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
deborah.raji@mail.utoronto.ca

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 [15]) 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 artificial intelligence technology, increasing transparency into how well artificial intelligence 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.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Social and professional topics** → *User characteristics*; • **Software and its engineering** → **Use cases**; *Documentation*; *Software evolution*; • **Human-centered computing** → *Walkthrough evaluations*;

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01.
https://doi.org/10.1145/3287560.3287596

KEYWORDS

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

ACM Reference Format:
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287596

1 INTRODUCTION

Currently, there are no standardized documentation procedures to communicate the performance characteristics of trained machine learning (ML) and artificial intelligence (AI) models. This lack of documentation is especially problematic when models are used in applications that have serious impacts on people's lives, such as in health care [14, 42, 44], employment [1, 13, 29], education [23, 45] and law enforcement [2, 7, 20, 34].

Researchers have discovered systematic biases in commercial machine learning models used for face detection and tracking [4, 9, 49], attribute detection [5], criminal justice [10], toxic comment detection [11], 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 [4], 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 [5, 41]. In spite of the potential negative effects of such reported biases, documentation accompanying 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 [3, 28] that report details of the datasets used to train and test machine learning models. Model cards are also similar to the TRIPOD statement proposal in medicine [25]. We provide two example model cards in Section 5: A smiling detection model trained on the CelebA dataset [36] (Figure 2), and a public toxicity detection model [32] (Figure 3). Where Datasheets highlight characteristics of the data feeding into the model, we

focus on trained model characteristics such as the type of model, intended use cases, information about attributes for which model performance may vary, and measures of model performance.

We advocate for measures of model performance that contain quantitative evaluation results to be broken down by individual cultural, demographic, or phenotypic groups, domain-relevant conditions, and intersectional analysis combining two (or more) groups and conditions. In addition to model evaluation results, model cards should detail the motivation behind chosen performance metrics, group definitions, and other relevant factors. Each model card could be accompanied with Datasheets [21], Nutrition Labels [28], Data Statements [3], or Factsheets [27], describing datasets that the model was trained and evaluated on. Model cards provide a way to inform users about what machine learning systems can and cannot do, the types of errors they make, and additional steps that could create more fair and inclusive outcomes with the technology.

2 BACKGROUND

Many mature industries have developed standardized methods of benchmarking various systems under different conditions. For example, as noted in [21], the electronic hardware industry provides datasheets with detailed characterizations of components' performances under different test conditions. By contrast, despite the broad reach and impact of machine learning models, there are no standard stress tests that are performed on machine learning based systems, nor standardized formats to report the results of these tests. Recently, researchers have proposed standardized forms of communicating characteristics of datasets used in machine learning [3, 21, 28] to help users understand the context in which the datasets should be used. We focus on the complementary task for machine learning models, proposing a standardized method to evaluate the performance of human-centric models: Disaggregated by unitary and intersectional groups such as cultural, demographic, or phenotypic population groups. A framework that we refer to as "Model Cards" can present such evaluation supplemented with additional considerations such as intended use.

Outside of machine learning, the need for population-based reporting of outcomes as suggested here has become increasingly evident. For example, in vehicular crash tests, dummies with prototypical female characteristics were only introduced after researchers discovered that women were more likely than men to suffer serious head injuries in real-world side impacts [18]. Similarly, drugs developed based on results of clinical trials with exclusively male participants have led to overdosing in women [17, 50]. In 1998, the U.S. Food and Drug Administration mandated that clinical trial results be disaggregated by groups such as age, race and gender [16].

While population-based analyses of errors and successes can be provided for unitary groups such as "men", "women", and "non-binary" gender groups, they should also be provided intersectionally, looking at two or more characteristics such as gender and age simultaneously. Intersectional analyses are linked to intersectionality theory, which describes how discrete experiences associated with characteristics like race or gender in isolation do not accurately reflect their interaction [8]. Kimberlé Crenshaw, who pioneered intersectional research in critical race theory, discusses the story of Emma DeGraffenreid, who was part of a failed lawsuit against General Motors in 1976, claiming that the company's hiring practices discriminated against Black women. In their court opinion, the judges noted that since General Motors hired many women for secretarial positions, and many Black people for factory roles, they could not have discriminated against Black women. However, what the courts failed to see was that only White women were hired into secretarial positions and only Black men were hired into factory roles. Thus, Black women like Emma DeGraffenreid had no chance of being employed at General Motors. This example highlights the importance of intersectional analyses: empirical analyses that emphasize the interaction between various demographic categories including race, gender, and age.

