Fairness in Machine Learning

SML310: Research Projects in Data Science, Fall 2019

Michael Guerzhoy

Content from Moritz Hardt, Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel

COMPAS

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html

https://www.propublica.org/article/technical-response-to-northpointe

https://www.liebertpub.com/doi/pdf/10.1089/big.2016.0047

COMPAS

- "Correctional Offender Management Profiling for Alternative Sanctions"
  - Developed by Northpointe (currently Equivant)
  - Used by *a lot* of probation departments to assess the likelihood of a defendant becoming a recidivist
  - Defendants who are defined as medium or high risk are more likely to be detained before trial
    - (N.B., this is only suggestive of importance)
  - Race is not an input to the algorithm

COMPAS Probation Risk and Needs Assessment Questionnaire

OFFENDER NAME:	NYSID:	STATUS:
RACE:	SEX:	DOB:
DATE OF ASSESSMENT:	MARITAL STATUS:	
SCALE SET: Full COMPAS Assessment v2	AGENCY/COUNTY NAME:	

PART ONE: CRIMINAL HISTORY / RISK ASSESSMENT

CURRENT CHARGES

What offenses are covered by the current charges (check all that apply)?

Homicide	Arson	Property/Larceny
Assault	Weapons	Fraud
Robbery	Drug Sales	DWI / DWAI
Sex Offense (with force)	Drug Possession	AUO
Sex Offense (without force)	Burglary	Other

1 Do any of the current offenses involve domestic violence?

Yes No

2 What offense category represents the most serious current charge?

Misdemeanor Non-Assault Felony Assaultive Felony

3 Was there any degree of physical injury to a victim in the current offense?

Yes No

4 Based on your judgment, after reviewing the history of the offender from all known sources of information (PSI, police reports, prior supervision, victim, etc.) does the defendant demonstrate a pattern of violent behavior against people resulting in physical injury?

Yes No

http://www.northpointeinc.com/downloads/research/DCJS_OPCA_COMPAS_Probation_Validity.pdf

PART TWO: NEEDS ASSESSMENT

A. ASSOCIATES / PEERS

17 The offender has peers and associates who *(check all that apply)* :

- Use illegal drugs
- Have been arrested
- Have been incarcerated
- None
- Lead law-abiding lifestyles
- Are gainfully employed
- Are involved in pro-social activities

18 What is the gang affiliation status of the offender :

- Current gang membership
- Previous gang membership
- Not a member but associates with gang members
- None

19 Does the offender have a criminal alias, a gang-related or street name?

Yes No

20 Does unstructured idle time contribute to the opportunity for the offender to commit criminal offenses?

Yes Unsure No

21 Does offender report boredom as a contributing factor to his or her criminal behavior?

Yes Unsure No

B. FAMILY

22 Are the offender 's family or household members able and willing to support a law abiding lifestyle?

Yes Unsure No

23 Is the offender's current household characterized by *(check all that apply)* :

COMPAS Probation Risk and Needs Assessment Questionnaire – *Continued*

PART THREE: OFFENDER QUESTIONNAIRE

NYSID : Name : DOB :

Please look at the following areas and let us know which of them you think will present the greatest problems for you. *Please check one response for each question in the column provided* .

	Please answer questions as either No, Yes or Don't Know	No	Yes	Don't Know
48	Do you feel you need assistance with finding or maintaining a steady job?			
49	Do you feel you need assistance with finding or maintaining a place to live?			
50	Will money be a problem for you over the next several months?			

	How difficult will it be for you to...	Not Difficult	Somewhat Difficult	Very Difficult
51	manage your money?			
52	keep a job once you have found one or if you currently have one?			
53	find or keep a steady place to live?			
54	have enough money to get by?			
55	find or keep people that you can trust?			
56	find or keep friends who will be a good influence on you?			
57	avoid risky situations?			
58	learn to control your temper?			
59	find things that interest you?			
60	learn better skills to get or keep a job?			
61	find a safe place to live where you won't be hassled or threatened?			
62	get along with people?			

