CITY OF OAKLAND

CITY HALL • 1 FRANK H. OGAWA PLAZA • OAKLAND, CALIFORNIA 94612

REBECCA KAPLAN
Council President
atlarge@oaklandnet.com

(510) 238-7008
FAX: (510) 238-6910
TDD: (510) 839-6451

Date: June 6, 2019
To: Members of City Council and Members of the Public
From: Council President Kaplan
Re: **ORDINANCE AMENDING OAKLAND MUNICIPAL CODE CHAPTER 9.64 TO PROHIBIT THE CITY OF OAKLAND FROM ACQUIRING AND/OR USING FACE RECOGNITION TECHNOLOGY**

Dear Colleagues on the City Council and Members of the Public,

We submit to you the attached Ordinance for your kind consideration.

RECOMMENDATION/ ACTION REQUESTED

Council President Kaplan requests that the City Council adopt language to define "Face Recognition Technology" as "an automated or semi-automated process that assists in identifying or verifying an individual based on an individual's face."

In addition, there is a request to add new language to our existing surveillance technology code to prohibit the City from acquiring, obtaining, retaining, requesting, or accessing Face Recognition Software. It also allows for City staff to not be penalized if they inadvertently or unintentionally receive information obtained from Face Recognition Technology.

BACKGROUND/LEGISLATIVE HISTORY

On January 19, 2016, the City Council adopted Ordinance NO.13349 which created and defined the duties of the Privacy Advisory Commission (PAC) which "provide(s) advice and technical assistance on best practices to protect citizen privacy rights in connection with the City's purchase and use of surveillance equipment and other technology that collects or stores citizen data." The PAC reviews, considers, and drafts legislation related surveillance equipment usage. On May 2, 2019, the PAC voted on a proposed amendment to prohibit use of face recognition technology. This proposal passed unanimously. This proposal was then brought to Oakland City Council President Kaplan's office by the Chairperson of the PAC, Brian Hofer, as a request to introduce as an ordinance.

ANALYSIS

Face recognition technology runs the risk of making Oakland residents less safe as the misidentification of individuals could lead to the misuse of force, false incarceration, and minority-based persecution. The City of Oakland should reject the use of this flawed technology on the following basis: 1) systems rely on biased datasets with high levels of inaccuracy; 2) a lack of standards around the use and sharing of this technology; 3) the

CITY OF OAKLAND

CITY HALL • 1 FRANK H. OGAWA PLAZA • OAKLAND, CALIFORNIA 94612

REBECCA KAPLAN
Council President
atlarge@oaklandnet.com

(510) 238-7008
FAX: (510) 238-6910
TDD: (510) 839-6451

invasive nature of the technology; and 4) and the potential abuses of data by our government that could lead to persecution of minority groups.

Data shows that this technology negatively and disproportionately misidentifies darker skinned women and the error rate has been widely studied. In a 2018 report by the MIT Lab, *Gender Shades: Intersection Accuracy Disparities in Commercial Gender Classification*, the study concluded, using a data set of 1,270 people, that facial recognitions systems worked best on white males and failed most often with the combination of female and dark-skin individuals with error rates of up to 34.7%. Last year, Amazon's Rekognition face surveillance product misidentified 28 members of Congress, disproportionately the Black and Latino representatives, as persons in a database of booking photos in a test conducted by the ACLU of Northern California. Apple, Inc., is currently being sued by an 18-year-old for misidentifying him as a thief, and in Sri Lanka, Face Recognition Technology falsely identified an American as a suspect in a terrorist bombing in April of 2019.

In addition, the misuse and lack of guidelines around the use of this technology by police departments in other jurisdictions has raised serious and ethical dilemmas. In May 2019, Georgetown Law's Center on Privacy and Technology (CPT) issued a report *Garbage in and Garbage Out*, detailing how law enforcement agencies across the country are feeding facial recognition software flawed data stating "when blurry or flawed photos of suspects have failed to turn up good leads, analysts have instead picked a celebrity they thought looked like the suspect, then run the celebrity's photo through their automated face recognition system looking for a lead." The report warns there are "no rules when it comes to what images police can submit to face recognition algorithms to generate investigative leads."

More recently concerns about privacy have been exacerbated by questions into how international, federal, and local government bodies use this data. In Baltimore, police agencies used face recognition technology to target activists in the aftermath of Freddie Gray's death, and the Chinese government is currently using face recognition software in the persecution of its Muslim minority population. This week, Detroit's face recognition system is under scrutiny, the placement of cameras near abortion clinics, and more frightening, there has been a slew of hackings, including most recently thousands of facial scans that were stolen from a subcontractor of the US Customs and Enforcement Agency.

FISCAL IMPACT
There is no anticipated fiscal impact to this legislation.

COORDINATION
This ordinance was unanimously recommended by the Oakland Privacy Commission. This ordinance was created with the assistance of the Oakland City Attorney's office.

CITY OF OAKLAND

CITY HALL • 1 FRANK H. OGAWA PLAZA • OAKLAND, CALIFORNIA 94612

REBECCA KAPLAN
Council President
atlarge@oaklandnet.com

(510) 238-7008
FAX: (510) 238-6910
TDD: (510) 839-6451

SUSTAINABLE OPPORTUNITIES

Economic: The information presented in this report presents no economic impact.

