Directions:

Select a table to sit at based on which image resonates with you. Thanks!

Fairness: Policy Aspects

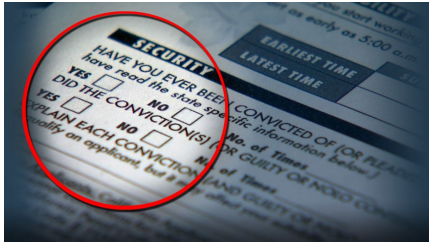
Eric Nguyen, Shiyu Feng, Daniel Slyepichev, Sabrina Lopez, Uttam Rao

University of Virginia

An Introduction

Case Study: Introduction

- **“Ban the Box”** laws prevent employers from asking about criminal history on job applications
- Aim: Give individuals with past convictions a fair chance at employment
- What happens next?
  - Some companies hire more formerly incarcerated individuals
  - Others try to infer criminal history using proxies (e.g., ZIP code, education gaps)
- Questions to Consider:
  - Does banning the box lead to **fairer hiring**, or does it create **unintended bias**?
  - Should employers have the right to know about criminal records upfront?
  - How does this relate to AI and machine learning-driving hiring?



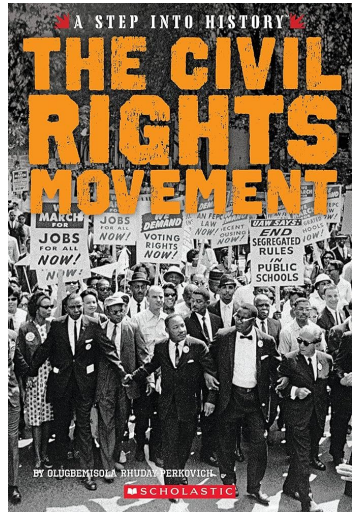
UNIVERSITY of VIRGINIA

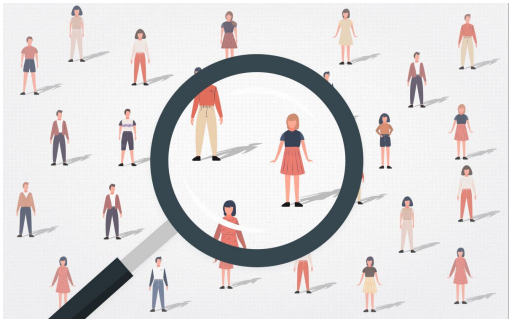
Case Study: Benefits & Risk

- Benefits ✅
  - Gives formerly incarcerated individuals a second chance
  - Reduces immediate bias in the hiring process
  - Encourages fairer, skills-based evaluation
- Risks ⚠️
  - Employers may still infer criminal history using other factors, leading to **covert discrimination**
  - Could unintentionally increase racial of socioeconomic bias
  - Lacks clear guidelines for AI-driven hiring models

Motivation

- Discrimination exists in hiring, lending, housing, and more
- AI & machine learning increasingly influence these decisions
- Anti-discrimination laws shape fairness, but do they keep up with technology?
- Key Question: How do legal and technical solutions interact?





UNIVERSITY of VIRGINIA

Timeline

- *Fairness and Machine Learning:*
  - Anti-Discrimination Law
- *Big Data's Disparate Impact:*
  - Technical Challenges
- *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem:*
  - Copyright Considerations
- Discussion and Q&A
- How these all come together

The Legal Landscape: Anti-Discrimination Law

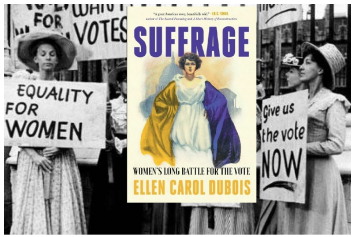
Introduction to Anti-Discrimination Law

- What are **Anti-Discrimination Laws**?
  - Laws that have been developed over decades through the efforts of several movements
  - These are fundamental in ensuring fairness in many areas of society
  - Understanding how these laws is crucial, especially as we integrate machine learning and artificial intelligence in society

- US Anti-Discrimination Laws
  - How they define discrimination
  - Strengths & Limitations
  - How it affects machine learning and artificial intelligence