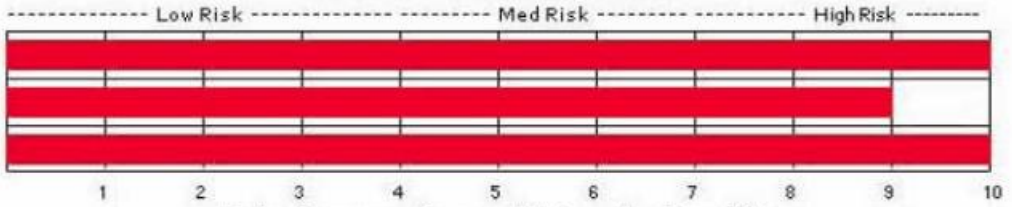
COMPAS Probation Risk Assessment

Offender: **Joe Sample** DOB: **2/2/1950** Gender: **Male**

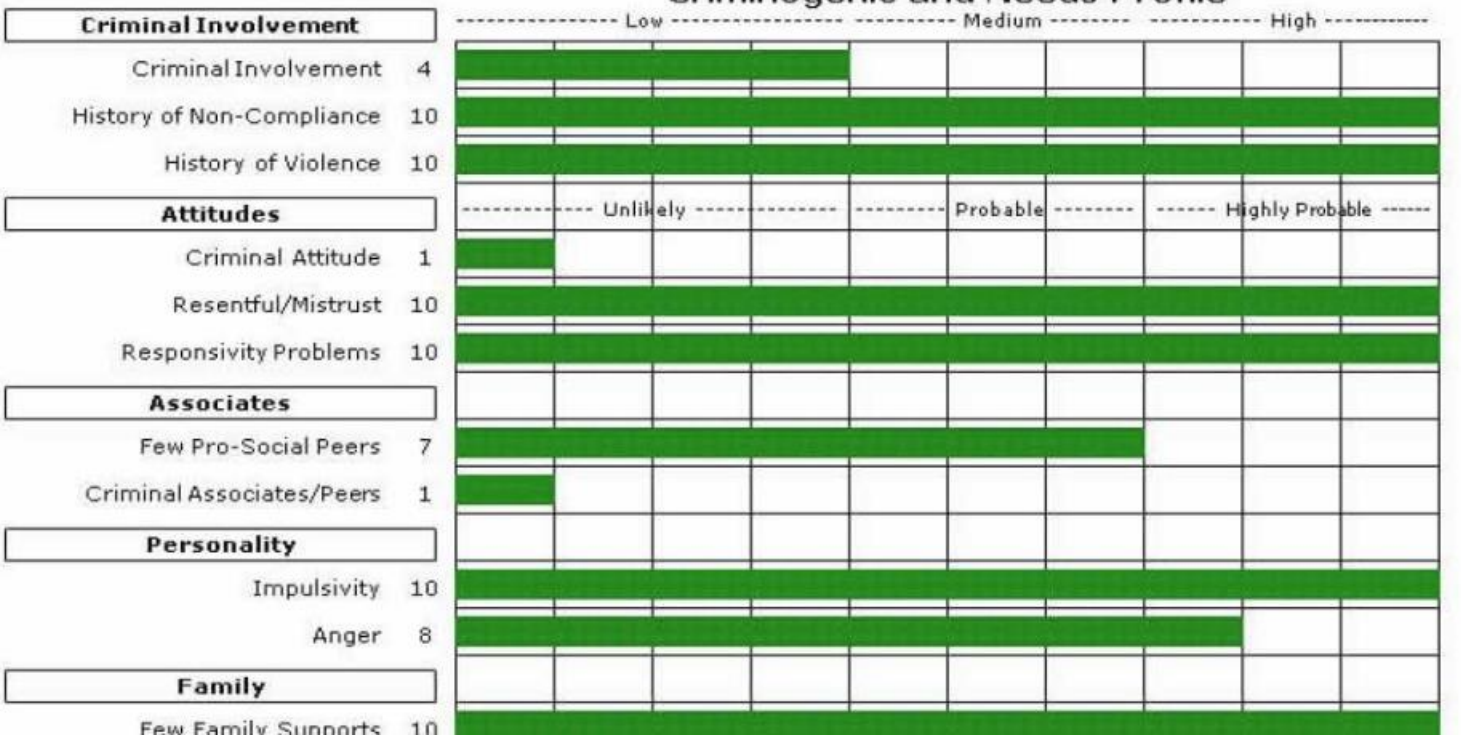
Screening Date: **9/13/2007** Screener: **Hellem, Dan** Ethnicity: **Native A**