Environmental: There are no environmental opportunities identified in this report.

Social Equity: The adoption of an Ordinance is paramount for Social Equity as Face Recognition Technology has been shown to have an algorithm bias. In other words, there is a disparate treatment of persons with the use of this technology based on skin color and gender.

For questions regarding this report, please contact Bobbi Lopez, Policy Director, at 510-238-7082.

Attached please find the following documents:

1. Ordinance
2. Study – Buolamwini, Joy & Timnit, Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, Proceedings of Machine Learning Research 81: 1-15, 2018.
3. Study - Garvie, Clare. *Garbage In, Garbage Out: Face Recognition on Flawed Data*, Georgetown University's Center on Privacy and Technology, May 2019.

Sincerely,

Rebecca Kaplan
Council President

Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Recent studies demonstrate that machine learning algorithms can discriminate based on classes like race and gender. In this work, we present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups. Using the dermatologist approved Fitzpatrick Skin Type classification system, we characterize the gender and skin type distribution of two facial analysis benchmarks, IJB-A and Adience. We find that these datasets are overwhelmingly composed of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience) and introduce a new facial analysis dataset which is balanced by gender and skin type. We evaluate 3 commercial gender classification systems using our dataset and show that darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%. The substantial disparities in the accuracy of classifying darker females, lighter females, darker males, and lighter males in gender classification systems require urgent attention if commercial companies are to build genuinely fair, transparent and accountable facial analysis algorithms.

Keywords: Computer Vision, Algorithmic Audit, Gender Classification

1. Introduction

Artificial Intelligence (AI) is rapidly infiltrating every aspect of society. From helping determine who is hired, fired, granted a loan, or how long an individual spends in prison, decisions that have traditionally been performed by humans are rapidly made by algorithms (O'Neil, 2017; Citron and Pasquale, 2014). Even AI-based technologies that are not specifically trained to perform high-stakes tasks (such as determining how long someone spends in prison) can be used in a pipeline that performs such tasks. For example, while face recognition software by itself should not be trained to determine the fate of an individual in the criminal justice system, it is very likely that such software is used to identify suspects. Thus, an error in the output of a face recognition algorithm used as input for other tasks can have serious consequences. For example, someone could be wrongfully accused of a crime based on erroneous but confident misidentification of the perpetrator from security video footage analysis.

Many AI systems, e.g. face recognition tools, rely on machine learning algorithms that are trained with labeled data. It has recently been shown that algorithms trained with biased data have resulted in algorithmic discrimination (Bolukbasi et al., 2016; Caliskan et al., 2017). Bolukbasi et al. even showed that the popular word embedding space, Word2Vec, encodes societal gender biases. The authors used Word2Vec to train an analogy generator that fills in missing words in analogies. The analogy man is to computer programmer as woman is to "X" was completed with "homemaker", conforming to the stereotype that programming is associated with men and homemaking with women. The biases in Word2Vec are thus likely to be propagated throughout any system that uses this embedding.

* Download our gender and skin type balanced PPB dataset at gendershades.org

Although many works have studied how to create fairer algorithms, and benchmarked discrimination in various contexts (Kilbertus et al., 2017; Hardt et al., 2016b,a), only a handful of works have done this analysis for computer vision. However, computer vision systems with inferior performance across demographics can have serious implications. Esteva et al. showed that simple convolutional neural networks can be trained to detect melanoma from images, with accuracies as high as experts (Esteva et al., 2017). However, without a dataset that has labels for various skin characteristics such as color, thickness, and the amount of hair, one cannot measure the accuracy of such automated skin cancer detection systems for individuals with different skin types. Similar to the well documented detrimental effects of biased clinical trials (Popejoy and Fullerton, 2016; Melloni et al., 2010), biased samples in AI for health care can result in treatments that do not work well for many segments of the population.

In other contexts, a demographic group that is underrepresented in benchmark datasets can nonetheless be subjected to frequent targeting. The use of automated face recognition by law enforcement provides such an example. At least 117 million Americans are included in law enforcement face recognition networks. A year-long research investigation across 100 police departments revealed that African-American individuals are more likely to be stopped by law enforcement and be subjected to face recognition searches than individuals of other ethnicities (Garvie et al., 2016). False positives and unwarranted searches pose a threat to civil liberties. Some face recognition systems have been shown to misidentify people of color, women, and young people at high rates (Klare et al., 2012). Monitoring phenotypic and demographic accuracy of these systems as well as their use is necessary to protect citizens' rights and keep vendors and law enforcement accountable to the public.

We take a step in this direction by making two contributions. First, our work advances gender classification benchmarking by introducing a new face dataset composed of 1270 unique individuals that is more phenotypically balanced on the basis of skin type than existing benchmarks. To our knowledge this is the first gender classification benchmark labeled by the Fitzpatrick (TB, 1988) six-point skin type scale, allowing us to benchmark the performance of gender classification algorithms by skin type. Second, this work introduces the first intersectional demographic and phenotypic evaluation of face-based gender classification accuracy. Instead of evaluating accuracy by gender or skin type alone, accuracy is also examined on 4 intersectional subgroups: darker females, darker males, lighter females, and lighter males. The 3 evaluated commercial gender classifiers have the lowest accuracy on darker females. Since computer vision technology is being utilized in high-stakes sectors such as healthcare and law enforcement, more work needs to be done in benchmarking vision algorithms for various demographic and phenotypic groups.