Before further discussing the details of the model card, it is important to note that at least two of the three characteristics discussed so far, race and gender, are socially sensitive. Although analyzing models by race and gender may follow from intersectionality theory, how "ground truth" race or gender categories should be labeled in a dataset, and whether or not datasets should be labeled with these categories at all, is not always clear. This issue is further confounded by the complex relationship between gender and sex. When using cultural identity categories such as race and gender to subdivide analyses, and depending on the context, we recommend either using datasets with self-identified labels or with labels clearly designated as *perceived* (rather than self-identified). When this is not possible, datasets of public figures with known public identity labels may be useful. Further research is necessary to expand how groups may be defined, for example, by automatically discovering groups with similarities in the evaluation datasets.

3 MOTIVATION

As the use of machine learning technology has rapidly increased, so too have reports of errors and failures. Despite the potentially serious repercussions of these errors, those looking to use trained machine learning models in a particular context have no way of understanding the systematic impacts of these models before deploying them.

The proposal of "Model Cards" specifically aims to standardize ethical practice and reporting - allowing stakeholders to compare candidate models for deployment across not only traditional evaluation metrics but also along the axes of ethical, inclusive, and fair considerations. This goes further than current solutions to aid stakeholders in different contexts. For example, to aid policy makers and regulators on questions to ask of a model, and known benchmarks around the suitability of a model in a given setting.

Model reporting will hold different meaning to those involved in different aspects of model development, deployment, and use. Below, we outline a few use cases for different stakeholders:

- **ML and AI practitioners** can better understand how well the model might work for the intended use cases and track its performance over time.
- **Model developers** can compare the model's results to other models in the same space, and make decisions about training their own system.
- **Software developers** working on products that use the model's predictions can inform their design and implementation decisions.

- **Policymakers** can understand how a machine learning system may fail or succeed in ways that impact people.
- **Organizations** can inform decisions about adopting technology that incorporates machine learning.
- **ML-knowledgeable individuals** can be informed on different options for fine-tuning, model combination, or additional rules and constraints to help curate models for intended use cases without requiring technical expertise.
- **Impacted individuals** who may experience effects from a model can better understand how it works or use information in the card to pursue remedies.

Not only does this practice improve model understanding and help to standardize decision making processes for invested stakeholders, but it also encourages forward-looking model analysis techniques. For example, slicing the evaluation across groups functions to highlight errors that may fall disproportionately on some groups of people, and accords with many recent notions of mathematical fairness (discussed further in the example model card in Figure 2). Including group analysis as part of the reporting procedure prepares stakeholders to begin to gauge the fairness and inclusion of future outcomes of the machine learning system. Thus, in addition to supporting decision-making processes for determining the suitability of a given machine learning model in a particular context, model reporting is an approach for responsible transparent and accountable practices in machine learning.

People and organizations releasing models may be additionally incentivized to provide model card details because it helps potential users of the models to be better informed on which models are best for their specific purposes. If model card reporting becomes standard, potential users can compare and contrast different models in a well-informed way. Results on several different evaluation datasets will additionally aid potential users, although evaluation datasets suitable for disaggregated evaluation are not yet common. Future research could include creating robust evaluation datasets and protocols for the types of disaggregated evaluation we advocate for in this work, for example, by including differential privacy mechanisms [12] so that individuals in the testing set cannot be uniquely identified by their characteristics.

4 MODEL CARD SECTIONS

Model cards serve to disclose information about a trained machine learning model. This includes how it was built, what assumptions were made during its development, what type of model behavior different cultural, demographic, or phenotypic population groups may experience, and an evaluation of how well the model performs with respect to those groups. Here, we propose a set of sections that a model card should have, and details that can inform the stakeholders discussed in Section 3. A summary of all suggested sections is provided in Figure 1.