Scale Set: **DMB-PSI** Case: **009943** Marital Status: **Single**

Overall Risk Potential



Criminogenic and Needs Profile



Observational measures of fairness

- C – output of the classifier
- Y – ground truth (rearrested/was not rearrested)
- D – demographic
  - For simplicity 0 or 1
- X – features
- Demographic parity
  - $P(C = 1|D = 0) = P(C = 1|D = 1)$
- False positive parity (“equal opportunity”)
  - $P(C = 1|D = 0, Y = 0) = P(C = 1|D = 1, Y = 0)$

Observational measures of fairness

- Demographic parity
  - $P(C = 1|D = 0) = P(C = 1|D = 1)$
  - Everyone is predicted to re-offend at the same rate, regardless of demographic
  - A type of "classification parity"
- False positive parity ("equal opportunity")
  - $P(C = 1|D = 0, Y = 0) = P(C = 1|D = 1, Y = 0)$
  - People who did not reoffend predicted to reoffend at the same rate, regardless of demographics
  - A type of "classification parity"
- Accuracy parity
  - $P(Y = 1|C = 1, D = 0) = P(Y = 1|C = 1, D = 1)$ and $P(Y = 1|C = 0, D = 0) = P(Y = 1|C = 0, D = 1)$
  - People predicted to reoffend actually reoffend at the same rate, regardless of demographics

Calibration

- $P(Y = 1|s(X) = s, D = 0) = P(Y = 1|s(X) = s, D = 1)$
  - The probability of re-arrest for people who got the same risk scores is the same
  - N.B.: if the score is 0/1, this reduces to

$$P(Y = 1|C = 1, D = 0) = P(Y = 1|C = 1, D = 1)$$
$$P(Y = 1|C = 0, D = 0) = P(Y = 1|C = 0, D = 1)$$

Anti-classification

- Protected characteristics are not considered
- $P(C = 1|X) = P(C = 1|X')$ if X and X’ only differ by protected demographic

Utility functions

- Can assign a cost to each of true positive/true negative/false positive/false negative, and then compute the expected utility for a rule for making decisions
- Optimal rules are of the form
  $P(Y = 1|X) \geq thr$
- Sketch of proof
  - An exchange argument: always better to predict C = 1 for riskier individuals

Generally, can’t satisfy two measures simultaneously

Accuracy parity vs. PPV Parity

Low-risk: 10% chance of re-arrest

High-risk: 80% chance of re-arrest

Group A	Group B
Low-risk: 40, High-risk: 60	Low-risk: 50, High-risk: 50

- Assume the system perfectly identifies low vs. high-risk
- Group A: Predict 60 will be arrested. 12/60 won't be.
- Group B: Predict 50 will be arrested. 10/50 won't be.
- Group A: error rate is $\frac{12+4}{100} = 16\%$
- Group B: error rate is $\frac{10+5}{100} = 15\%$
- Equalizing the error rates (perhaps by randomly erring when deciding about group B, if the user is acting in bad faith) will mess up the false-positive parity

Accuracy disparity when False Positive Parity holds

- The mix of False Positives is different for different populations
  - Mix of high-risk individuals and low-risk individuals who did not end up re-offending

