MY474: Applied Machine Learning for Social Science

Lecture 10: Bias, Fairness, and Ethics in Machine Learning

Blake Miller

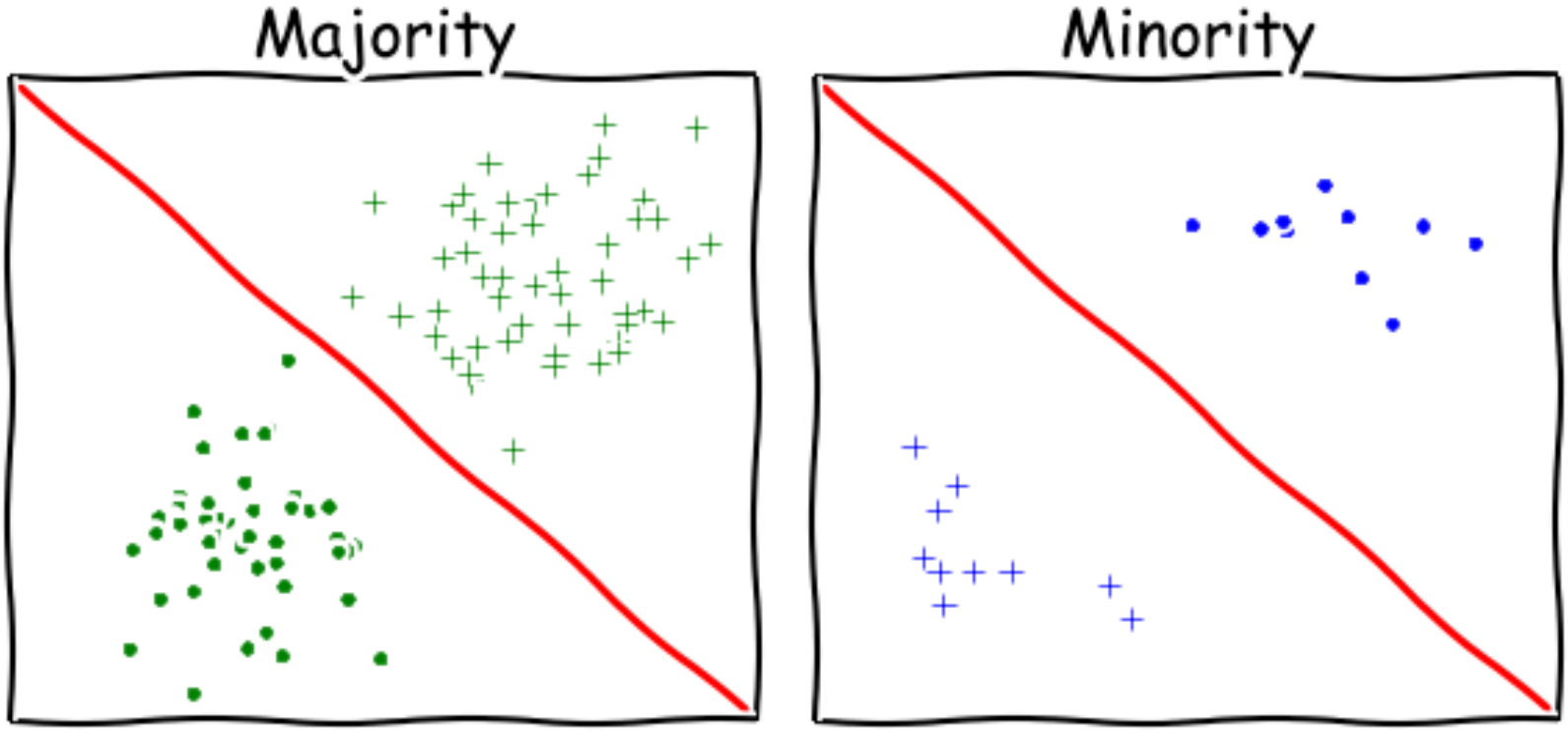
11 December 2019

Agenda

1. What is algorithmic bias?
2. How do we address bias?
3. Ethical concerns with ML

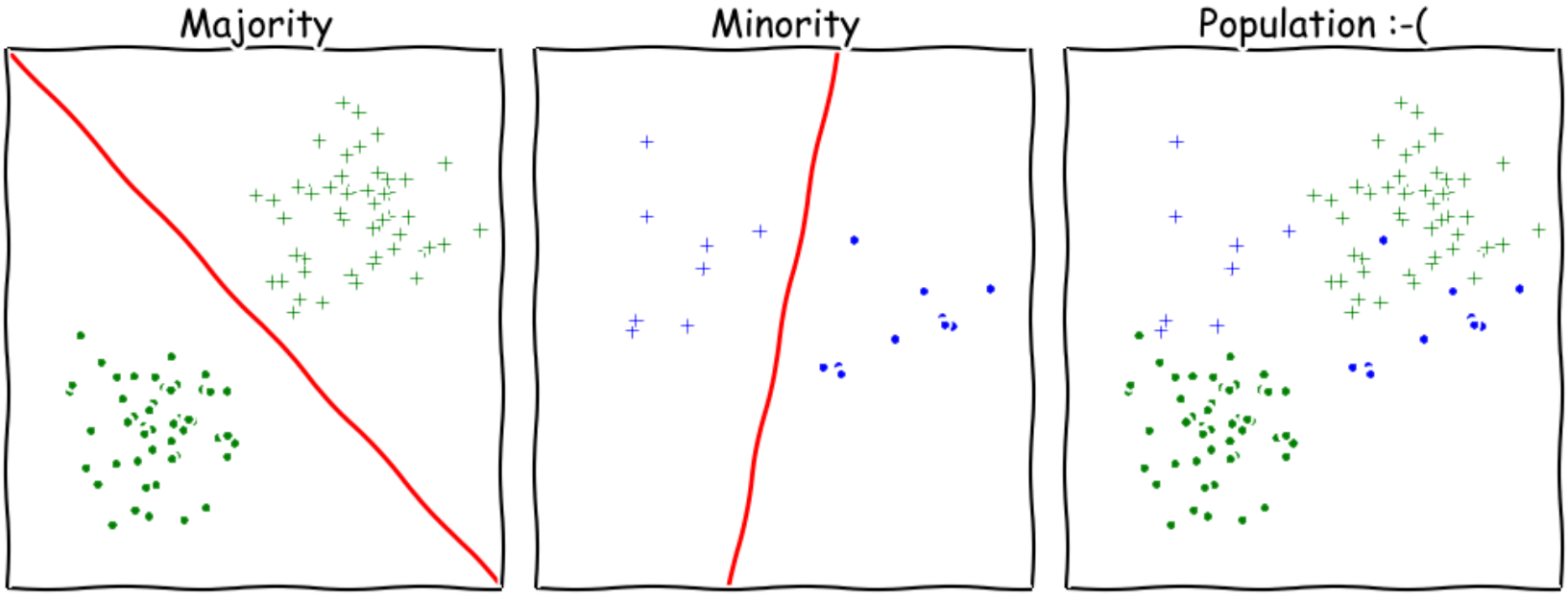
Algorithmic Bias

Example of Bias in Classification



Positively labeled examples are on opposite sides of the classifier for the two groups.

Example of Bias in Classification