The proposed set of sections below are intended to provide relevant details to consider, but are not intended to be complete or exhaustive, and may be tailored depending on the model, context, and stakeholders. Additional details may include, for example, interpretability approaches, such as saliency maps, TCAV [33], and Path-Integrated Gradients [38, 43]); stakeholder-relevant explanations (e.g., informed by a careful consideration of philosophical, psychological, and other factors concerning what is as a good explanation in different contexts [22]); and privacy approaches used in model training and serving.

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 1: Summary of model card sections and suggested prompts for each.

4.1 Model Details

This section of the model card should serve to answer basic questions regarding the model version, type and other details.

Person or organization developing model: What person or organization developed the model? This can be used by all stakeholders to infer details pertaining to model development and potential

conflicts of interest.
Model date: When was the model developed? This is useful for all stakeholders to become further informed on what techniques and data sources were likely to be available during model development.
Model version: Which version of the model is it, and how does it differ from previous versions? This is useful for all stakeholders to track whether the model is the latest version, associate known bugs to the correct model versions, and aid in model comparisons.
Model type: What type of model is it? This includes basic model architecture details, such as whether it is a Naive Bayes classifier, a Convolutional Neural Network, etc. This is likely to be particularly relevant for software and model developers, as well as individuals knowledgeable about machine learning, to highlight what kinds of assumptions are encoded in the system.
Paper or other resource for more information: Where can resources for more information be found?
Citation details: How should the model be cited?
License: License information can be provided.
Feedback on the model: E.g., what is an email address that people may write to for further information?

There are cases where some of this information may be sensitive. For example, the amount of detail corporations choose to disclose might be different from academic research groups. This section should not be seen as a requirement to compromise private information or reveal proprietary training techniques; rather, a place to disclose basic decisions and facts about the model that the organization can share with the broader community in order to better inform on what the model represents.

4.2 Intended Use

This section should allow readers to quickly grasp what the model should and should not be used for, and why it was created. It can also help frame the statistical analysis presented in the rest of the card, including a short description of the user(s), use-case(s), and context(s) for which the model was originally developed. Possible information includes:
Primary intended uses: This section details whether the model was developed with general or specific tasks in mind (e.g., plant recognition worldwide or in the Pacific Northwest). The use cases may be as broadly or narrowly defined as the developers intend. For example, if the model was built simply to label images, then this task should be indicated as the primary intended use case.
Primary intended users: For example, was the model developed for entertainment purposes, for hobbyists, or enterprise solutions? This helps users gain insight into how robust the model may be to different kinds of inputs.
Out-of-scope uses: Here, the model card should highlight technology that the model might easily be confused with, or related contexts that users could try to apply the model to. This section may provide an opportunity to recommend a related or similar model that was designed to better meet that particular need, where possible. This section is inspired by warning labels on food and toys, and similar disclaimers presented in electronic datasheets. Examples include "not for use on text examples shorter than 100 tokens" or "for use on black-and-white images only; please consider our research group's full-color-image classifier for color images."

4.3 Factors

Model cards ideally provide a summary of model performance across a variety of relevant factors including *groups*, *instrumentation*, and *environments*. We briefly describe each of these factors and their relevance followed by the corresponding prompts in the model card.

4.3.1 Groups. "Groups" refers to distinct categories with similar characteristics that are present in the evaluation data instances. For human-centric machine learning models, "groups" are people who share one or multiple characteristics. Intersectional model analysis for human-centric models is inspired by the sociological concept of intersectionality, which explores how an individual's identity and experiences are shaped not just by unitary personal characteristics – such as race, gender, sexual orientation or health – but instead by a complex combination of many factors. These characteristics, which include but are not limited to cultural, demographic and phenotypic categories, are important to consider when evaluating machine learning models. Determining which groups to include in an intersectional analysis requires examining the intended use of the model and the context under which it may be deployed. Depending on the situation, certain groups may be more vulnerable than others to unjust or prejudicial treatment.

For human-centric computer vision models, the visual presentation of age, gender, and Fitzpatrick skin type [15] may be relevant. However, this must be balanced with the goal of preserving the privacy of individuals. As such, collaboration with policy, privacy, and legal experts is necessary in order to ascertain which groups may be responsibly inferred, and how that information should be stored and accessed (for example, using differential privacy [12]).

Details pertaining to groups, including who annotated the training and evaluation datasets, instructions and compensation given to annotators, and inter-annotator agreement, should be provided as part of the data documentation made available with the dataset. See [3, 21, 28] for more details.