2. Related Work

Automated Facial Analysis. Automated facial image analysis describes a range of face perception tasks including, but not limited to, face detection (Zafeiriou et al., 2015; Mathias et al., 2014; Bai and Ghanem, 2017), face classification (Reid et al., 2013; Levi and Hassner, 2015a; Rothe et al., 2016) and face recognition (Parkhi et al., 2015; Wen et al., 2016; Ranjan et al., 2017). Face recognition software is now built into most smart phones and companies such as Google, IBM, Microsoft and Face++ have released commercial software that perform automated facial analysis (IBM; Microsoft; Face++; Google).

A number of works have gone further than solely performing tasks like face detection, recognition and classification that are easy for humans to perform. For example, companies such as Affectiva (Affectiva) and researchers in academia attempt to identify emotions from images of people's faces (Dehghan et al., 2017; Srinivasan et al., 2016; Fabian Benitez-Quiroz et al., 2016). Some works have also used automated facial analysis to understand and help those with autism (Leo et al., 2015; Palestra et al., 2016). Controversial papers such as (Kosinski and Wang, 2017) claim to determine the sexuality of Caucasian males whose profile pictures are on Facebook or dating sites. And others such as (Wu and Zhang, 2016) and Israeli based company Faception (Faception) have developed software that purports to determine an individual's characteristics (e.g. propensity towards crime, IQ, terrorism) solely from

their faces. The clients of such software include governments. An article by (Aguera Y Arcas et al., 2017) details the dangers and errors propagated by some of these aforementioned works.

Face detection and classification algorithms are also used by US-based law enforcement for surveillance and crime prevention purposes. In "The Perpetual Lineup", Garvie and colleagues provide an in-depth analysis of the unregulated police use of face recognition and call for rigorous standards of automated facial analysis, racial accuracy testing, and regularly informing the public about the use of such technology (Garvie et al., 2016). Past research has also shown that the accuracies of face recognition systems used by US-based law enforcement are systematically lower for people labeled female, Black, or between the ages of 18—30 than for other demographic cohorts (Klare et al., 2012). The latest gender classification report from the National Institute for Standards and Technology (NIST) also shows that algorithms NIST evaluated performed worse for female-labeled faces than male-labeled faces (Ngan et al., 2015).

The lack of datasets that are labeled by ethnicity limits the generalizability of research exploring the impact of ethnicity on gender classification accuracy. While the NIST gender report explored the impact of ethnicity on gender classification through the use of an ethnic proxy (country of origin), none of the 10 locations used in the study were in Africa or the Caribbean where there are significant Black populations. On the other hand, Farinella and Dugelay claimed that ethnicity has no effect on gender classification, but they used a binary ethnic categorization scheme: Caucasian and non-Caucasian (Farinella and Dugelay, 2012). To address the underrepresentation of people of African-descent in previous studies, our work explores gender classification on African faces to further scholarship on the impact of phenotype on gender classification.

Benchmarks. Most large-scale attempts to collect visual face datasets rely on face detection algorithms to first detect faces (Huang et al., 2007; Kemelmacher-Shlizerman et al., 2016). Megaface, which to date is the largest publicly available set of facial images, was composed utilizing Head Hunter (Mathias et al., 2014) to select one million images from the Yahoo Flicker 100M image dataset (Thomee et al., 2015; Kemelmacher-Shlizerman et al., 2016). Any systematic error found in face detectors will inevitably affect the composition of the benchmark. Some datasets collected in this manner have already been documented to contain significant demographic bias. For example, LFW, a dataset composed of celebrity faces which has served as a gold standard benchmark for face recognition, was estimated to be 77.5% male and 83.5% White (Han and Jain, 2014). Although (Taigman et al., 2014)'s face recognition system recently reported 97.35% accuracy on the LFW dataset, its performance is not broken down by race or gender. Given these skews in the LFW dataset, it is not clear that the high reported accuracy is applicable to people who are not well represented in the LFW benchmark. In response to these limitations, Intelligence Advanced Research Projects Activity (IARPA) released the IJB-A dataset as the most geographically diverse set of collected faces (Klare et al., 2015). In order to limit bias, no face detector was used to select images containing faces. In comparison to face recognition, less work has been done to benchmark performance on gender classification. In 2015, the Adience gender and age classification benchmark was released (Levi and Hassner, 2015b). As of 2017, The National Institute of Standards and Technology is starting another challenge to spur improvement in face gender classification by expanding on the 2014-15 study.