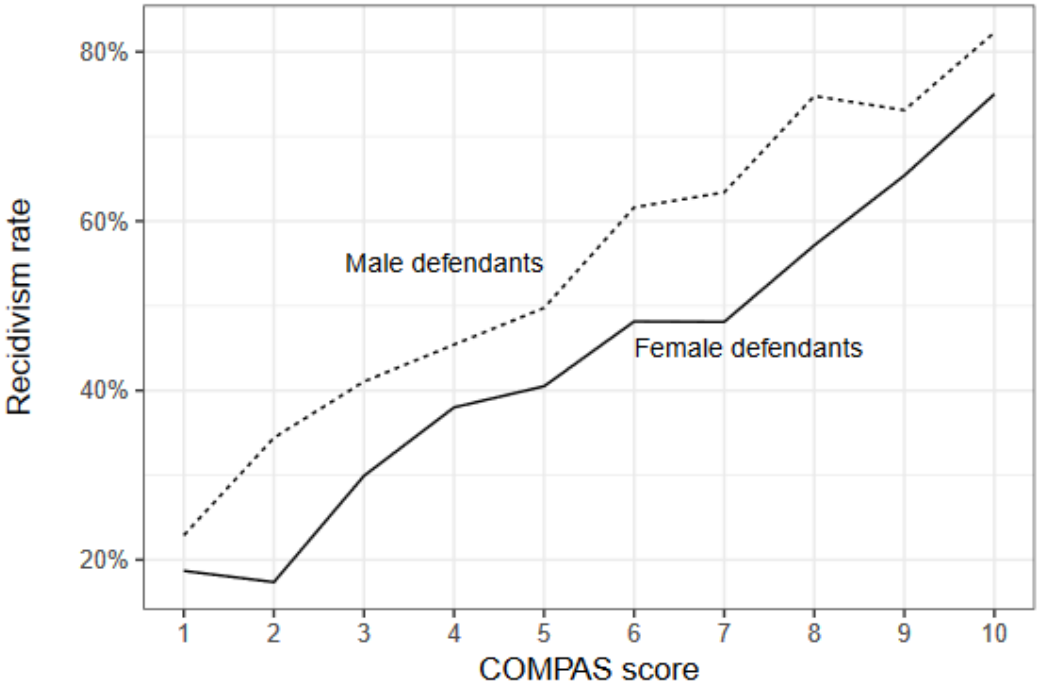
Discrimination before Fairness in ML

- Statistical discrimination
  - Charging male drivers more for insurance
  - Predicting younger people are more likely to reoffend
  - Predicting male defendants are more likely to reoffend
- “Taste-based discrimination”
  - Discrimination by the decision-maker that decrease an objective measure of the decision-maker’s utility (the decision-maker has a “taste for discrimination”) (Gary Becker 1957)

Discrimination before Fairness in ML

- Law usually focuses on the *intent* of the decision-maker to commit taste-based discrimination
  - If there is an observed disparity, that can trigger "strict scrutiny": the decision-maker needs to justify their decision
- In housing and employment, statistical disparities can be illegal unless they are justified
  - Griggs v Duke Power: the company could not require a high-school diploma for promotion since it was found there was no relation between job performance and having a diploma, because of racial disparity in promotion/having a diploma
  - "Unjustified disparate impact": intent to discriminate *not* needed for the requirement to be illegal

Limitations of Anti-Classification



Sometimes need to consider demographics to get the best probability. COMPAS didn't, So there's no calibration wrt gender

Limitations of demographic parity/FP parity/etc

- Not necessarily compatible with each other
- Not compatible with calibration
  - (Again, calibration: scores mean the same thing regardless of demographic)

Limitations of calibration

Presence of discrimination despite calibration

- Redlining: the practice of not approving loan applications for predominantly black neighborhoods
- When predicting default rates just based on the zip code, calibration could be satisfied
  - If black neighborhoods are also generally poorer
  - There can be discriminatory intent in neglecting to use other features of the individuals

Label bias

- The y's (outcomes) in the training set might not be labelled correctly
  - In the COMPAS data, y = 1 if there was re-arrest
  - But we *want* to measure violent crime
    - Racial bias in the amount of policing in different neighborhoods
      - But could downweight e.g. drug arrests
    - Some arrests are not for violent crime
  - We don't have counterfactual information
    - We observe data that's conditioned on a judge's past decision
      - But can look at the two years after the release

Sample bias

- If the training set is not representative of new data, that is a problem

Simple and transparent models

- Advantages:
  - More likely to be adopted/trusted
  - Less sensitive to changes in data
- Disadantages
  - Worse accuracy

Externalities + Equilibrium Effects

- Sometimes useful to think of decisions on a group level rather individual level
  - E.g. diversity is a measure of the group rather than individuals
- Predictive policing may create a feedback loop
  - More predicted crime => more policing => more detected crime => more predicted crime

Beyond observational measures

- Want to model the causal structures directly, and eliminate consideration of the causes of discrimination
- Requires very strong modelling assumptions