4.3.2 Instrumentation. In addition to groups, the performance of a model can vary depending on what instruments were used to capture the input to the model. For example, a face detection model may perform differently depending on the camera's hardware and software, including lens, image stabilization, high dynamic range techniques, and background blurring for portrait mode. Performance may also vary across real or simulated traditional camera settings such as aperture, shutter speed and ISO. Similarly, video and audio input will be dependent on the choice of recording instruments and their settings.

4.3.3 Environment. A further factor affecting model performance is the environment in which it is deployed. For example, face detection systems are often less accurate under low lighting conditions or when the air is humid [51]. Specifications across different lighting and moisture conditions would help users understand the impacts of these environmental factors on model performance.

4.3.4 Card Prompts. We propose that the Factors section of model cards expands on two prompts:

Relevant factors: What are foreseeable salient factors for which model performance may vary, and how were these determined?
Evaluation factors: Which factors are being reported, and why were these chosen? If the relevant factors and evaluation factors are different, why? For example, while Fitzpatrick skin type is a relevant factor for face detection, an evaluation dataset annotated by skin type might not be available until reporting model performance across groups becomes standard practice.

4.4 Metrics

The appropriate metrics to feature in a model card depend on the type of model that is being tested. For example, classification systems in which the primary output is a class label differ significantly from systems whose primary output is a score. In all cases, the reported metrics should be determined based on the model's structure and intended use. Details for this section include:
Model performance measures: What measures of model performance are being reported, and why were they selected over other measures of model performance?
Decision thresholds: If decision thresholds are used, what are they, and why were those decision thresholds chosen? When the model card is presented in a digital format, a threshold slider should ideally be available to view performance parameters across various decision thresholds.
Approaches to uncertainty and variability: How are the measurements and estimations of these metrics calculated? For example, this may include standard deviation, variance, confidence intervals, or KL divergence. Details of how these values are approximated should also be included (e.g., average of 5 runs, 10-fold cross-validation).

4.4.1 Classification systems. For classification systems, the error types that can be derived from a confusion matrix are *false positive rate*, *false negative rate*, *false discovery rate*, and *false omission rate*. We note that the relative importance of each of these metrics is system, product and context dependent.

For example, in a surveillance scenario, surveillors may value a low false negative rate (or the rate at which the surveillance system fails to detect a person or an object when it should have). On the other hand, those being surveilled may value a low false positive rate (or the rate at which the surveillance system detects a person or an object when it should not have). We recommend listing all values and providing context about which were prioritized during development and why.

Equality between some of the different confusion matrix metrics is equivalent to some definitions of fairness. For example, equal false negative rates across groups is equivalent to fulfilling Equality of Opportunity, and equal false negative and false positive rates across groups is equivalent to fulfilling Equality of Odds [26].

4.4.2 Score-based analyses. For score-based systems such as pricing models and risk assessment algorithms, describing differences in the distribution of measured metrics across groups may be helpful. For example, reporting measures of central tendency such as the mode, median and mean, as well as measures of dispersion or variation such as the range, quartiles, absolute deviation, variance and standard deviation could facilitate the statistical commentary necessary to make more informed decisions about model development. A model card could even extend beyond these summary statistics to reveal other measures of differences between distributions such as cross entropy, perplexity, KL divergence and pinned area under the curve (pinned AUC) [11].

There are a number of applications that do not appear to be score-based at first glance, but can be considered as such for the purposes of intersectional analysis. For instance, a model card for a translation system could compare BLEU scores [40] across demographic groups, and a model card for a speech recognition system could compare word-error rates. Although the primary outputs of these systems are not scores, looking at the score differences between populations may yield meaningful insights since comparing raw inputs quickly grows too complex.

4.4.3 Confidence. Performance metrics that are disaggregated by various combinations of instrumentation, environments and groups makes it especially important to understand the confidence intervals for the reported metrics. Confidence intervals for metrics derived from confusion matrices can be calculated by treating the matrices as probabilistic models of system performance [24].

4.5 Evaluation Data

All referenced datasets would ideally point to any set of documents that provide visibility into the source and composition of the dataset. Evaluation datasets should include datasets that are publicly available for third-party use. These could be existing datasets or new ones provided alongside the model card analyses to enable further benchmarking. Potential details include:
Datasets: What datasets were used to evaluate the model?
Motivation: Why were these datasets chosen?
Preprocessing: How was the data preprocessed for evaluation (e.g., tokenization of sentences, cropping of images, any filtering such as dropping images without faces)?