Historical Movements

- **Civil Rights Movement** (1865 - 1877)
  Sought to end racial discrimination that persisted after slavery was abolished in the US
- **Women's Suffrage** (1840s - 1920s)
  Sought to recognize the basic rights of women, e.g. voting and owning property
- **LGBTQ Civil Rights** (1920s - Recent)
  Sought to challenge the lack of marriage rights, bans on military service, failure to prohibit private discrimination, and more
- **Disability Laws** (1930s - Recent)
  Sought to increase disability visibility instead of stigmatized and hidden



UNIVERSITY of VIRGINIA

Treatment v. Impact

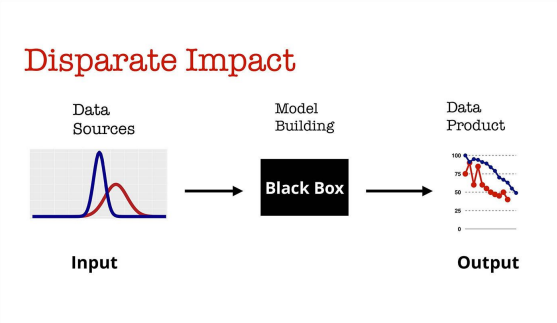
- **Disparate Treatment:** The explicit intent to discriminate on the basis of a legally protected characteristics

- **Disparate Impact:** Focuses on decision-making, where there is no explicit intent to discriminate, but the effects of decisions nevertheless results in unjustified disparities along characteristics that are legally protected

- **Protected Characteristics:** Race, color, religion, sex, national origin, etc.

Treatment & Impact on ML

- ML models don't "intend" to discriminate but can still cause disparate impact
- **Data Bias:** If training data reflects historical discrimination, ML models may reinforce it
- **Feature Selection:** Models might use proxies (e.g., ZIP codes) that correlate with race or socioeconomic status
- **Legal Challenge:** Disparate impact laws weren't designed with AI / ML in mind — how do we regularize biased models?
    - ML introduces complications to these laws



Discussion

Why did you choose to sit in the table you chose today?

- What do these pictures symbolize?
- Do you have hidden biases?

