Fairness in Machine Learning II: Beyond observational measures

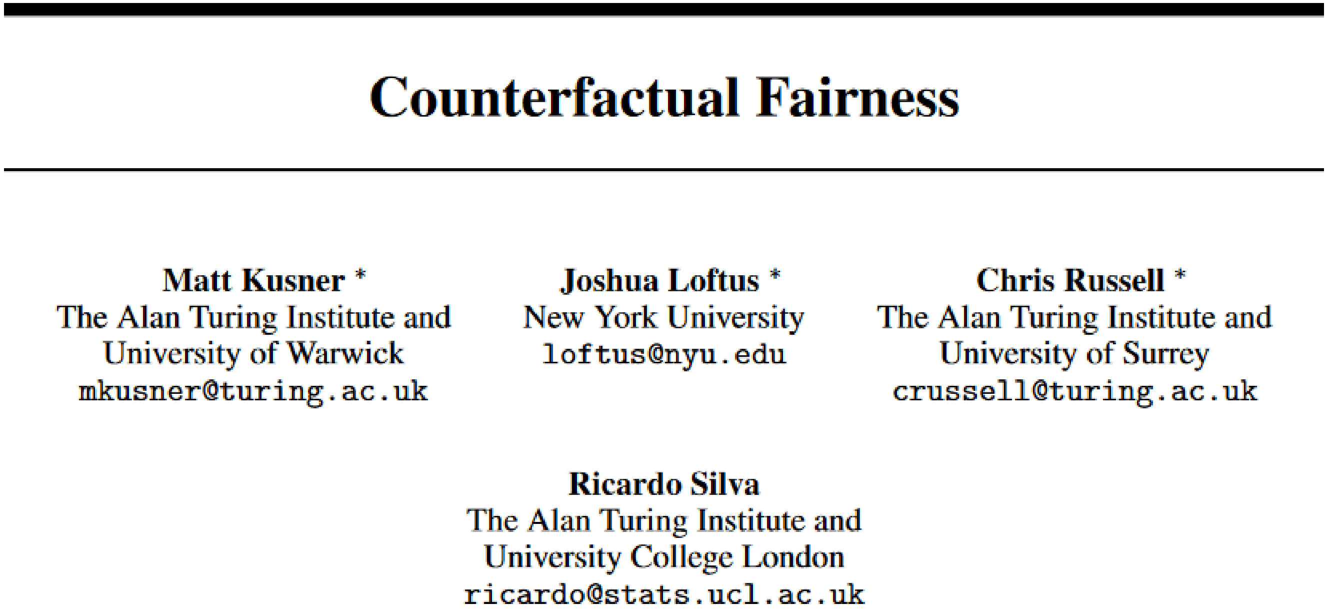
ECE324, Winter 2023

Michael Guerzhoy

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

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

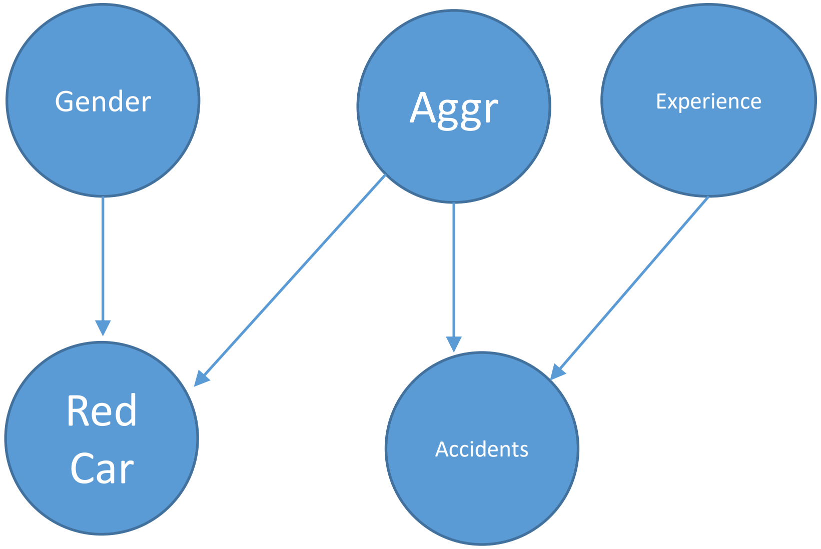
- Require

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

  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

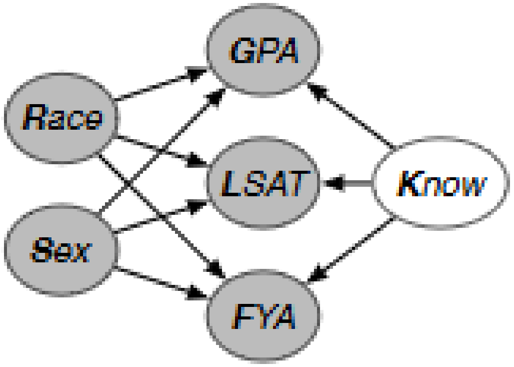
$$P(RedCar = 1 | do(Gender = 1), Exp) \neq P(RedCar = 1 | do(Gender = 0), Exp)$$