3. Intersectional Benchmark

An evaluation of gender classification performance currently requires reducing the construct of gender into defined classes. In this work we use the sex labels of "male" and "female" to define gender classes since the evaluated benchmarks and classification systems use these binary labels. An intersectional evaluation further requires a dataset representing the defined genders with a range of phenotypes that enable subgroup accuracy analysis. To assess the suitability of existing datasets for intersectional benchmarking, we provided skin type annotations for unique subjects within two selected datasets, and compared the distribution of darker females, darker males, lighter females, and lighter males. Due to phenotypic imbalances in existing benchmarks, we

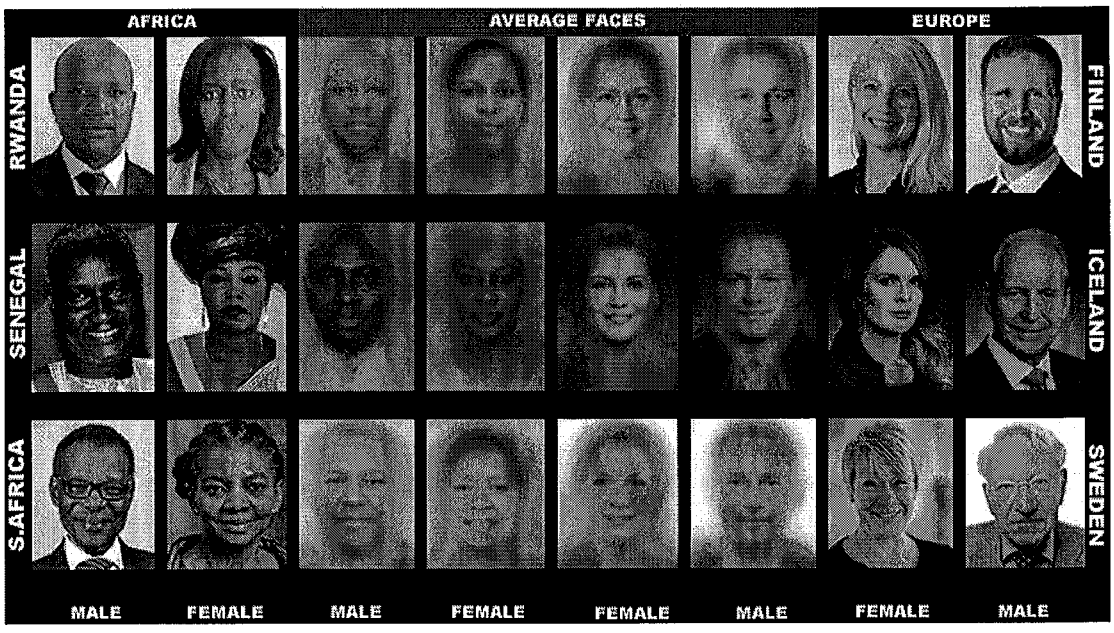

Figure 1: Example images and average faces from the new Pilot Parliaments Benchmark (PPB). As the examples show, the images are constrained with relatively little variation in pose. The subjects are composed of male and female parliamentarians from 6 countries. On average, Senegalese subjects are the darkest skinned while those from Finland and Iceland are the lightest skinned.

created a new dataset with more balanced skin type and gender representations.

3.1. Rationale for Phenotypic Labeling

Though demographic labels for protected classes like race and ethnicity have been used for performing algorithmic audits (Friedler et al., 2016; Angwin et al., 2016) and assessing dataset diversity (Han and Jain, 2014), phenotypic labels are seldom used for these purposes. While race labels are suitable for assessing potential algorithmic discrimination in some forms of data (e.g. those used to predict criminal recidivism rates), they face two key limitations when used on visual images. First, subjects' phenotypic features can vary widely within a racial or ethnic category. For example, the skin types of individuals identifying as Black in the US can represent many hues. Thus, facial analysis benchmarks consisting of lighter-skinned Black individuals would not adequately represent darker-skinned ones. Second, racial and ethnic categories are not consistent across geographies: even within countries these categories change over time.

Since race and ethnic labels are unstable, we decided to use skin type as a more visually precise label to measure dataset diversity. Skin type is one phenotypic attribute that can be used to more objectively characterize datasets along with eye and nose shapes. Furthermore, skin type was chosen as a phenotypic factor of interest because default camera settings are calibrated to expose lighter-skinned individuals (Roth, 2009). Poorly exposed images that result from sensor optimizations for lighter-skinned subjects or poor illumination can prove challenging for automated facial analysis. By labeling faces with skin type, we can increase our understanding of performance on this important phenotypic attribute.

3.2. Existing Benchmark Selection Rationale

IJB-A is a US government benchmark released by the National Institute of Standards and Tech-

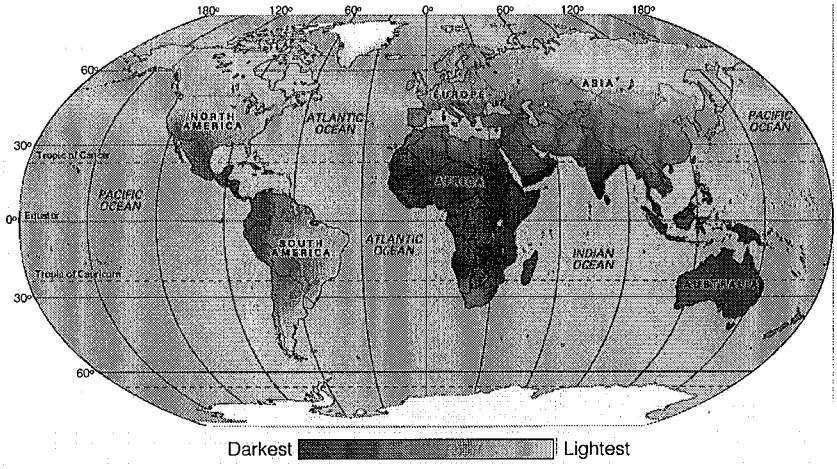

Figure 2: The global distribution of skin color. Most Africans have darker skin while those from Nordic countries are lighter-skinned. Image from (Encyclopedia Britannica) ©Copyright 2012 Encyclopedia Britannica.

nology (NIST) in 2015. We chose to evaluate this dataset given the government's involvement and the explicit development of the benchmark to be geographically diverse (as mentioned in Sec. 2). At the time of assessment in April and May of 2017, the dataset consisted of 500 unique subjects who are public figures. One image of each unique subject was manually labeled with one of six Fitzpatrick skin types (TB, 1988).