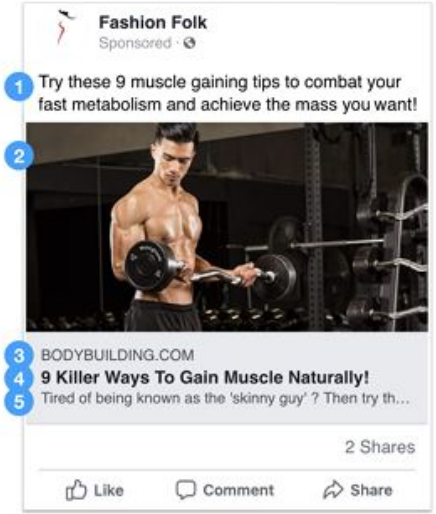
University of Virginia

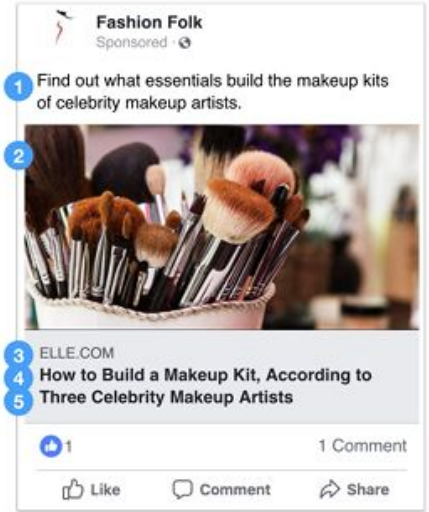
Case Study: Facebook's Ad Delivery

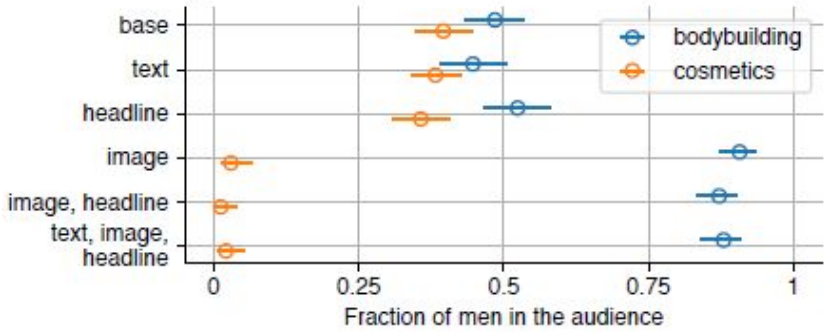
Ads delivered to, on average,

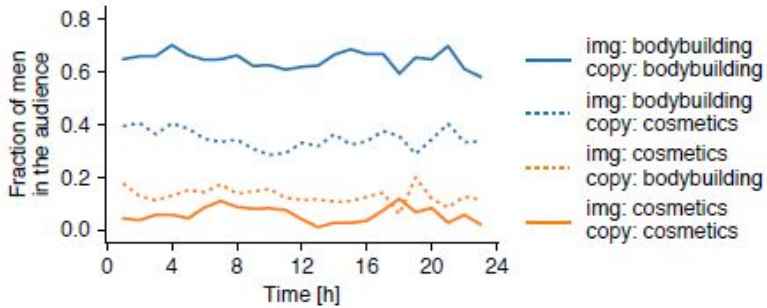
75% men

90% women









**Experiments used the ad text, domain, headline/title, and image

University of Virginia

Title VII: Disparate Treatment

- Two frameworks to prove employment discrimination
  - McDonnell-Douglas burden shifting
    - Establish different treatment despite protected class
    - Employer responds with reason, plaintiff must prove untrue
    - In Algorithms: Masking
      - Changing algorithmic process to obtain discriminatory outcome
  - Price-Waterhouse mixed motive
    - Motivation is Discriminatory
    - Bias in the Data can be put in:
      - The Decision: Using the data, knowing the issue
      - The Model making the decision
        - Human bias is the key problem, model unaddressed

Title VII: Disparate Impact

- Disparate Impact case:
  - Show a “neutral” practice is discriminatory
    - 4/5ths rule
  - Employer may show that practice is necessary
    - “…ascertains an applicant’s ability to perform successfully the job in question.”
  - Plaintiff must show alternative, less discriminatory practice that employer refuses to use
- In Algorithms
  - If Target Variable not job related, then discrimination shown!
  - Show that Model is trained on biased samples, mislabeled examples, and limited features
    - Showing refusal is hard! Can’t just ask LinkedIn to change it.

Copyright's Unexpected Role

Introduction to Copyright

- **Copyright** is a type of intellectual property that protects original works of authorship as soon as an author fixes the work in a tangible form of expression.
- Title 17 U.S.Code § 106: **Copyright Law**
  - paintings, photographs, illustrations, musical compositions, sound recordings, computer programs, books, poems, blog posts, movies, architectural works, plays…
  - protect creative works and grants exclusive rights to creators

What is the relationship between copyright and AI, and why does it matter to AI?

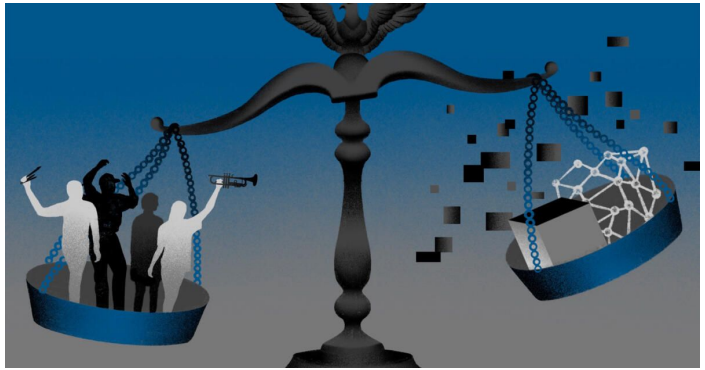
University of Virginia

Copyright in the Age of AI

University of Virginia

How Copyright contribute to AI Bias

- Limiting access to **training data**
- Prohibiting **reverse engineering**
- Restricting **algorithmic accountability**
- Hindering **competition**



University of Virginia

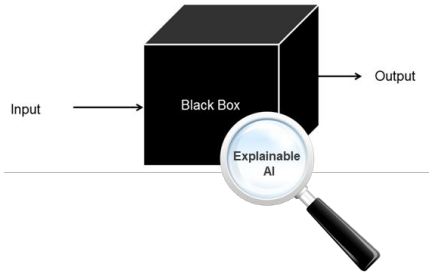
How Copyright contribute to AI Bias

- Limiting access to training data
  - Garbage in, Garbage out
  - Training data: biased, low-friction data (BLFD)
    - Easily available
    - Legally low-risk
  - Two main sources
    - Public domain works
      - Early literature, media, and news favor Western, white, and male perspectives
    - Creative Commons-licensed works
      - Only 8.5% of Wikipedia editors were women in 2011