- (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

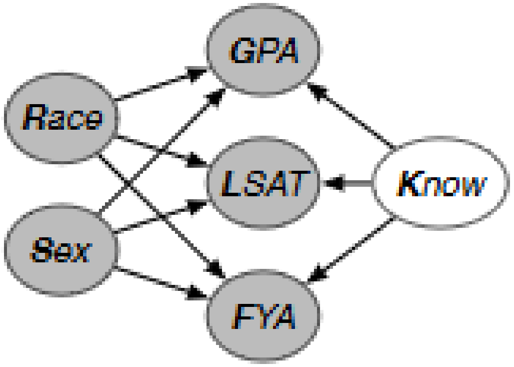
$$\text{GPA} \sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G), \qquad \text{FYA} \sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1),$$

$$\text{LSAT} \sim \text{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S)), \qquad \text{K} \sim \mathcal{N}(0, 1)$$

- 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

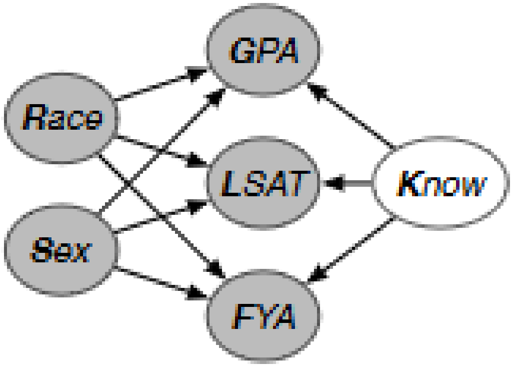
Counterfactual Fairness

- 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?