Even if two groups of the population can be simply classified, the whole population may be more difficult to classify.

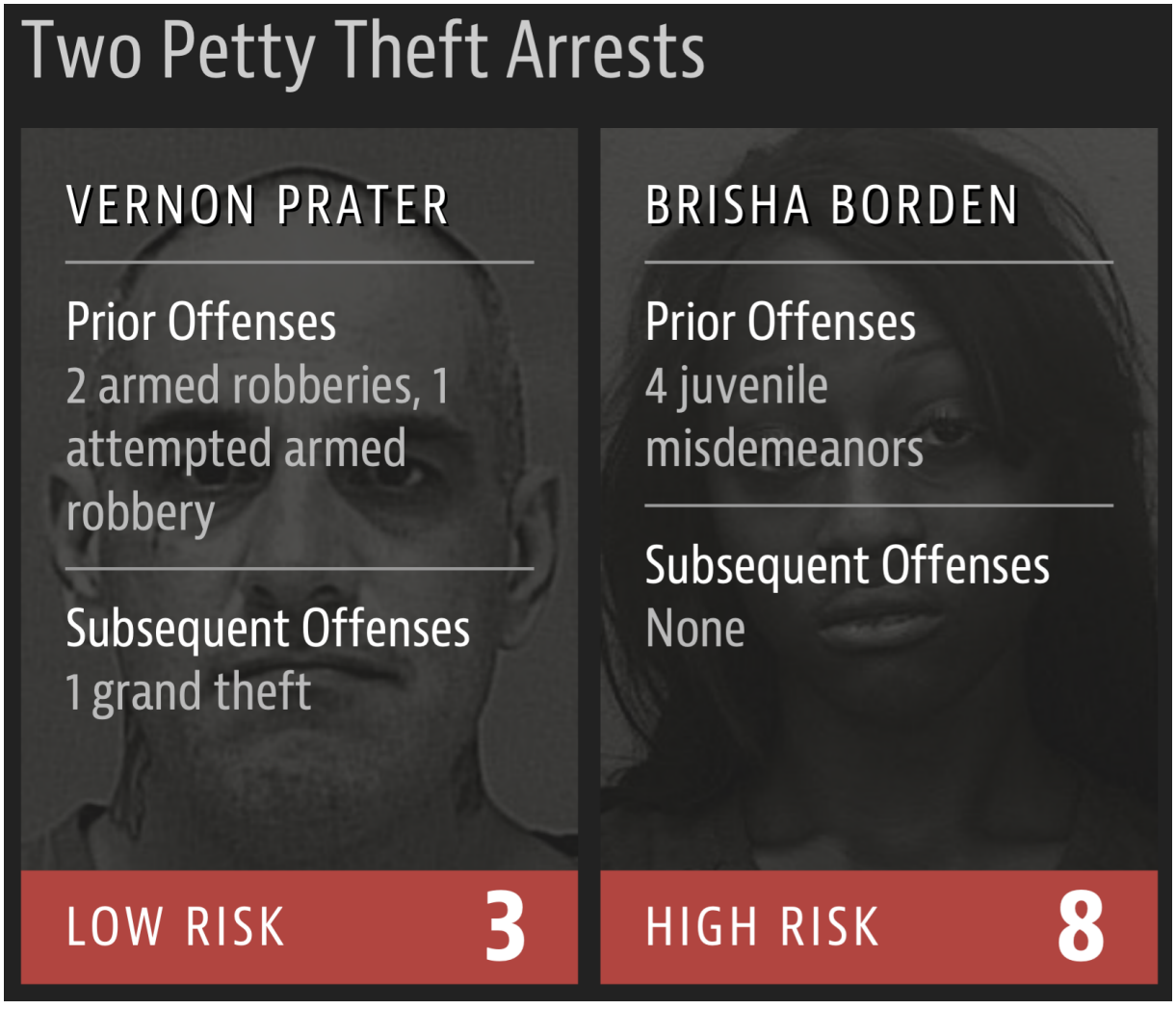
Data Reflect Societal Biases

- Data by their nature are biased; we must make arbitrary decisions about how to measure and record the world around us.
- These decisions will reflect our culture, language, and history
- Social data also represent the world we live in, biases can be captured in our data even if we do not intend for them to be.

