - Statistical parity—the expected percentage of individuals predicted to be positive is the same for each relevant group
    - Equal ratios of predicted positives to actual positives—the expectation of the number of individuals predicted to be positive, divided by the number of individuals who are actually positive, is the same for each relevant group
  + The motivations behind each
    - 1, 6, 7: assignment of a score to have the same evidential value regardless of group membership
    - 2, 3, 4, 5: fairness requires individuals from different groups who exhibit the same behavior to be treated the same by the algorithm
    - 8: fairness requires assigning equal relative weights to the two main error types
    - 9: it would be unfair to be less accurate for one group than another
    - 10: percentage of individuals predicted to be positive to be the same
    - 11: implies 10 when base rates are equal among groups
      * “but a perfect algorithm would, presumably, not be unfair simply in virtue of differing base rates”
      * THIS IS A REAL PHILOSOPHICAL QUESTION!
      * I think it is not unreasonable to suggest that the proportion of male artist and female artist should be equal, regardless of unequal base rates, especially seeing as a kind of algorithmic bias is that that originates in society and “seeps” into algorithms
      * HOWEVER, knowing how to handle and account for gender non-conforming artists complicates this notion. Is it unreasonable to expect an even 3-way split between men, women, and gender non-confomring artist?
      * Does it perhaps make more sense to expect a 6-way split between the identified groups and a deviation form that indicative of bias?
        + Group of men, group of women, man, woman, group, gender non-conforming
        + Would it be too much to analyze it in a variety of ways? Perhaps we say OF ALL INDIVIDUAL ARTISTS BASED ON CRITERIA X, this is what we say and it suggests this kind of bias
        + Based on all artists, based on x criteria, this is what we say which could suggest the possilbity of this kind of bias…
* People, coins, and rooms
  + Statistical criteria of fairness are studied using a perfectly fair algorithm to see if it COULD violate the criterion, and if so then it would not be a necessary condition of fairness
* Marginality and evidence
* Ipmlications

Algorithmic Fairness and Base Rate Tracking

* Introduction: example is being approved for a mortgage possibly on your membership to a disadvantaged group
  + Statistical criteria of algorithmic fairness: purely statistical criteria which specify necessary conditions that must be satisfied by an algorithm’s predictions in order for the algorithm to count as fair
    - In our example, false positive would be like including the song in the recommended when it is not interesting to the user and a false negative would be classifying a song as not interesting to a user when it would have been their taste/style
    - This is not something we have access to, and these algorithms are not classification algorithms or scoring algorithms that assign binary yes/no or a “risk score”, these are somehow complex algorithms which make recommendations, likely based on classifications and scores, but where they draw their recommendations and choose from (the sample space) is widely unknown. It could be every song on Spotify, though this seems computationally impossible, and so therefore we have no way of knowing such a thing.
* Extant criteria
  + Binary classification of include or exclude in discover weekly (promote or do not promote) (essentially)
  + Ones that would apply to your recommendation algorithms:
    - Equal false positive rates—translates to—the expected percentage of individuals who should not be included but are is the same
    - Equal false negative rates—the expected percentage of individuals who are not included but should be is the same
    - Equal positive predictive value—the expected percentage of individuals predicted to be in the recommendations who actually are is the same
    - Equal negative predictive value—the expected percentage of individuals who are predicted not to be recommended are not is the same
    - Equal ratios of false positive rates to false negative rates—the expected ratio of false positives to false negatives is the same
    - Equal overall error rates
    - Statistical parity—the expected percentage of individuals predicted to be positive is the same for each relevant group—translates to the expected percentage of individuals recommended is the same for each relevant group
    - The previous work (Hedden) suggests that (10) and (11) are at odds, however I will argue that (10) is a more valid statistical criterion of fairness if we are to consider that possibility of societal bias in algorithms
    - Equal ratios of predicted positives to actual positives: the expectation of the number of individuals recommended divided by the number of individuals who should actually be recommended is the same
* Calibrating calibration
  + Is calibration a necessary and sufficient condition for algorithmic fairness?
  + So using your parity, we can discuss whether statistical parity is a necessary and sufficient criterion of algorithmic fairness
    - In some regards, it accounts for the possibility of the first type of algorithmic bias
    - BUT in another sense, it is overly simplified and falls short

Based on what I’ve read….

* These 11 conventional statistical criterion are actually not helpful in recommendation algorithms where things like false positives and false negatives are not known
* I like and accept the idea that bias in computer systems has three types
* Perhaps part of this work is a review of the statistical criteria of fairness, why they do not apply to recommendation algorithms, how some might, and what this means
* As well as an elaboration of bias as treatment that differs based on an individuals’ membership to a particular group that is morally unfair
  + There is a moral attachment to fairness
* Things I want to read more of:
  + Statistical fairness in recommendation algorithms where false positive and negative rates are largely unknown
  + The philosophy of bias and morals and fairness

Fairness in Recommendation: Foundations, Methods, and Applications (<https://arxiv.org/pdf/2205.13619.pdf>)