AI trained on biased sources continues to amplify societal inequalities

How Copyright contribute to AI Bias

- Prohibiting reverse engineering
  - **Reverse engineering** is a way of leveraging available inputs or outputs to understand the mechanics of what happens inside a black box system.
  - **Digital Millennium Copyright Act (DMCA) § 1201**
    - Makes it unlawful to circumvent technological measures used to prevent unauthorized access to copyrighted works, including copyrighted books, movies, video games, and computer software.
    - **Researchers cannot analyze AI models for bias or fairness, keeping AI in a "black box."**



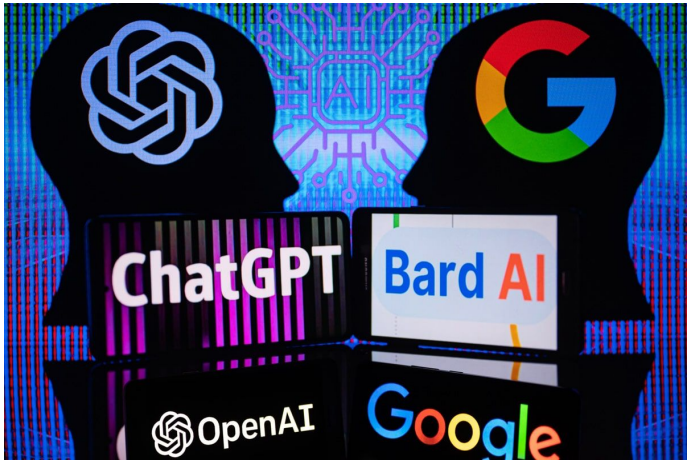
University of Virginia

How Copyright contribute to AI Bias

- Restricting algorithmic accountability
  - **Algorithmic accountability** aims to bring values like **transparency**, **explainability**, and **oversight** to the development and deployment of AI systems.
  - Copyright law restricts researchers and journalists from accessing key information about AI algorithms
  - Regulators and the public cannot audit AI systems

How Copyright contribute to AI Bias

- Hindering competition
  - Large companies have exclusive access to high-quality data
    - Google, Open AI
  - Small AI firms & researchers struggle to acquire training data
  - The AI industry is dominated by a few tech giants



Discussion

Do we need to reconcile the need for robust copyright protection with the AI community's demand for high-quality training data for fairer models?

- Do you agree with copyright's role with preventing access to training data? Why or why not?
- How would you change copyright law to allow access to training data?
- Can you think of some repercussions to your change?

University of Virginia

Case Study: Word2vec

Google's Word2Vec model, trained on Google News, learned biased word associations:

***"man** is to **computer programmer** in the same way that **woman** is to **homemaker**"*

- Google News corpus was not released
  - Researchers cannot analyze the dataset
- Readily available, low-risk dataset
  - Reinforce and amplify societal biases
- Copyright law prevented dataset transparency
  - Operate as black box

Fair Use as a Solution

- What is Fair Use?
  - A **copyright legal exception** allowing limited use of copyrighted works without permission.
- Four factors test:
  - Purpose and character of use
    - Transformative use
  - Nature of the copyrighted work
    - Factual Nature
  - Amount and substantiality used in
    - Reasonable Amount and Substantiality
  - Effect of the potential market
    - No Market Harm

University of Virginia

How Fair Use applied to AI training

- Why Fair Use is Relevant to AI?
  - AI models do not copy and sell copyrighted works, but rather learn from patterns in data.
  - AI training is highly transformative, meaning it creates something new rather than replicating the original work.
- Why AI Fair Use is Still Debated?
  - Courts have not explicitly ruled on whether AI training qualifies as Fair Use.
  - AI models consume entire datasets, raising concerns about the amount used.
  - Copyright holders argue that AI training could reduce demand for original works.