Data Reflect Societal Biases (Example)

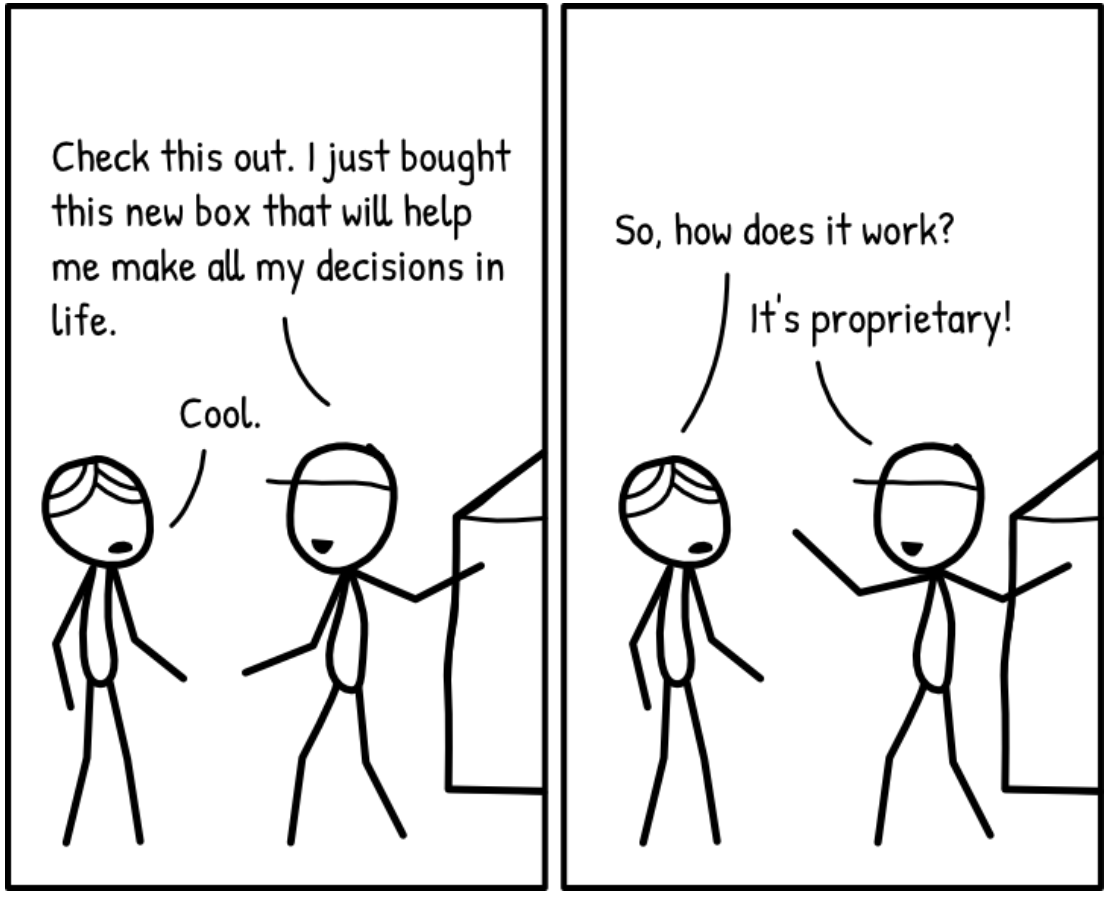
- Even if data on individuals do not include race as a predictor, it can still be represented through variables such as postal code, which provide information on race due to racial segregation, historical policies such as redlining, etc.
- COMPAS, a predictive policing software used across the United States to predict future criminals is biased against blacks; the systems predictions about the risks of reoffending are given to judges during criminal sentencing.

Data Reflect Societal Biases (Example)