Adience is a gender classification benchmark released in 2014 and was selected due to its recency and unconstrained nature. The Adience benchmark contains 2,284 unique individual subjects. 2,194 of those subjects had reference images that were discernible enough to be labeled by skin type and gender. Like the IJB-A dataset, only one image of each subject was labeled for skin type.

3.3. Creation of Pilot Parliaments Benchmark

Preliminary analysis of the IJB-A and Adience benchmarks revealed overrepresentation of lighter males, underrepresentation of darker females, and underrepresentation of darker individuals in general. We developed the Pilot Parliaments Benchmark (PPB) to achieve better intersectional representation on the basis of gender and skin type. PPB consists of 1270 individuals from three African countries (Rwanda, Senegal, South Africa) and three European countries (Iceland, Finland, Sweden) selected for gender parity in the national parliaments.

Property	PPB	IJB-A	Adience
Release Year	2017	2015	2014
#Subjects	1270	500	2284
Avg. IPD	63 pixels	-	-
BBox Size	141 (avg)	≥36	-
IM Width	160-590	-	816
IM Height	213-886	-	816

Table 1: Various image characteristics of the Pilot Parliaments Benchmark compared with prior datasets. #Subjects denotes the number of unique subjects, the average bounding box size is given in pixels, and IM stands for image.

Figure 1 shows example images from PPB as well as average faces of males and females in each country represented in the datasets. We decided to use images of parliamentarians since they are public figures with known identities and photos available under non-restrictive licenses posted on government websites. To add skin

type diversity to the dataset, we chose parliamentarians from African and European countries. Fig. 2 shows an approximated distribution of average skin types around the world. As seen in the map, African countries typically have darker-skinned individuals whereas Nordic countries tend to have lighter-skinned citizens. Colonization and migration patterns nonetheless influence the phenotypic distribution of skin type and not all Africans are darker-skinned. Similarly, not all citizens of Nordic countries can be classified as lighter-skinned.

The specific African and European countries were selected based on their ranking for gender parity as assessed by the Inter Parliamentary Union (Inter Parliamentary Union Ranking). Of all the countries in the world, Rwanda has the highest proportion of women in parliament. Nordic countries were also well represented in the top 10 nations. Given the gender parity and prevalence of lighter skin in the region, Iceland, Finland, and Sweden were chosen. To balance for darker skin, the next two highest-ranking African nations, Senegal and South Africa, were also added.

Table 1 compares image characteristics of PPB with IJB-A and Adience. PPB is highly constrained since it is composed of official profile photos of parliamentarians. These profile photos are taken under conditions with cooperative subjects where pose is relatively fixed, illumination is constant, and expressions are neutral or smiling. Conversely, the images in the IJB-A and Adience benchmarks are unconstrained and subject pose, illumination, and expression by construction have more variation.

3.4. Intersectional Labeling Methodology

Skin Type Labels. We chose the Fitzpatrick six-point labeling system to determine skin type labels given its scientific origins. Dermatologists use this scale as the gold standard for skin classification and determining risk for skin cancer (TB, 1988).

The six-point Fitzpatrick classification system which labels skin as Type I to Type VI is skewed towards lighter skin and has three categories that can be applied to people perceived as White (Figure 2). Yet when it comes to fully representing the sepia spectrum that characterizes the rest of the world, the categorizations are fairly coarse. Nonetheless, the scale provides a scientifically based starting point for auditing algorithms and datasets by skin type.

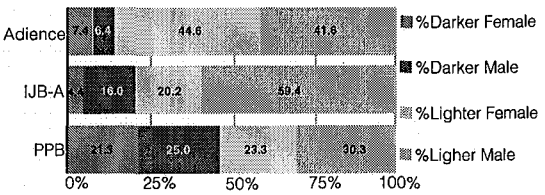

Figure 3: The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB.

Gender Labels. All evaluated companies provided a "gender classification" feature that uses the binary sex labels of female and male. This reductionist view of gender does not adequately capture the complexities of gender or address transgender identities. The companies provide no documentation to clarify if their gender classification systems which provide sex labels are classifying gender identity or biological sex. To label the PPB data, we use female and male labels to indicate subjects perceived as women or men respectively.

Labeling Process. For existing benchmarks, one author labeled each image with one of six Fitzpatrick skin types and provided gender annotations for the IJB-A dataset. The Adience benchmark was already annotated for gender. These preliminary skin type annotations on existing datasets were used to determine if a new benchmark was needed.

More annotation resources were used to label PPB. For the new parliamentarian benchmark, 3 annotators including the authors provided gender and Fitzpatrick labels. A board-certified surgical dermatologist provided the definitive labels for the Fitzpatrick skin type. Gender labels were determined based on the name of the parliamentarian, gendered title, prefixes such as Mr or Ms, and the appearance of the photo.