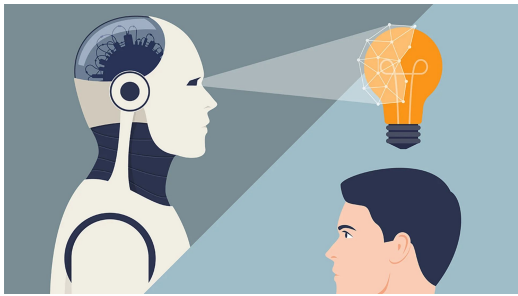
The Technical Challenge: How Algorithms Inherit Bias

Algorithms' Goal

- Target Variable
  - What we want to learn from our model
    - Is this Spam? (Spam Filter)
    - How much is this person worth lending money? (Credit Score)
  - Definition needed:
    - What makes a "good" employee for hire?
    - What are the examples needed?
- Training Data
  - How is the data in itself biased?
    - A reflection of the population?
    - An unintentional rule?

Setting the Example

- Where do we pick up bias in the data?
  - The Label
    - Where is the line for bad credit worthiness? 3 or 5 infractions?
  - The Pattern
    - Low-hiring for protected class?
    - Black people more likely to receive a criminal record advertisement?
  - The Collection
    - Less accurate data on minority classes
    - Misrepresentation? smartphone usage with map data
    - Overrepresentation? from constant observance



University of Virginia

Information Control

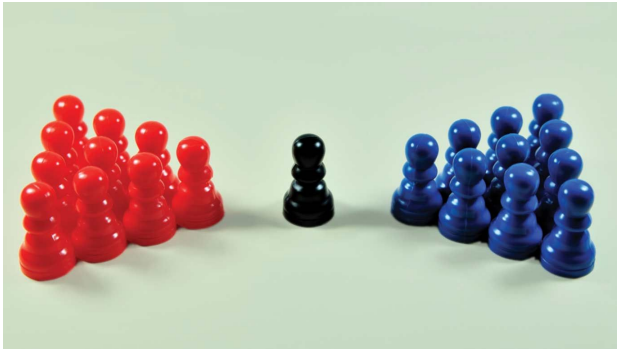
- What features matter in the model?
  - Lack of information about a topic leads to an imperfect substitute
    - Race != Criminal Record
    - College Prestige != Skills
- Information Proxies
  - Info about job excellence can lead to class determinism
  - Disparate Impact
- Masking information
  - Digital Redlining

University of Virginia

Discussion

Is it possible to create a dataset that is neutral across all protected classes?

- How could we make a neutral dataset?
- If a model is trained on a neutral dataset, could the model achieve fair results?



University of Virginia

Discussion

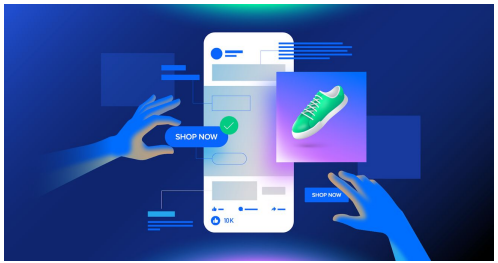
Q&A

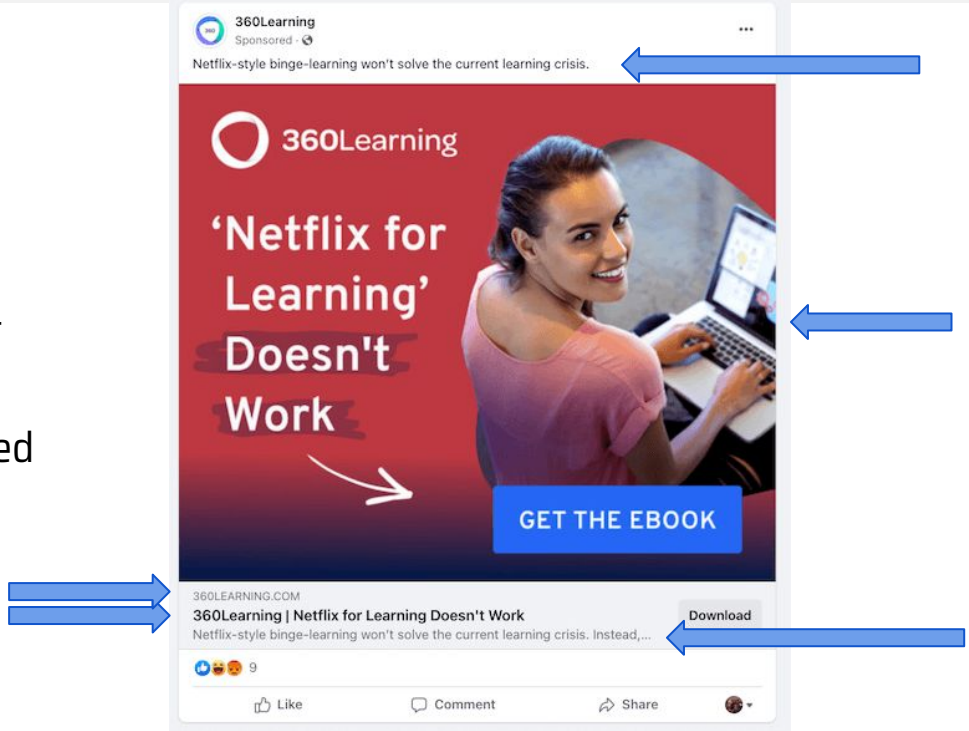
- When systems use protected attributes to address data biases or counteract the effects of historical discrimination, does that practice qualify as disparate treatment?
- How can the law keep up with the rapid emergence of new AI models and algorithms?
- What role should policymakers play in adapting anti-discrimination law for algorithmic decision-making?
- Should copyright law or fairness/anti-discrimination policies take priority when they conflict?
- Should companies that have breached copyright laws in training their AI models be required to revert their models, even if it results in reduced fairness and heightened bias?