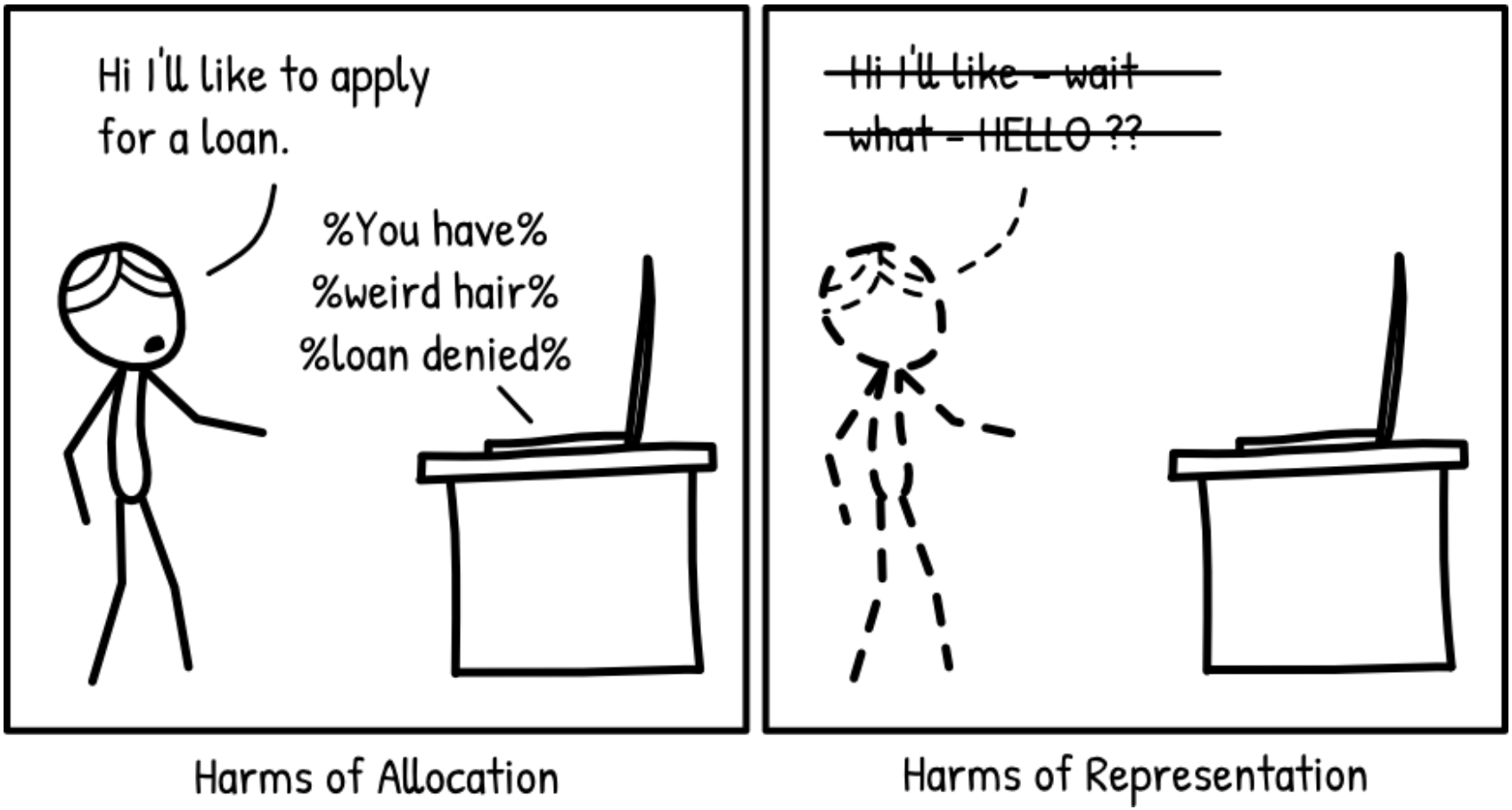
Example: Predictive policing systems such as COMPAS may perpetuate racial bias even though race is not an input variable. Source: "Machine Bias," Propublica.

Challenges in Addressing Bias: Disclosure



Systems like COMPAS are not transparent and they do not have to disclose how their systems work. We only find out about bias due to excellent investigative reporting. Source: https://machinesgonewrong.com/bias_i/

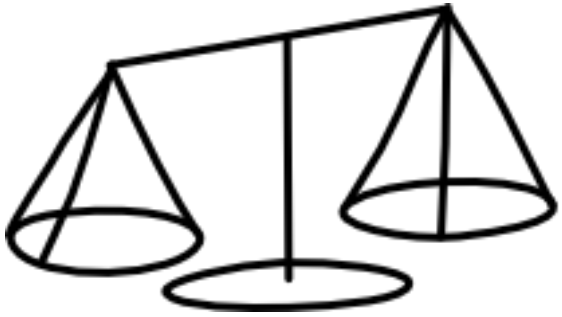
Types of Harm from Bias

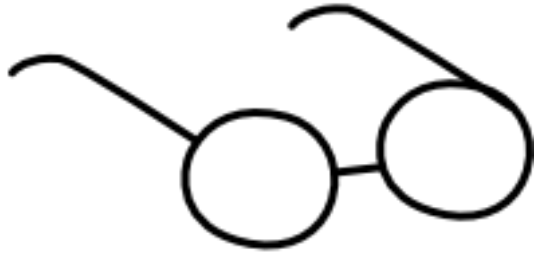


- **Harms of allocation:** opportunities and resources are unfairly allocated due to algorithmic intervention.
- **Harms of representation:** algorithms produce representations that are discriminatory