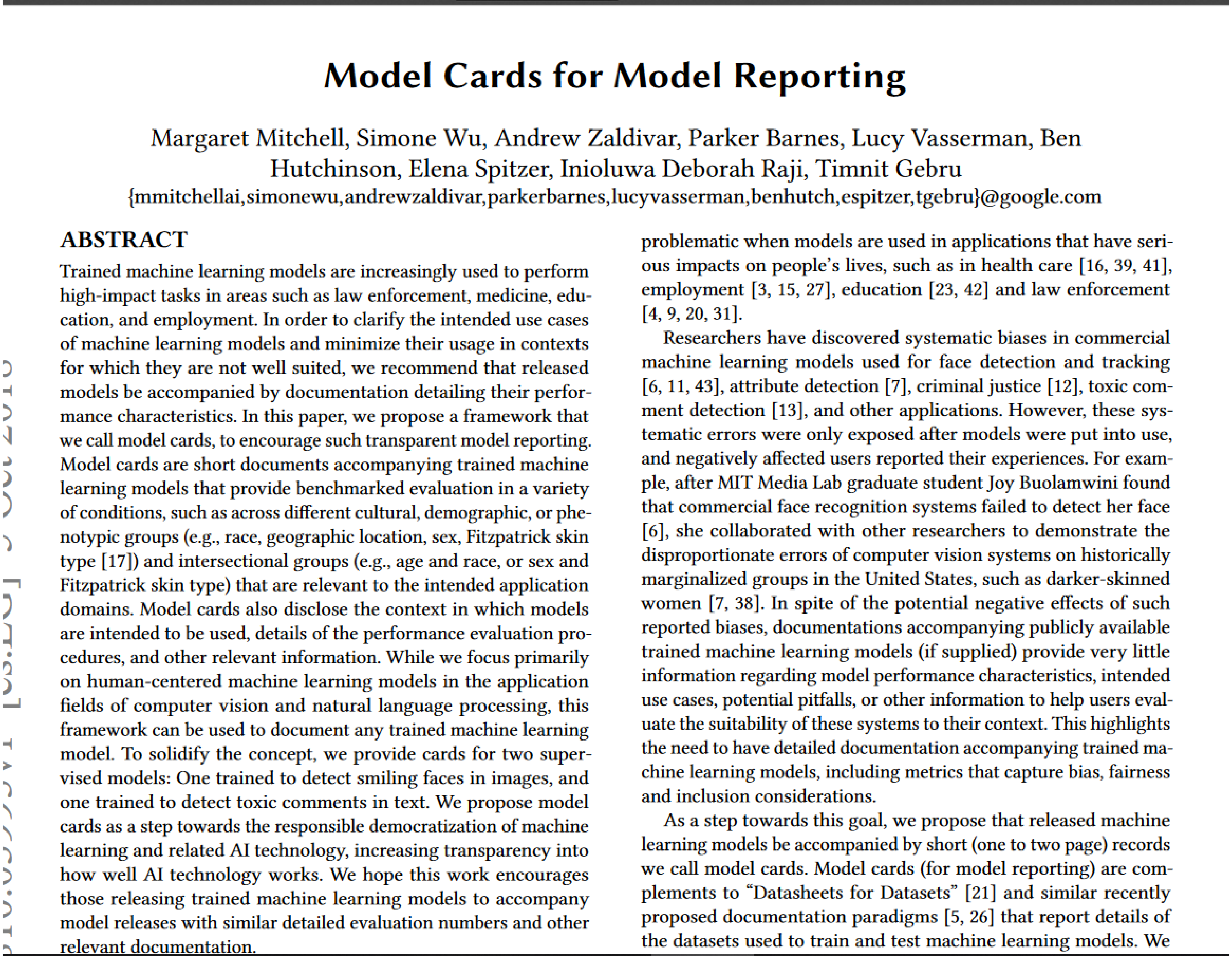


Level 2

$$\text{GPA} \sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G), \qquad \text{FYA} \sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1),$$
$$\text{LSAT} \sim \text{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S)), \qquad \text{K} \sim \mathcal{N}(0, 1)$$

(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

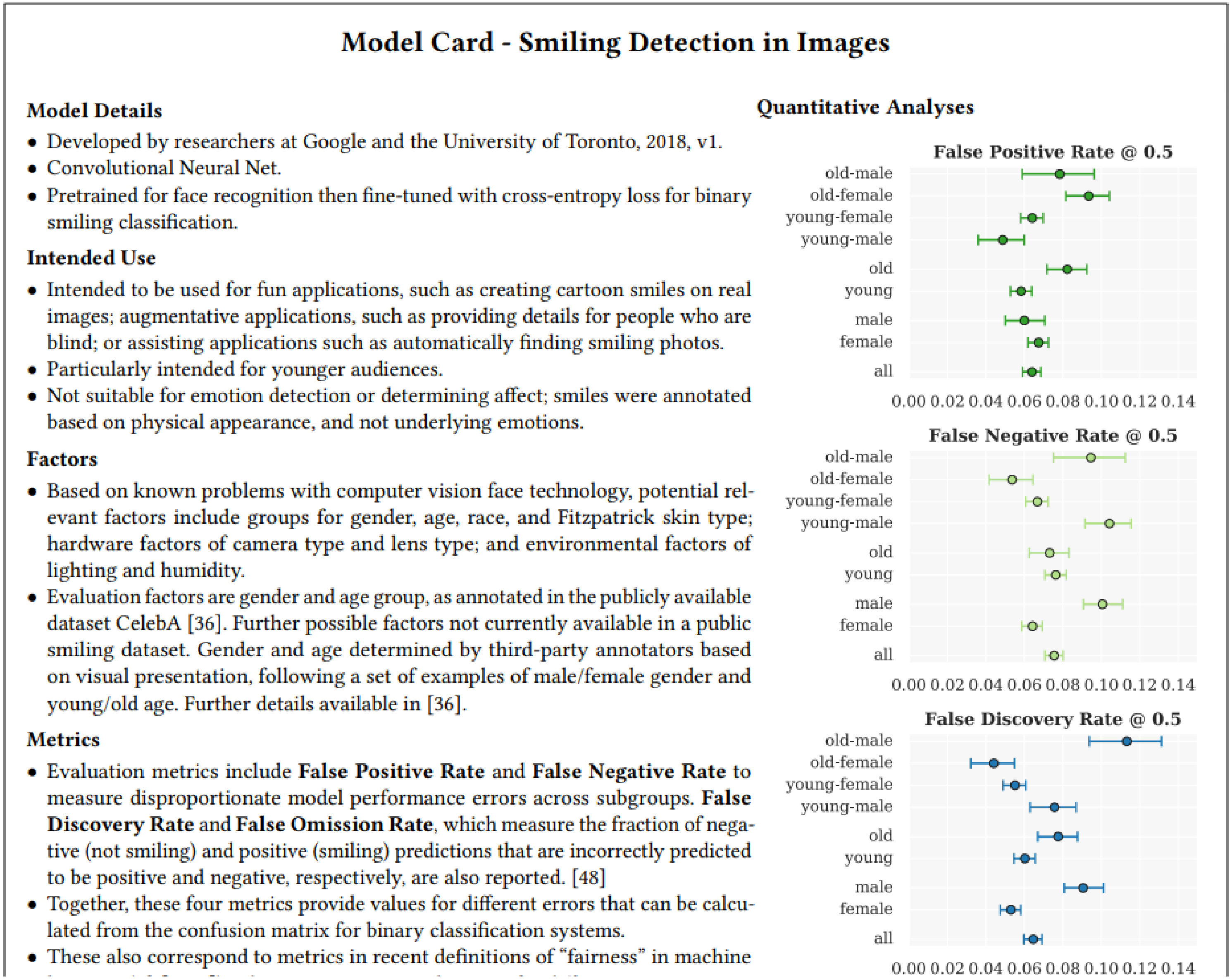
ABSTRACT

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



Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine

Quantitative Analyses