Set	n	F	M	Darker	Lighter	DF	DM	LF	LM
All Subjects	1270	44.6%	55.4%	46.4%	53.6%	21.3%	25.0%	23.3%	30.3%
Africa	661	43.9%	56.1%	86.2%	13.8%	39.8%	46.4%	4.1%	9.7%
South Africa	437	41.4%	58.6%	79.2%	20.8%	35.2%	43.9%	6.2%	14.6%
Senegal	149	43.0%	57.0%	100.0%	0.0%	43.0%	57.0%	0.0%	0.0%
Rwanda	75	60.0%	40.0%	100.0%	0.0%	60.0%	40.0%	0.0%	0.0%
Europe	609	45.5%	54.5%	3.1%	96.9%	1.3%	1.8%	44.2%	52.7%
Sweden	349	46.7%	53.3%	4.9%	95.1%	2.0%	2.9%	44.7%	50.4%
Finland	197	42.6%	57.4%	1.0%	99.0%	0.5%	0.5%	42.1%	56.9%
Iceland	63	47.6%	52.4%	0.0%	100.0%	0.0%	0.0%	47.6%	52.4%

Table 2: Pilot Parliaments Benchmark decomposition by the total number of female subjects denoted as F, total number of male subjects (M), total number of darker and lighter subjects, as well as female darker/lighter (DF/LF) and male darker/lighter subjects (DM/LM). The group compositions are shown for all unique subjects, Africa, Europe and the countries in our dataset located in each of these continents.

Dataset	Lighter (I,II,III)		Darker (IV, V, VI)		Total
PPB	53.6%	681	46.4%	589	1270
IJB-A	79.6%	398	20.4%	102	500
Adience	86.2%	1892	13.8%	302	2194

Table 3: The distributions of lighter and darker-skinned subjects (according to the Fitzpatrick classification system) in PPB, IJB-A, and Adience datasets. Adience has the most skewed distribution with 86.2% of the subjects consisting of lighter-skinned individuals whereas PPB is more evenly distributed between lighter (53.6%) and darker (46.4%) subjects.

3.5. Fitzpatrick Skin Type Comparison

For the purposes of our analysis, lighter subjects will refer to faces with a Fitzpatrick skin type of I,II, or III. Darker subjects will refer to faces labeled with a Fitzpatrick skin type of IV,V, or VI. We intentionally choose countries with majority populations at opposite ends of the skin type scale to make the lighter/darker dichotomy more distinct. The skin types are aggregated to account for potential off-by-one errors since the skin type is estimated using images instead of employing a standard spectrophotometer and Fitzpatrick questionnaire.

Table 2 presents the gender, skin type, and intersectional gender by skin type composition of PPB. And Figure 3 compares the percentage of images from darker female, darker male, lighter female and lighter male subjects from Adience, IJB-A, and PBB. PPB provides the most balanced representation of all four groups whereas IJB-A has the least balanced distribution.

Darker females are the least represented in IJB-A (4.4%) and darker males are the least represented in Adience (6.4%). Lighter males are the most represented unique subjects in all datasets. IJB-A is composed of 59.4% unique lighter males whereas this percentage is reduced to 41.6% in Adience and 30.3% in PPB. As seen in Table 3, Adience has the most skewed distribution by skin type.

While all the datasets have more lighter-skinned unique individuals, PPB is around half light at 53.6% whereas the proportion of lighter-skinned unique subjects in IJB-A and Adience

is 79.6% and 86.2% respectively. PPB provides substantially more darker-skinned unique subjects than IJB-A and Adience. Even though Adience has 2194 labeled unique subjects, which is nearly twice that of the 1270 subjects in PPB, it has 302 darker subjects, nearly half the 589 darker subjects in PPB. Overall, PPB has a more balanced representation of lighter and darker subjects as compared to the IJB-A and Adience datasets.

4. Commercial Gender Classification Audit

We evaluated 3 commercial gender classifiers. Overall, male subjects were more accurately classified than female subjects replicating previous findings (Ngan et al., 2015), and lighter subjects were more accurately classified than darker individuals. An intersectional breakdown reveals that all classifiers performed worst on darker female subjects.

4.1. Key Findings on Evaluated Classifiers

- All classifiers perform better on male faces than female faces (8.1% – 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% – 19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8% – 34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

4.2. Commercial Gender Classifier Selection: Microsoft, IBM, Face++

We focus on gender classifiers sold in API bundles made available by Microsoft, IBM, and Face++ (Microsoft; IBM; Face++). Microsoft's Cognitive Services Face API and IBM's Watson Visual Recognition API were chosen since both companies have made large investments in artificial intelligence, capture significant market shares in the machine learning services domain, and provide public demonstrations of their facial analysis technology. At the time of evaluation, Google did not provide a publicly available gender classifier. Previous studies have shown that face recognition systems developed in Western nations and those developed in Asian nations tend to perform better on their respective populations (Phillips et al., 2011). Face++, a computer vision company headquartered in China with facial analysis technology previously integrated with some Lenovo computers, was thus chosen to see if this observation holds for gender classification. Like Microsoft and IBM, Face++ also provided a publicly available demonstration of their gender classification capabilities at the time of evaluation(April and May 2017).