Image: https://machinesgonewrong.com/bias_i/

Types of Harm from Bias

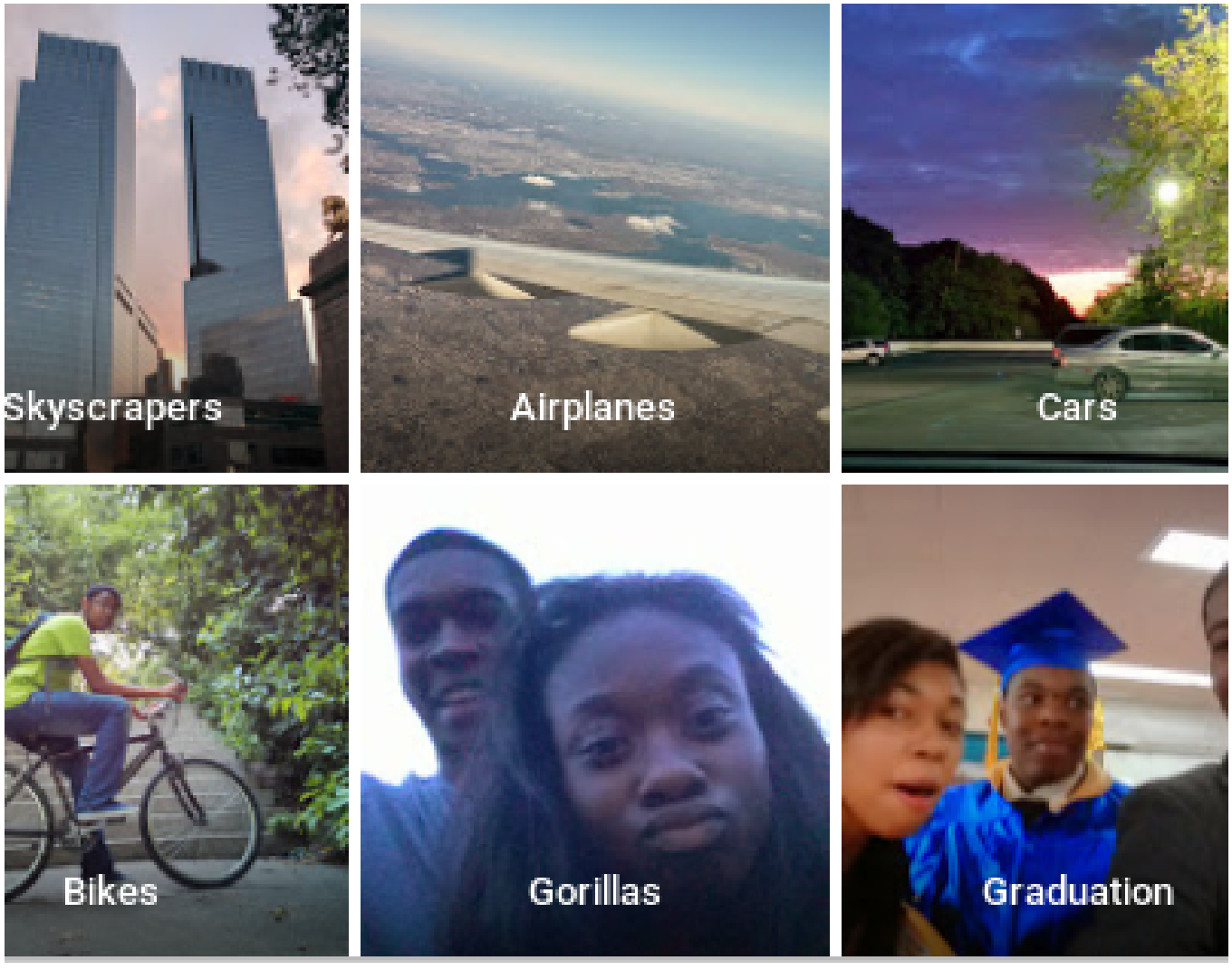




Harms of Allocation	**Harms of Representation**
Immediate	Long term
Easily quantifiable	Difficult to formalize
Discrete	Diffuse
Transactional	Cultural

Table: Kate Crawford, NIPS 2017 Keynote, Image: https://machinesgonewrong.com/bias_i/

Harm from Bias Example



Google's image recognition algorithm idenified images of black people as gorillas.

Harm from Bias Example

Joy Buolamwini, researcher of algorithmic fairness, needed to wear a white mask to be recognized by a common facial reconition tool.

Types of Harm from Bias

- What are some examples of representative harm in ML?
- What are some examples of allocative harm in ML?

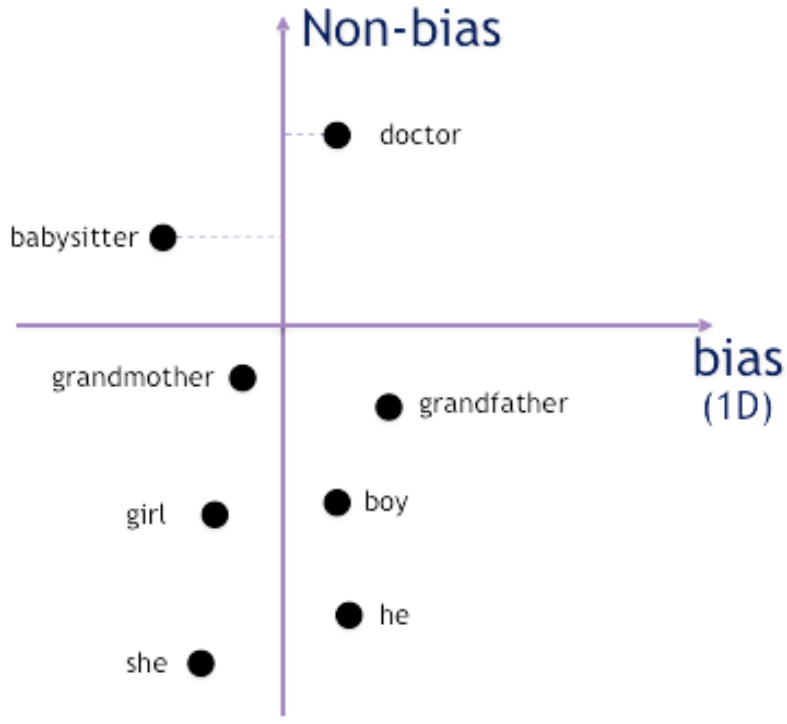
Addressing Bias

Bolukbasi et. al.