To ensure that model cards are statistically accurate and verifiable, the evaluation datasets should not only be representative of the model's typical use cases but also anticipated test scenarios and challenging cases. For instance, if a model is intended for use in a workplace that is phenotypically and demographically homogeneous, and trained on a dataset that is representative of the expected use case, it may be valuable to evaluate that model on two evaluation sets: one that matches the workplace's population, and another set that contains individuals that might be more challenging for the model (such as children, the elderly, and people from outside the typical workplace population). This methodology can highlight pathological issues that may not be evident in more routine testing.

It is often difficult to find datasets that represent populations outside of the initial domain used in training. In some of these situations, synthetically generated datasets may provide representation for use cases that would otherwise go unevaluated [35]. Section 5.2 provides an example of including synthetic data in the model evaluation dataset.

4.6 Training Data

Ideally, the model card would contain as much information about the training data as the evaluation data. However, there might be cases where it is not feasible to provide this level of detailed information about the training data. For example, the data may be proprietary, or require a non-disclosure agreement. In these cases, we advocate for basic details about the distributions over groups in the data, as well as any other details that could inform stakeholders on the kinds of biases the model may have encoded.

4.7 Quantitative Analyses

Quantitative analyses should be *disaggregated*, that is, broken down by the chosen factors. Quantitative analyses should provide the results of evaluating the model according to the chosen metrics, providing confidence interval values when possible. Parity on the different metrics across disaggregated population subgroups corresponds to how *fairness* is often defined [37, 48]. Quantitative analyses should demonstrate the metric variation (e.g., with error bars), as discussed in Section 4.4 and visualized in Figure 2.
The disaggregated evaluation includes:
Unitary results: How did the model perform with respect to each factor?
Intersectional results: How did the model perform with respect to the intersection of evaluated factors?

4.8 Ethical Considerations

This section is intended to demonstrate the ethical considerations that went into model development, surfacing ethical challenges and solutions to stakeholders. Ethical analysis does not always lead to precise solutions, but the process of ethical contemplation is worthwhile to inform on responsible practices and next steps in future work.

While there are many frameworks for ethical decision-making in technology that can be adapted here [19, 30, 46], the following are specific questions you may want to explore in this section:
Data: Does the model use any sensitive data (e.g., protected classes)?
Human life: Is the model intended to inform decisions about matters central to human life or flourishing – e.g., health or safety? Or could it be used in such a way?
Mitigations: What risk mitigation strategies were used during model development?
Risks and harms: What risks may be present in model usage? Try to identify the potential recipients, likelihood, and magnitude of harms. If these cannot be determined, note that they were considered but remain unknown.
Use cases: Are there any known model use cases that are especially fraught? This may connect directly to the intended use section of the model card.

If possible, this section should also include any additional ethical considerations that went into model development, for example, review by an external board, or testing with a specific community.

4.9 Caveats and Recommendations

This section should list additional concerns that were not covered in the previous sections. For example, did the results suggest any further testing? Were there any relevant groups that were not represented in the evaluation dataset? Are there additional recommendations for model use? What are the ideal characteristics of an evaluation dataset for this model?

5 EXAMPLES

We present worked examples of model cards for two models: an image-based classification system and a text-based scoring system.

5.1 Smiling Classifier

To show an example of a model card for an image classification problem, we use the public CelebA dataset [36] to examine the performance of a trained "smiling" classifier across both age and gender categories. Figure 2 shows our prototype.

These results demonstrate a few potential issues. For example, the false discovery rate on older men is much higher than that for other groups. This means that many predictions incorrectly classify older men as smiling when they are not. On the other hand, men (in aggregate) have a higher false negative rate, meaning that many of the men that are in fact smiling in the photos are incorrectly classified as not smiling.

The results of these analyses give insight into contexts the model might not be best suited for. For example, it may not be advisable to apply the model on a diverse group of audiences, and it may be the most useful when detecting the presence of a smile is more important than detecting its absence (for example, in an application that automatically finds 'fun moments' in images). Additional fine-tuning, for example, with images of older men, may help create a more balanced performance across groups.