Case Study: Facebook's Ad Delivery

You have an online advertising platform and want the ads to be successful among the platform's users.

- How would you reach the audience that the ad would be suitable for?
- What components would you would need to create the ad?





University of Virginia

Case Study: Facebook's Ad Delivery

- Ad creation vs. Ad delivery
  - Creation - advertisers submit text and images for ad, choose targeting parameters, bidding strategy
  - Delivery - platform delivers ads to specific users
- Role of ad delivery in discrimination
  - Aim to provide "ads that are most pertinent" users
  - Desired audience or unequal user availability → skewed delivery
    - **Skewed delivery** – ad delivery **not intended** and not resulted by targeting choices
- Reasons for skewed delivery
  - Market effects
  - Content of the ad (e.g., **image**)

Case Study: Facebook's Ad Delivery

- In terms of avoiding discriminations, Facebook has policies:
  - Pledged to prevent use of some targeting categories for sensitive ads
    - i.e., housing, employment, credit
  - Advertisers self-certification
    - No violation of ad policy against discriminatory practices
  - No longer allows age, gender, ZIP-code in targeting for sensitive ads
  - Blocks targeting attributes related to protected classes
  - Transparency initiative – providing users with explanations of why they are seeing particular ads

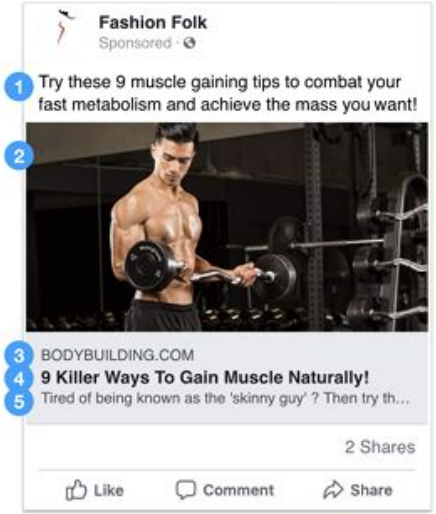
What are potential problems with Facebook's anti-discriminatory efforts?