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

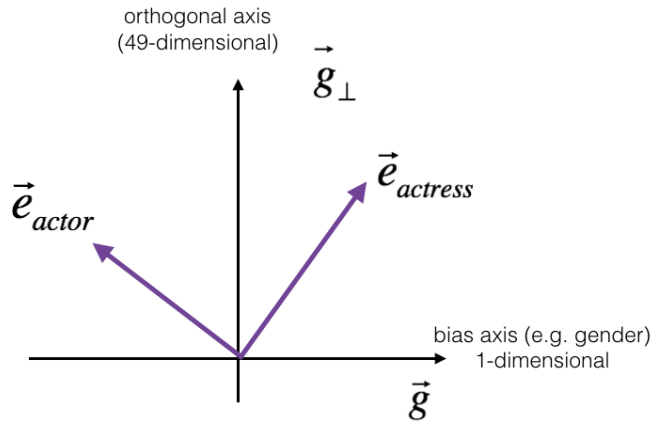
How should we deal with problematic associations encoded into word embeddings.

Bolukbasi et. al.

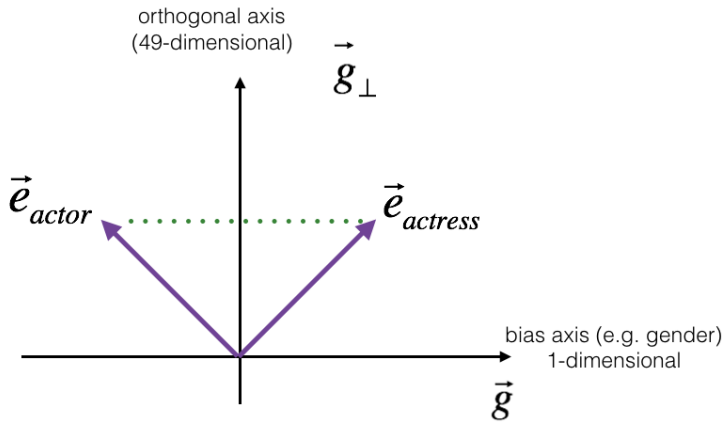


1. Identify the direction of bias (e.g. the gender subspace)
2. Neutralize: For all non-definitional words, project to get rid of bias.
3. Equalize pairs of words we want to differ only through the gender property (e.g. he, she)

Bolukbasi et. al.



before equalizing,
"actress" and "actor" differ
in many ways beyond the
direction of $\vec{g}$



after equalizing,
"actress" and "actor" differ
only in the direction of $\vec{g}$, and further
are equal in distance from $\vec{g}_\perp$

Debiasing applied to word pairs "actress" and "actor."

- **Equalization** makes sure that a particular pair of words are equi-distant from the orthogonal non-bias vector
- This ensures that the words actress and actor are equidistant from any other word that has been neutralized (e.g. babysitter).

Discussion

What are some limitations of technical approaches to dealing with bias such as those offered in Bolukbasi et. al.?

Discussion

Recall the "CEO" Google Image search discussed by Kate Crawford in the assigned NeurIPS keynote video. Racial and gender representation in the results arguably reflect the de facto distributions we observe in CEOs today. **How are we to decide upon the appropriate representation in this example? Should we privilege accuracy or idealism?**

Discussion

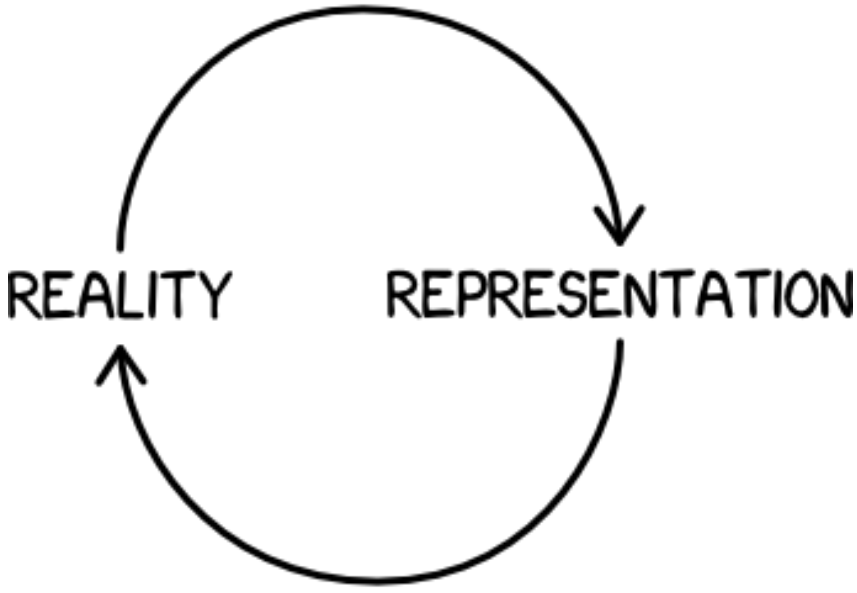


Image: https://machinesgonewrong.com/bias_i/

Characterizing the Population

- To mitigate bias we need to be clear what the population looks like
- Imagine an access system to the subway that uses facial recognition; what does the population of subway passengers look like?
- How is the population of subway passengers distributed amongst the features?
- The population's characteristics should inform our sampling for the training data.
- Underrepresenting one group could have serious consequences for this group

Beware of Bias of Labelers

- Most supervised learning algorithms require human-labeled data
- For example, some social networks hire humans to manually label sexually explicit content or pornography
- What happens when one of the labelers thinks gay people showing affection (kissing, hugging) is sexually explicit?
- Solution 1: track coder reliability through routine audits of labelers. How well do coders agree?
- Solution 2: allow user feedback to alert the social network when the algorithm gets it wrong.
- Question: what are potential problems with each of these approaches?