5.2 Toxicity Scoring

Our second example provides a model card for Perspective API's TOXICITY classifier built to detect 'toxicity' in text [32], and is presented in Figure 3. To evaluate the model, we use an intersectional version of the open source, synthetically created Identity Phrase Templates test set published in [11]. We show two versions of the quantitative analysis: one for TOXICITY v. 1, the initial version of the this model, and one for TOXICITY v. 5, the latest version.

This model card highlights the drastic ways that models can change over time, and the importance of having a model card that is updated with each new model release. TOXICITY v. 1 has low performance for several terms, especially "lesbian", "gay", and "homosexual". This is consistent with what some users of the initial TOXICITY model found, as reported by the team behind Perspective API in [47]. Also in [47], the Perspective API team shares the bias mitigation techniques they applied to the TOXICITY v. 1 model, in order to create the more equitable performance in TOXICITY v. 5. By making model cards a standard part of API launches, teams like the Perspective API team may be able to find and mitigate some of these biases earlier.

6 DISCUSSION & FUTURE WORK

We have proposed frameworks called model cards for reporting information about what a trained machine learning model is and how well it works. Model cards include information about the context of the model, as well as model performance results disaggregated by different unitary and intersectional population groups. Model

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 learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

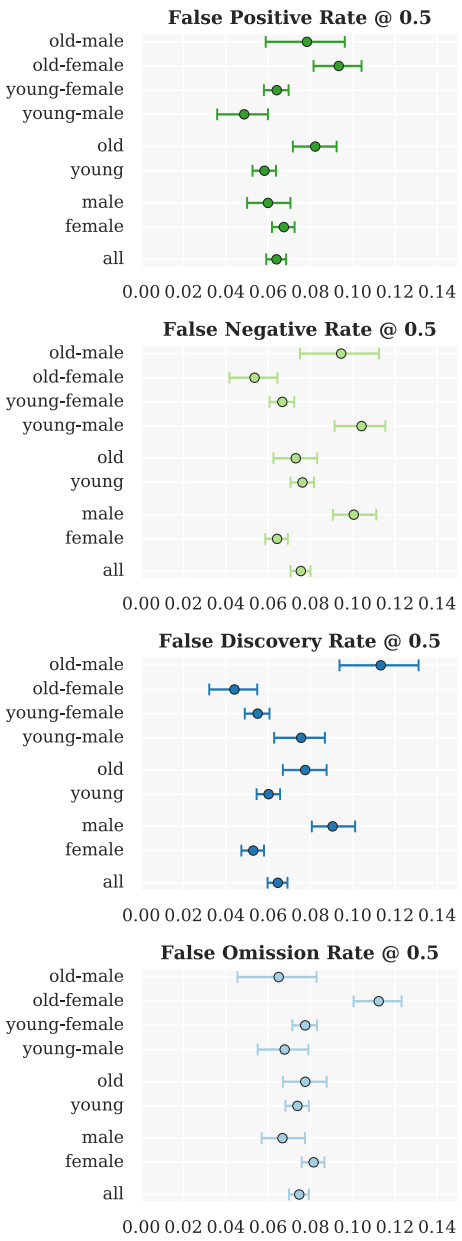
Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses



Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses

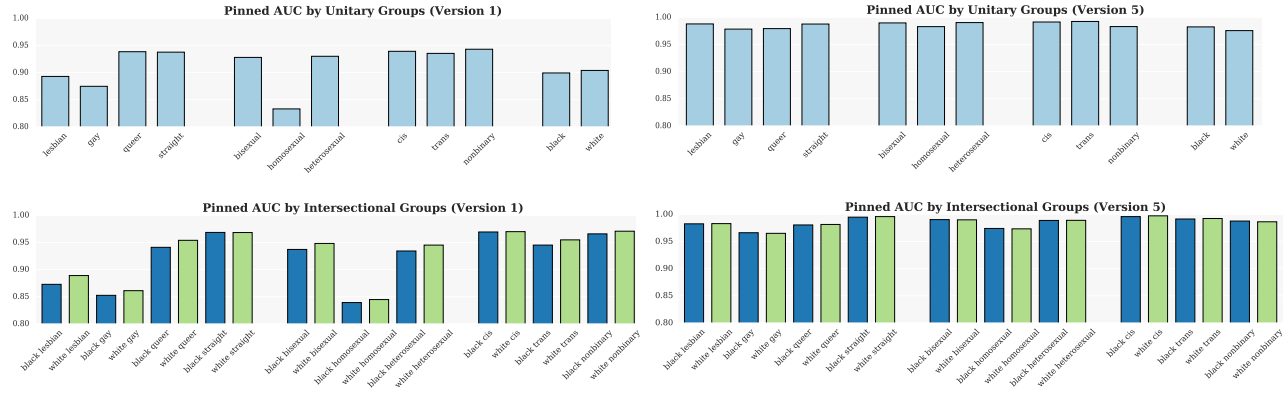


Figure 3: Example Model Card for two versions of Perspective API's toxicity detector.

cards are intended to accompany a model after careful review has determined that the foreseeable benefits outweigh the foreseeable risks in the model's use or release.

To demonstrate the use of model cards in practice, we have provided two examples: A model card for a smiling classifier tested on the CelebA dataset, and a model card for a public toxicity detector tested on the Identity Phrase Templates dataset. We report confusion matrix metrics for the smile classifier and Pinned AUC for the toxicity detector, along with model details, intended use, pointers to information about training and evaluation data, ethical considerations, and further caveats and recommendations.

The framework presented here is intended to be general enough to be applicable across different institutions, contexts, and stakeholders. It also is suitable for recently proposed requirements for analysis of algorithmic decision systems in critical social institutions, for example, for models used in determining government benefits, employment evaluations, criminal risk assessment, and criminal DNA analysis [39].

Model cards are just one approach to increasing transparency between developers, users, and stakeholders of machine learning models and systems. They are designed to be flexible in both scope and specificity in order to accommodate the wide variety of machine learning model types and potential use cases. Therefore the usefulness and accuracy of a model card relies on the integrity of the creator(s) of the card itself. It seems unlikely, at least in the near term, that model cards could be standardized or formalized to a degree needed to prevent misleading representations of model results (whether intended or unintended). It is therefore important to consider model cards as one transparency tool among many, which could include, for example, algorithmic auditing by third-parties (both quantitative and qualitative), "adversarial testing" by technical and non-technical analysts, and more inclusive user feedback mechanisms. Future work will aim to refine the methodology of creating model cards by studying how model information is interpreted and used by different stakeholders. Researchers should also explore how model cards can strengthen and complement other transparency methods

7 ACKNOWLEDGEMENTS

Thank you to Joy Buolamwini, Shalini Ananda and Shira Mitchell for invaluable conversations and insight.