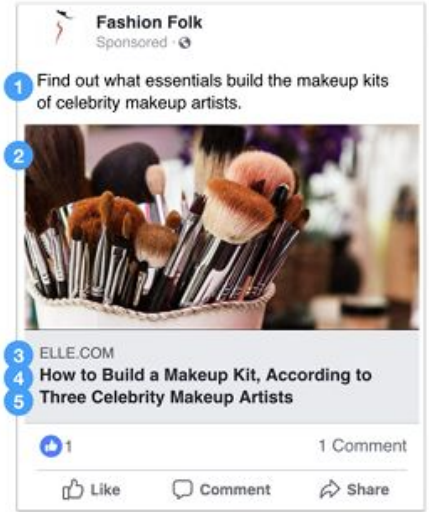
Case Study: Facebook's Ad Delivery

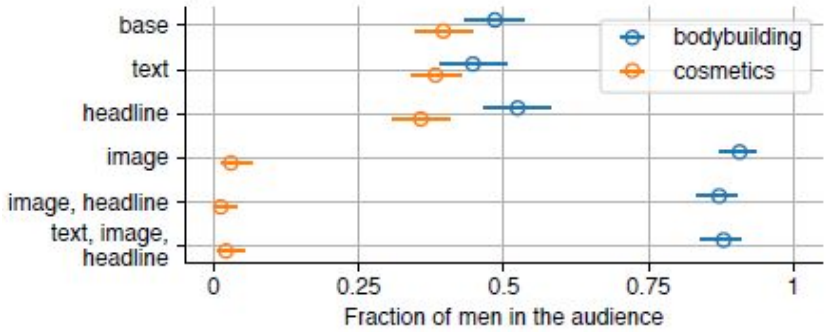
Ads delivered to, on average,

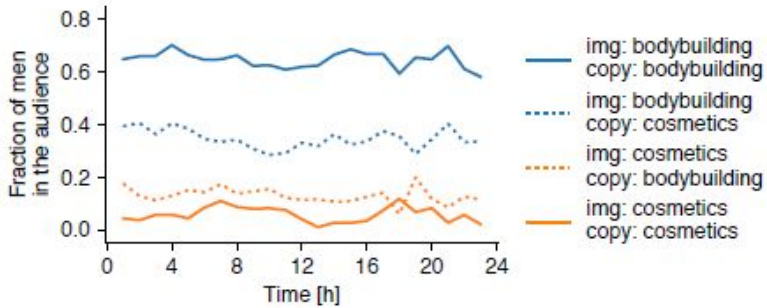
75% men

90% women







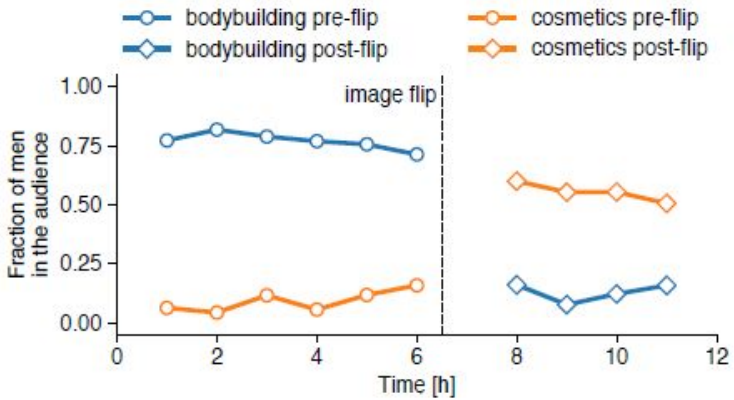


**Experiments used the ad text, domain, headline/title, and image

University of Virginia

Case Study: Facebook's Ad Delivery

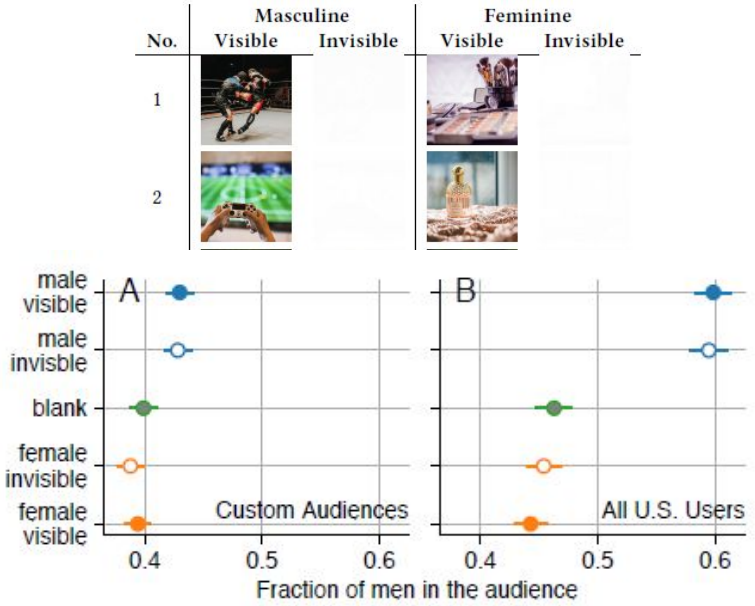
Experiment with switching the images of the body-building and cosmetic ads half way through an ad campaign:



**Invisibility refers to adding alpha channel to images with 98% opacity

** Not all invisible ad image shown in presentation

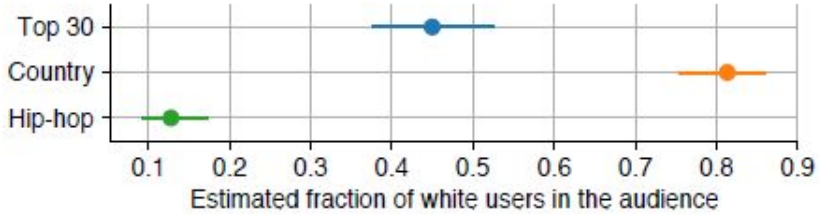
Experiment making ad images invisible:



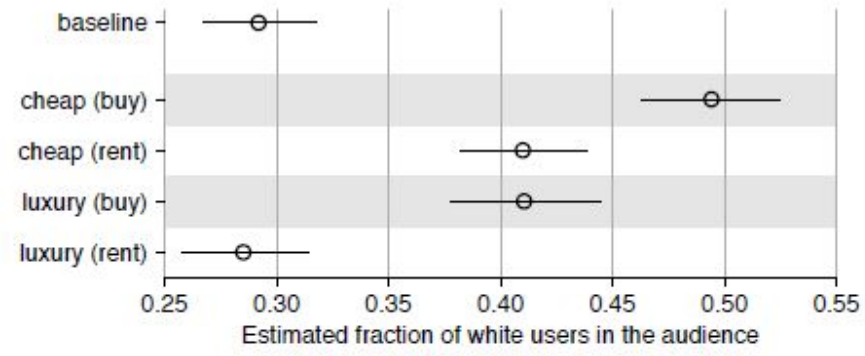
University of Virginia

Case Study: Facebook's Ad Delivery

Experiment with music ads based on race:



Experiment with housing ads based on race:



What features could be resulting in the racial skewed ad delivery for housing ads?

The custom audiences are from NC, but would the trend of these results change if the custom audiences were from another geographical location? Why?

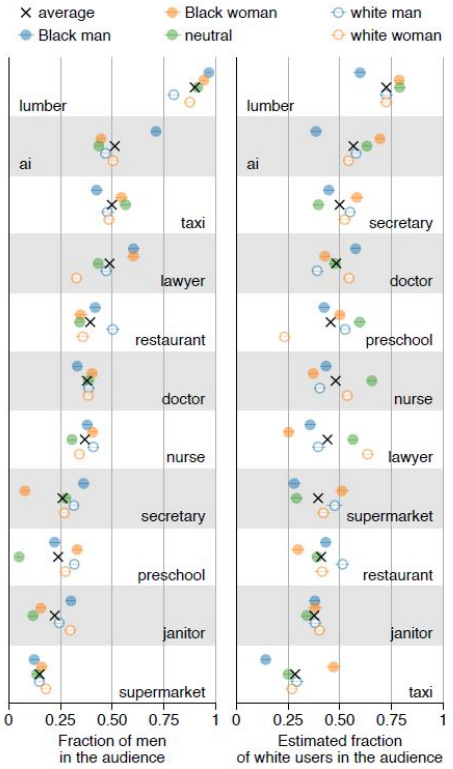
University of Virginia

Case Study: Facebook's Ad Delivery

Is it possible to create a dataset that is neutral across all protected classes?

If a model is trained on a neutral dataset, could the model achieve fair results?

Experiment with employment ads:



Putting It All Together: A Multi-Faceted Approach

Key Takeaways

Big data and ML necessitate a redefinition of fairness and discrimination and new regulation that fit the modern context

- U.S. laws provide a foundation for fairness, but are inadequate for addressing bias in ML, which may perpetuate or amplify existing inequalities
- Copyright law may contribute to AI bias by limiting access to diverse data, questioning the current balance between intellectual property rights and AI fairness
- Algorithmic decision-making may unintentionally reinforce societal biases causing disparate impact that current legal frameworks struggle to address

Thank You

CS 6501: Responsible AI

References

- Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. Proceedings of the ACM on human-computer interaction, 3(CSCW), 1-30.

- Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press

- Barocas, S., & Selbst, A.D. (2016). Big Data's Disparate Impact. California Law Review, 104, 671.

- Levendowski, Amanda, "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem" (2018). Georgetown Law Faculty Publications and Other Works. 2439. https://scholarship.law.georgetown.edu/facpub/2439