Beware of Proxy Features

- **Proxy features** are features that measure **protected traits** such as race, ethnicity, gender identity, sexual orientation, age, etc.
- An example of a proxy feature is postal code. Postal codes may end up being a proxy for race due to segregation of neighborhoods in the United States.

Beware of Mismatched Domains

Ground truth: Soap **Nepal, 288 $/month**

Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment

Ground truth: Soap **UK, 1890 $/month**

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave

Comparing images of soap from different cultures in the Dollar Street dataset, from DeVries et al..

Beware of Mismatched Domains

- Oftentimes there is a mismatch between the **domain** of training data and the population of individuals we are predict outcomes for/classifying
- In the absense of training data from the **target population** computer scientists often train on **convenience datasets**
- Convenience datasets may have been labeled for a similar task, but they may not be a perfect fit for your target population. Using a convenience dataset in this scenario is called **transfer learning**
- Imagine we used the movie review dataset from problem set 4 to measure sentiment of legislative debates in Congress or Parliament? Would it do an ok job? What problems might arise?

Social Approaches to Mitigating Bias

- Legal approaches
  - DPR: Right to explanation
  - Antidiscrimination laws
- **Social conditions** surrounding the **data generating process** affect predictions
  - Are certain groups over-represented?
  - Does the data encode biases in human behavior, language, or power structures?
  - Are data representative of the population they will serve, classify, or evaluate?
  - When were data generated? What historical biases might be encoded in the data? (different values, ideologies, etc.)

Technical Approaches to Mitigating Bias

- Use **explainable** models in certain circumstances.
- **Debiasing** training data
- "Fairness Forensics": measuring discrimination
  - Group-level comparisons: Is there statistical parity among groups? (e.g. are blacks and whites misclassified at the same rate?)
  - Individual-level comparisons: Do two otherwise similar individuals in different groups (race, gender, age, etc.)

Checklist for Analysts

A good reference for how to ensure fairness and mitigate bias in ML projects:

https://machinesgonewrong.com/checklist/

Ethics

Three Painful Truths (Diebert)

1) Social media is built around personal-data surveillance, with products ultimately designed to spy on us in order to push advertising in our direction.
2) We have consented to this, but not entirely wittingly: Social media are designed as addiction machines, expressly programmed to draw upon our emotions.
3) Attention-grabbing algorithms underlying social media also propel authoritarian practices that aim to sow confusion, ignorance, prejudice, and chaos, thereby facilitating manipulation and undermining accountability. Moreover, the fine-grained surveillance that companies perform for economic reasons is a valuable proxy for authoritarian control.

Discussion

- How do we weigh the costs and benefits of ML technology that might prove harmful to democracy, freedom of speech, etc.?
- Is all knowledge production and all technological progress necessarily good? Where do we draw the line?

Do the benefits outweigh the risks?