REFERENCES

[1] Avrio AI. 2018. Avrio AI: AI Talent Platform. (2018). https://www.goavrio.com/
[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[3] Emily M. Bender and Batya Friedman. 2018. "Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science". *Transactions of the ACL (TACL)* (2018).
[4] Joy Buolamwini. 2016. How I'm fighting Bias in Algorithms. (2016). https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms#t-63664
[5] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html
[6] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[7] Federal Trade Commission. 2016. Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues. (2016). https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report
[8] Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *U. Chi. Legal F.* (1989), 139.
[9] Black Desi. 2009. HP computers are racist. (2009). https://www.youtube.com/watch?v=t4DT3tQqgRM
[10] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. (2016). https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html
[11] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2018).
[12] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *Theory and Applications of Models of Computation*, Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–19.
[13] Entelo. 2018. Recruitment Software | Entelo. (2018). https://www.entelo.com/
[14] Daniel Faggella. 2018. Follow the Data: Deep Learning Leads the Transformation of Enterprise - A Conversation with Naveen Rao. (2018).
[15] Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* 124, 6 (1988), 869–871.
[16] Food and Drug Administration. 1989. Guidance for the Study of Drugs Likely to Be Used in the Elderly. (1989).
[17] U.S. Food and Drug Administration. 2013. FDA Drug Safety Communication: Risk of next-morning impairment after use of insomnia drugs; FDA requires lower recommended doses for certain drugs containing zolpidem (Ambien, Ambien CR, Edluar, and Zolpimist). (2013). https://web.archive.org/web/20170428150213/https://www.fda.gov/drugs/drugsafety/ucm352085.htm
[18] IIHS (Insurance Institute for Highway Safety: Highway Loss Data Institute). 2003. Special Issue: Side Impact Crashworthiness. *Status Report* 38, 7 (2003).
[19] Institute for the Future, Omidyar Network's Tech, and Society Solutions Lab. 2018. Ethical OS. (2018). https://ethicalos.org/
[20] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. 2016. The Perpetual Line-Up. (2016). https://www.perpetuallineup.org/
[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR* abs/1803.09010 (2018). http://arxiv.org/abs/1803.09010
[22] Google. 2018. Responsible AI Practices. (2018). https://ai.google/education/responsible-ai-practices
[23] Gooru. 2018. Navigator for Teachers. (2018). http://gooru.org/about/teachers
[24] Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval*. Springer, 345–359.
[25] Collins GS, Reitsma JB, Altman DG, and Moons KM. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. *Annals of Internal Medicine* 162, 1 (2015), 55–63. DOI: http://dx.doi.org/10.7326/M14-0697
[26] Moritz Hardt, Eric Price, , and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 3315–3323. http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf
[27] Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, and Kush R. Varshney. 2018. Increasing Trust in AI Services through Supplier's Declarations of Conformity. *CoRR* abs/1808.07261 (2018).
[28] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *CoRR* abs/1805.03677 (2018). http://arxiv.org/abs/1805.03677
[29] Ideal. 2018. AI For Recruiting Software | Talent Intelligence for High-Volume Hiring. (2018). https://ideal.com/
[30] DrivenData Inc. 2018. An Ethics Checklist for Data Scientists. (2018). http://deon.drivendata.org/
[31] Jigsaw. 2017. Conversation AI Research. (2017). https://conversationai.github.io/
[32] Jigsaw. 2017. Perspective API. (2017). https://www.perspectiveapi.com/
[33] B. Kim, Wattenberg M., J. Gilmer, Cai C., Wexler J., , F. Viegas, and R. Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) . *ICML* (2018).
[34] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6 (2012), 1789–1801. DOI: http://dx.doi.org/10.1109/TIFS.2012.2214212
[35] Der-Chiang Li, Susan C Hu, Liang-Sian Lin, and Chun-Wu Yeh. 2017. Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets. *PloS one* 12, 8 (2017), e0181853.

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
[37] Shira Mitchell, Eric Potash, and Solon Barocas. 2018. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *arXiv:1811.07867* (2018).
[38] Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the Model Understand the Question? *Proceedings of the Association for Computational Linguistics* (2018).
[39] AI Now. 2018. Litigating Algorithms: Challenging Government Use Of Algorithmic Decision Systems. AI Now Institute.
[40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.
[41] Inioluwa Raji. 2018. Black Panther Face Scorecard: Wakandans Under the Coded Gaze of AI. (2018).
[42] Microsoft Research. 2018. Project InnerEye - Medical Imaging AI to Empower Clinicians. (2018). https://www.microsoft.com/en-us/research/project/medical-image-analysis/
[43] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. PMLR, Sydney, Australia.
[44] Digital Reasoning Systems. 2018. AI-Enabled Cancer Software | Healthcare AI : Digital Reasoning. (2018). https://digitalreasoning.com/solutions/healthcare/
[45] Turnitin. 2018. Revision Assistant. (2018). http://turnitin.com/en_us/what-we-offer/revision-assistant
[46] Shannon Vallor, Brian Green, and Irina Raicu. 2018. Ethics in Technology Practice: An Overview. (22 6 2018). https://www.scu.edu/ethics-in-technology-practice/overview-of-ethics-in-tech-practice/
[47] Lucy Vasserman, John Li, CJ Adams, and Lucas Dixon. 2018. Unintended bias and names of frequently targeted groups. *Medium* (2018). https://medium.com/the-false-positive/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23
[48] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. (2018).
[49] Joz Wang. 2010. Flickr Image. (2010). https://www.flickr.com/photos/jozjozjoz/3529106844
[50] Amy Westervelt. 2018. The medical research gender gap: how excluding women from clinical trials is hurting our health. (2018).
[51] Mingyuan Zhou, Haiting Lin, S Susan Young, and Jingyi Yu. 2018. Hybrid sensing face detection and registration for low-light and unconstrained conditions. *Applied optics* 57, 1 (2018), 69–78.