All of the companies offered gender classification as a component of a set of proprietary facial analysis API services (Microsoft; IBM; Face++). The description of classification methodology lacked detail and there was no mention of what training data was used. At the time of evaluation, Microsoft's Face Detect service was described as using advanced statistical algorithms that "may not always be 100% precise" (Microsoft API Reference). IBM Watson Visual Recognition and Face++ services were said to use deep learning-based algorithms (IBM API Reference; Face++ Terms of Service). None of the commercial gender classifiers chosen for this analysis reported performance metrics on existing gender estimation benchmarks in their provided documentation. The Face++ terms of use explicitly disclaim any warranties of accuracy. Only IBM provided confidence scores (between 0 and 1) for face-based gender classification labels. But it did not report how any metrics like true positive rates (TPR) or false positive rates (FPR) were balanced.

4.3. Evaluation Methodology

In following the gender classification evaluation precedent established by the National Institute for Standards and Technology (NIST), we assess

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Classifier	Metric	DF	DM	LF	LM
MSFT	PPV(%)	76.2	**100**	**100**	**100**
	Error Rate(%)	**23.8**	0.0	0.0	0.0
	TPR(%)	**100**	84.2	**100**	**100**
	FPR(%)	**15.8**	0.0	0.0	0.0
Face++	PPV(%)	64.0	99.5	**100**	**100**
	Error Rate(%)	**36.0**	0.5	0.0	0.0
	TPR(%)	99.0	77.8	**100**	96.9
	FPR(%)	**22.2**	1.03	3.08	0.0
IBM	PPV(%)	66.9	94.3	**100**	98.4
	Error Rate(%)	**33.1**	5.7	0.0	1.6
	TPR(%)	90.4	78.0	96.4	**100**
	FPR(%)	**22.0**	9.7	0.0	3.6

Table 5: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the South African subset of the PPB dataset. Results for South Africa follow the overall trend with the highest error rates seen on darker-skinned females.

the overall classification accuracy, male classification accuracy, and female classification accuracy as measured by the positive predictive value (PPV). Extending beyond the NIST Methodology we also evaluate the true positive rate, false positive rate, and error rate (1-PPV) of the following groups: all subjects, male subjects, female subjects, lighter subjects, darker subjects, darker females, darker males, lighter females, and lighter males. See Table 2 in supplementary materials for results disaggregated by gender and each Fitzpatrick Skin Type.

4.4. Audit Results

MALE AND FEMALE ERROR RATES

To conduct a demographic performance analysis, the differences in male and female error rates for each gender classifier are compared first in aggregate (Table 4) and then for South Africa (Table 5). The NIST Evaluation of Automated Gender Classification Algorithms report revealed that gender classification performance on female faces was 1.8% to 12.5% lower than performance on male faces for the nine evaluated algorithms (Ngan et al., 2015). The gender misclassification rates on the Pilot Parliaments Benchmark replicate this trend across all classifiers. The differences between female and male classification error rates range from 8.1% to 20.6%. The relatively high true positive rates for females indicate that when a face is predicted to be female the estimation is more likely to be correct than when a face is predicted to be male. For the Microsoft and IBM classifiers, the false positive rates (FPR) for females are double or more than the FPR for males. The FPR for females is more than 13 times that of males with the Face++ classifier.

DARKER AND LIGHTER ERROR RATES

To conduct a phenotypic performance analysis, the differences in darker and lighter skin type error rates for each gender classifier are compared first in aggregate (Table 4) and then for South Africa (Table 5). All classifiers perform better on lighter subjects than darker subjects in PPB. Microsoft achieves the best result with error rates of 12.9% on darker subjects and 0.7% on lighter individuals. On darker subjects, IBM achieves the worst classification accuracy with an error rate of 22.4%. This rate is nearly 7 times higher than the IBM error rate on lighter faces.

INTERSECTIONAL ERROR RATES

To conduct an intersectional demographic and phenotypic analysis, the error rates for four intersectional groups (darker females, darker males, lighter females and lighter males) are compared in aggregate and then for South Africa.

Across the board, darker females account for the largest proportion of misclassified subjects. Even though darker females make up 21.3% of the PPB benchmark, they constitute between 61.0% to 72.4.1% of the classification error. Lighter males who make up 30.3% of the benchmark contribute only 0.0% to 2.4% of the total errors from these classifiers (See Table 1 in supplementary materials).

We present a deeper look at images from South Africa to see if differences in algorithmic performance are mainly due to image quality from each parliament. In PPB, the European parliamentary images tend to be of higher resolution with less pose variation when compared to images from African parliaments. The South African parliament, however, has comparable image resolution and has the largest skin type spread of all the parliaments. Lighter subjects makeup 20.8% (n=91) of the images, and darker subjects make up the remaining 79.2% (n=346) of images. Table 5 shows that all algorithms perform worse on female and darker subjects when compared to their counterpart male and lighter subjects. The Microsoft gender classifier performs the best, with zero errors on classifying all males and lighter females.