Counterfactual Fairness

Matt Kusner *
The Alan Turing Institute and
University of Warwick
mkusner@turing.ac.uk

Joshua Loftus *
New York University
loftus@nyu.edu

Chris Russell *
The Alan Turing Institute and
University of Surrey
crussell@turing.ac.uk

Ricardo Silva
The Alan Turing Institute and
University College London
ricardo@stats.ucl.ac.uk

Counterfactual Fairness

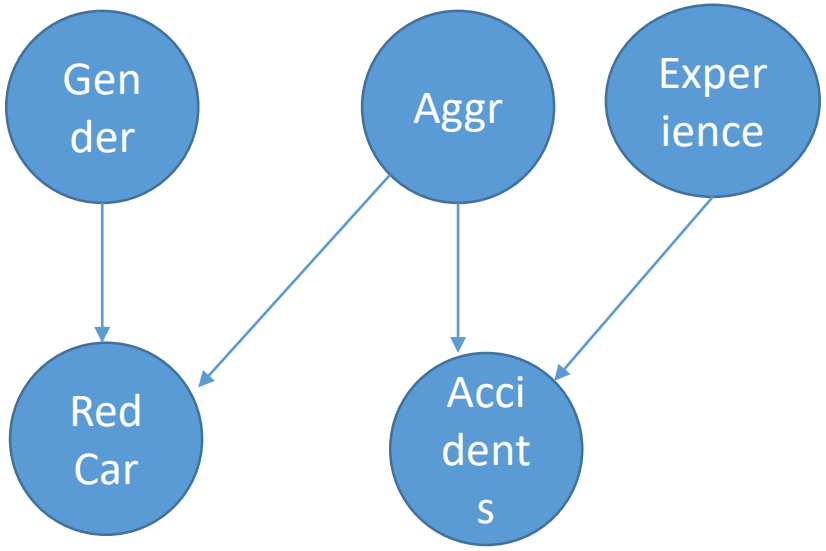
- Require

$$P(C = 1|X = x, do(A = 0)) = P(C = 1|do(A = 1))$$

A is the sensitive characteristic

- Interpretation: treat person with characteristic *A=0* the same as you would treat that person with the characteristic changed to *A=1*
- ***Not*** the same as anti-classification/fairness through unawareness!
  - In general, if A affects X, the probability P(C=1) will change if we apply do(A=0)

Counterfactual Fairness: Red Car



- We are setting insurance rates
- Want to be counterfactually fair w.r.t. gender
- Aggressivness and Gender are both related to driving a red car
- Aggressiveness is related to risk of accidents
- Cannot measure aggressiveness directly

- No direct relationship between Gender and Accidents, but Gender and Aggressiveness both cause driving red cars
- If we use Red Car as a variable (or any other variables that Gender causes, directly or indirectly), our estimates will in general not satisfy counterfactual fairness

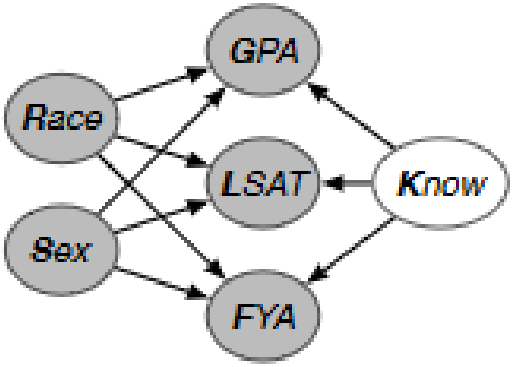
$$P(f(RedCar) = 1|do(Gender = 1), Exp) = P(f(RedCar) = 1|do(Gender = 0), Exp)$$

- (But we can fix this by considering Gender as an input as well)

Counterfactual fairness: recipe

- Idea: in a causal graph, exclude any node that's caused directly or indirectly by the sensitive characteristic
- This implies counterfactual fairness

Predicting The Final Year Average in Law School



Level 2

$$\text{GPA} \sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G),$$
$$\text{LSAT} \sim \text{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S)),$$