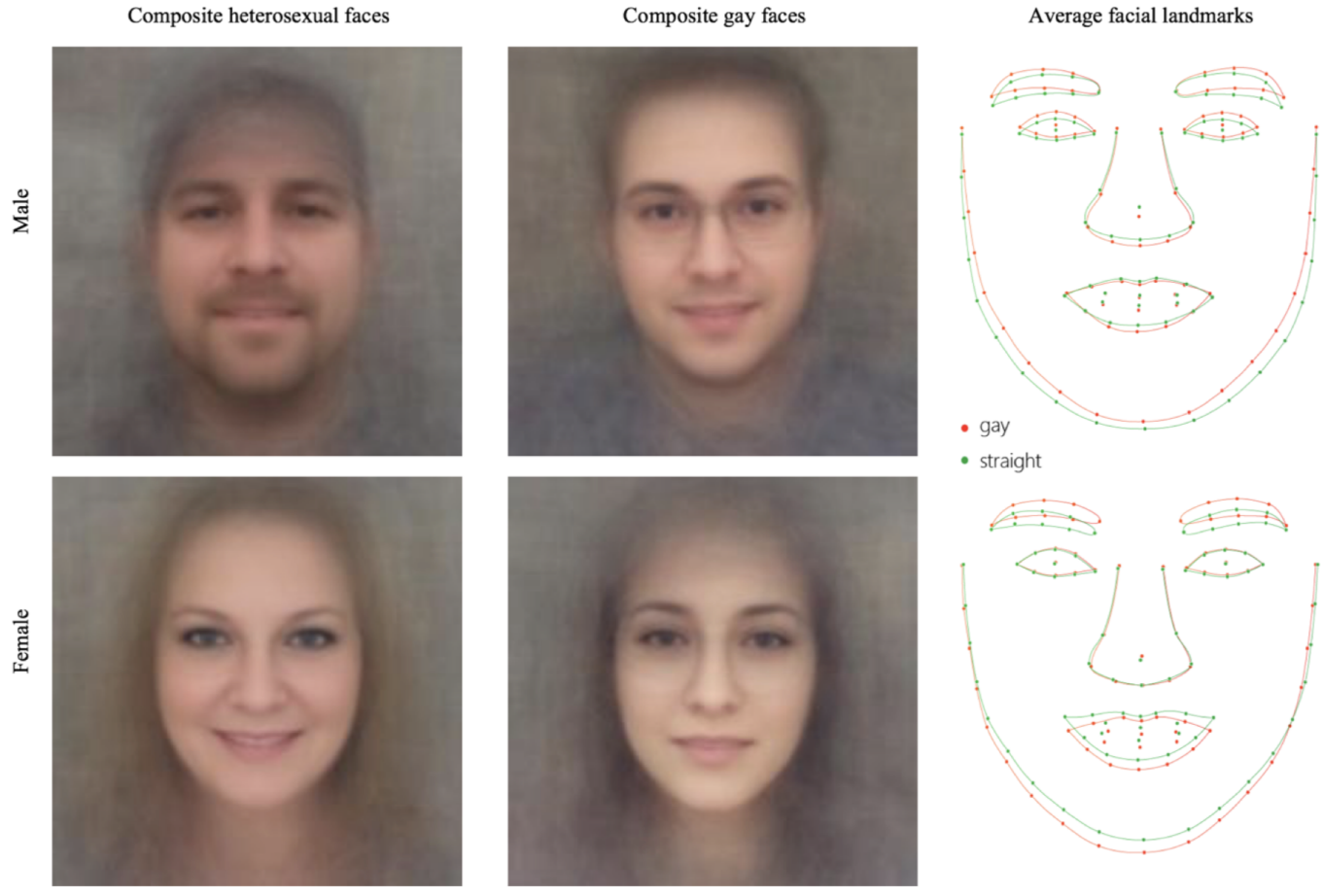


Figure 4. Composite faces and the average facial landmarks built by averaging faces classified as most and least likely to be gay.

Source: Wang and Kosinski, 2017

Serving digital authoritarianism

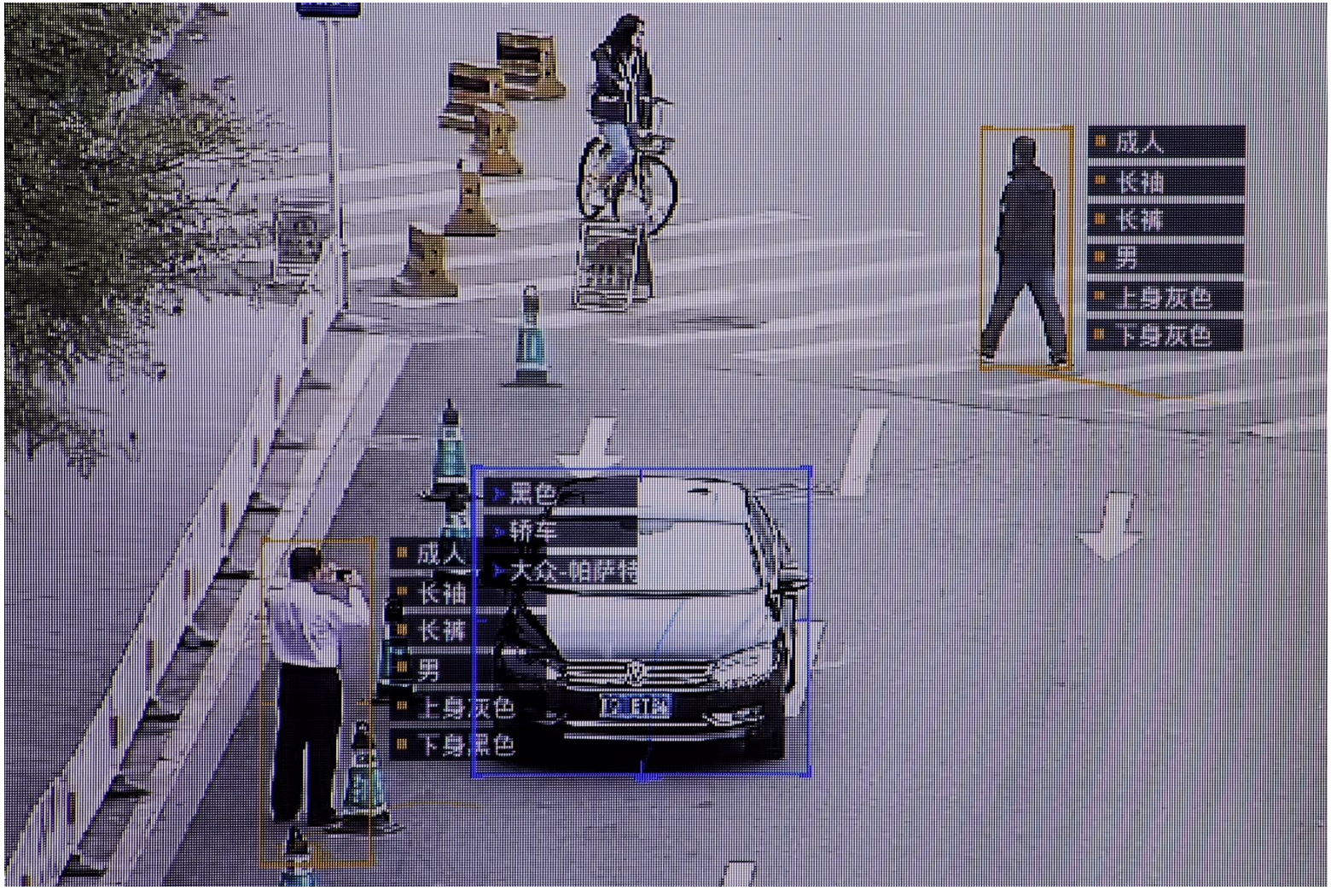
(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Source: Wu and Zhang, 2016

Serving digital authoritarianism



Sensetime, a Hong Kong-based company provides real time crowd analytics and facial recognition to mainland Chinese domestic security bureaucracies.

The Politics of Classification: Discussion

What are the potential harms caused by government systems that classify citizens into groups? How can we mitigate them?