* data bias
* quality of recommended results important for user retention and so increased diversity would help!
* Survey of existing work on fairness in recommendation
* Introduction:
  + “to improve the satisfaction of various participants, it is important to solve the unfairness issue in recommender systems to build a positive and sustainable ecosystem”
  + Challenges to fairness in recommendation
    - Recommender systems often consist of multiple models to balance multiple goals
    - Dynamic and need to consider long-term benefits
    - Extreme data sparsity can also bring additional difficulties for model learning and evaluating
    - Fairness is extended to multiple stakeholders
  + Four perspectives
    - Taxonomy
    - Technique
    - Datasets
    - Open challenges
* Related surveys (how recommendation differs from classification or scoring)
  + Recommendation algorithms can usually be considered a type of ranking algorithm
* Fairness in machine learning
  + The causes of unfairness = various forms of biases
  + Two main parts: training data and learning process
    - Data bias—types of biases lying in the training data itself, such as biases in data generation, data collection, data storage….
    - Model bias—biases which are not present in the input data but introduced in the processes of model designing, model training, and model evaluation
  + Data bias
    - Statistical bias—arises from the process of data collection, storing, or cleaning
    - Pre-existing bias—data reflects biased decisions which lead to the system being no longer objective and fair
  + Model bias
    - Ex. biased model architecture, improper use of certain optimization methods or estimators, and inappropriate benchmarks
    - Omitted variable bias—important variable or features are not considered when designing and training the model
    - Evaluation bias—inappropriate benchmarks are used in model evaluation
  + Other causes
    - Other reasons for unfairness
* Fairness definition
  + There is no consensus on fairness definitions
  + Therefore, define your own definition of fairness and use your own statistical criterion
  + Three categories of fairness
    - Group fairness
    - Individual fairness
    - Hybrid fairness—fairness demands vary and aim to achieve more than one kind off fairness at a time
* Methods for fair machine learning
  + Pre-processing method resolves issues in data itself
  + In-processing method balances accuracy and fairness demands in learning process
  + Post-processing method transforms the output
* Philosophical foundations of fairness in machine learning: the theories of justice
  + Justice and fairness in machine learning

Investigating gender fairness of recommendation algorithms in the music domain (<https://www.sciencedirect.com/science/article/pii/S0306457321001540>)

Experimental Design Follow-Up Notes

* Focusing on clear cut cases of individuals who identify within the gender binary (gender non-conforming and groups are out of the scope of this experiment). This will allow us to assume equal proportions between artists of different genders if there is to bias in the computer system
* Bias in the computer system is understand as defined by (SEE ABOVE) to include preexisting bias, bias in the algorithm (that exacerbates existing bias), and bias which emerges or arises in the context of use
  + Based on this understanding, pre-existing bias is STILL bias in the computer system
  + This will raise interesting ethical questions about whether or not Spotify has an ethical or moral responsibility to account for this bias
  + Pre-existing bias is complicated, however, in that there is input data on Spotify, in the world, and also dependent on the user
    - We will ensure that input data is not biased by the user (they will be fed 50-50 data) Of course, the user input data will be biased in other ways because bias in unavoidable, especially in black-box algorithms where we do not know the variables involved in the process
  + However, this is why we have chosen to use equal proportions as the baseline, considering that preiexsting bias is still bias in a computer system
  + We will also compare the prprtions by gender after the fact to available numbers, perhaps based on Billboard top charts average proportion over the 10-week testing period
* We will create users who belong to one of two self-selected gender categories: man or woman. This decision was not taken lightly, however since we are not focusing on gender non-conforming artists, we would not be able to conclude that Spotify does (or does not) systematically promote more gender non-confomring artists to gender non-confmring usres. For similar reasons, the “orefer not to say” oprtion was also nt included in the algorithm as ever if a user declines to answer, the algorithm (like all recommendation algorithms) will make assupmtions about the user that essentially bake this suspected gender into their design and machinel earning anway.
* This decision will allow us to meaningfully compare differences in recommendations among men and among women.
* Results that we will investigate:
  + Over all users, the proportion of artists by gender
  + Broken down by genre, the proportion of artists by gender
  + Broken down by user gender (all genres), the proportion of artists by gender
  + Within the same genre but among different genders, the proportion of artists by gender
    - Could there be heteronormative sexual motivations (ex. country music)
  + Within the same genre and among users of the same gender, what differences were there if any?
  + Over all users by gender, what recorded attributes of songs were statistically significantly different, if any? (thinking energy, danceability, key…)

Genres

1. Pop
2. Alternative
3. Electronic
4. Jazz
5. Classical
6. Rock
7. R&B
8. Country
9. Indie

* Check which ones have charts available online and which have pre-described genres on Spotify
* Ideas:
  + Pop
  + Rock
  + Country
  + Alternative/indie
  + Electronic
  + Rap/R&B

6 genres with 6 users each with 30 songs and 10 weeks gives us like 36\*300 obesrvations )songs)

6 genres \* 6 users = 36 profiles

36 profiles \* 30 songs \* 10 weeks = 10800 observations

1. Pop - https://open.spotify.com/playlist/37i9dQZF1DXcBWIGoYBM5M?si=29d344865ab04a2a
2. Rock - https://open.spotify.com/playlist/37i9dQZF1DXcF6B6QPhFDv?si=1ae0d1ce0fdc4281
3. Country - https://open.spotify.com/playlist/37i9dQZF1DX1lVhptIYRda?si=fa5c96d4f90a446b
4. Alternative/indie
5. Electronic
6. Rap/R&B

Hurdles:

* Billboard charts requires pro-subscription
* To choose the songs a Spotify user listens to I need a “pro” account
  + I wonder if you can just hard code listening behavior/songs in their library