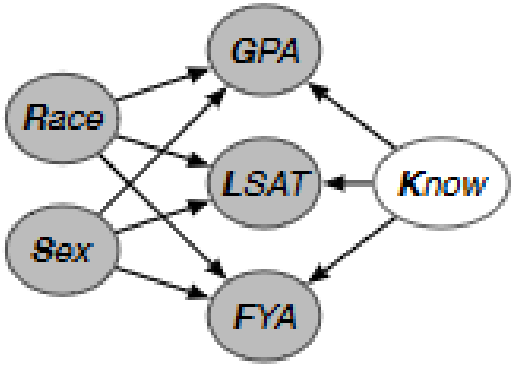
$$\text{FYA} \sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1),$$
$$\text{K} \sim \mathcal{N}(0, 1)$$

- Infer K for each individual
- Now can use K as predictor of success
- Idea: for an individual, the prediction will be the same in the actual world, and in a counterfactual world where they have different demographics
- Requires a causal model of the world

RStan code: https://github.com/mkusner/counterfactual-fairness

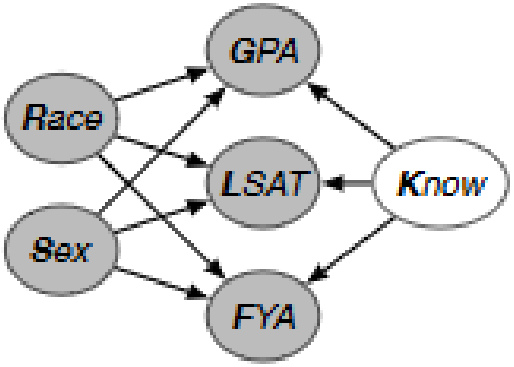
Counterfactual Fairness

- Somewhat analogous to demographic parity: want the same success rate to be the same for individual regardless demographics (basically)



Level 2

Predicting The Final Year Average in Law School: What's wrong with this picture?



Level 2

$$\text{GPA} \sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G),$$
$$\text{LSAT} \sim \text{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S)),$$

$$\text{FYA} \sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1),$$
$$\text{K} \sim \mathcal{N}(0, 1)$$

(Opinionated) Conclusions

- Most fairness measures are not compatible
- Should always consider various fairness criteria when designing/deploying opaque systems
- Observational fairness criteria are all questionable and incompatible – more about posing questions than answering them
- Tension between requiring calibration (same scores mean the same thing for everyone), considering group effects and feedback effects, and considering label and inputs bias
- Causal fairness is the right thing to do *if we understand all the mechanisms that generate all the data*. But we don't

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type [17]) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. While we focus primarily on human-centered machine learning models in the application fields of computer vision and natural language processing, this framework can be used to document any trained machine learning model. To solidify the concept, we provide cards for two supervised models: One trained to detect smiling faces in images, and one trained to detect toxic comments in text. We propose model cards as a step towards the responsible democratization of machine learning and related AI technology, increasing transparency into how well AI technology works. We hope this work encourages those releasing trained machine learning models to accompany model releases with similar detailed evaluation numbers and other relevant documentation.

problematic when models are used in applications that have serious impacts on people's lives, such as in health care [16, 39, 41], employment [3, 15, 27], education [23, 42] and law enforcement [4, 9, 20, 31].

Researchers have discovered systematic biases in commercial machine learning models used for face detection and tracking [6, 11, 43], attribute detection [7], criminal justice [12], toxic comment detection [13], and other applications. However, these systematic errors were only exposed after models were put into use, and negatively affected users reported their experiences. For example, after MIT Media Lab graduate student Joy Buolamwini found that commercial face recognition systems failed to detect her face [6], she collaborated with other researchers to demonstrate the disproportionate errors of computer vision systems on historically marginalized groups in the United States, such as darker-skinned women [7, 38]. In spite of the potential negative effects of such reported biases, documentations accompanying publicly available trained machine learning models (if supplied) provide very little information regarding model performance characteristics, intended use cases, potential pitfalls, or other information to help users evaluate the suitability of these systems to their context. This highlights the need to have detailed documentation accompanying trained machine learning models, including metrics that capture bias, fairness and inclusion considerations.

As a step towards this goal, we propose that released machine learning models be accompanied by short (one to two page) records we call model cards. Model cards (for model reporting) are complements to "Datasheets for Datasets" [21] and similar recently proposed documentation paradigms [5, 26] that report details of the datasets used to train and test machine learning models. We