On the South African subset of the PPB benchmark, all the error for Microsoft arises from misclassifying images of darker females. Table 5 also shows that all classifiers perform worse on darker females. Face++ is flawless on lighter males and lighter females. IBM performs best on lighter females with 0.0% error rate. Examining classification performance on the South African subset of PPB reveals trends that closely match the algorithmic performance on the entire dataset. Thus, we conclude that variation in performance due to the image characteristics of each country does not fully account for the differences in misclassification rates between intersectional subgroups. In other words, the presence of more darker individuals is a better explanation for error rates than a deviation in how images of parliamentarians are composed and produced. However, darker skin alone may not be fully responsible for misclassification. Instead, darker skin may be highly correlated with facial geometries or gender display norms that were less represented in the training data of the evaluated classifiers.

4.5. Analysis of Results

The overall gender classification accuracy results show the obfuscating nature of single perfor-

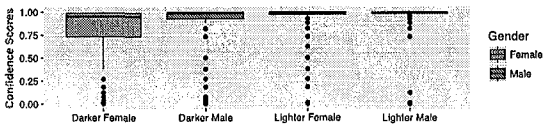

Figure 4: Gender classification confidence scores from IBM (IBM). Scores are near 1 for lighter male and female subjects while they range from $\sim 0.75 - 1$ for darker females.

mance metrics. Taken at face value, gender classification accuracies ranging from 87.9% to 93.7% on the PPB dataset, suggest that these classifiers can be used for all populations represented by the benchmark. A company might justify the market readiness of a classifier by presenting performance results in aggregate. Yet a gender and phenotypic breakdown of the results shows that performance differs substantially for distinct subgroups. Classification is $8.1\% - 20.6\%$ worse on female than male subjects and $11.8\% - 19.2\%$ worse on darker than lighter subjects.

Though helpful in seeing systematic error, gender and skin type analysis by themselves do not present the whole story. Is misclassification distributed evenly amongst all females? Are there other factors at play? Likewise, is the misclassification of darker skin uniform across gender?

The intersectional error analysis that targets gender classification performance on darker female, lighter female, darker male, and lighter male subgroups provides more answers. Darker females have the highest error rates for all gender classifiers ranging from $20.8\% - 34.7\%$. For Microsoft and IBM classifiers lighter males are the best classified group with 0.0% and 0.3% error rates respectively. Face++ classifies darker males best with an error rate of 0.7%. When examining the gap in lighter and darker skin classification, we see that even though darker females are most impacted, darker males are still more misclassified than lighter males for IBM and Microsoft. The most improvement is needed on darker females specifically. More broadly, the error gaps between male and female classification along with lighter and darker classification should be closed.

4.6. Accuracy Metrics

Microsoft and Face++ APIs solely output single labels indicating whether the face was classified as female or male. IBM's API outputs an additional number which indicates the confidence with which the classification was made. Figure 4 plots the distribution of confidence values for each of the subgroups we evaluate (i.e. darker females, darker males, lighter females and lighter males). Numbers near 0 indicate low confidence whereas those close to 1 denote high confidence in classifying gender. As shown in the box plots, the API is most confident in classifying lighter males and least confident in classifying darker females.

While confidence values give users more information, commercial classifiers should provide additional metrics. All 3 evaluated APIs only provide gender classifications, they do not output probabilities associated with the likelihood of being a particular gender. This indicates that companies are choosing a threshold which determines the classification: if the prediction probability is greater than this threshold, the image is determined to be that of a male (or female) subject, and viceversa if the probability is less than this number. This does not give users the ability to analyze true positive (TPR) and false positive (FPR) rates for various subgroups if different thresholds were to be chosen. The commercial classifiers have picked thresholds that result in specific TPR and FPR rates for each subgroup. And the FPR for some groups can be much higher than those for others. By having APIs that fail to provide the ability to adjust these thresholds, they are limiting users' ability to pick their own TPR/FPR trade-off.

4.7. Data Quality and Sensors

It is well established that pose, illumination, and expression (PIE) can impact the accuracy of automated facial analysis. Techniques to create robust systems that are invariant to pose, illumination, expression, occlusions, and background have received substantial attention in computer vision research (Kakadiaris et al., 2017; Ganguly et al., 2015; Ahmad Radzi et al., 2014). Illumination is of particular importance when doing an evaluation based on skin type. Default camera settings are often optimized to expose lighter skin

better than darker skin (Roth, 2009). Underexposed or overexposed images that present significant information loss can make accurate classification challenging.

With full awareness of the challenges that arise due to pose and illumination, we intentionally chose an optimistic sample of constrained images that were taken from the parliamentarian websites. Each country had its peculiarities. Images from Rwanda and Senegal had more pose and illumination variation than images from other countries (Figure 1). The Swedish parliamentarians all had photos that were taken with a shadow on the face. The South African images had the most consistent pose and illumination. The South African subset was also composed of a substantial number of lighter and darker subjects. Given the diversity of the subset, the high image resolution, and the consistency of illumination and pose, our finding that classification accuracy varied by gender, skin type, and the intersection of gender with skin type do not appear to be confounded by the quality of sensor readings. The disparities presented with such a constrained dataset do suggest that error rates would be higher on more challenging unconstrained datasets. Future work should explore gender classification on an inclusive benchmark composed of unconstrained images.

5. Conclusion

We measured the accuracy of 3 commercial gender classification algorithms on the new Pilot Parliaments Benchmark which is balanced by gender and skin type. We annotated the dataset with the Fitzpatrick skin classification system and tested gender classification performance on 4 subgroups: darker females, darker males, lighter females and lighter males. We found that all classifiers performed best for lighter individuals and males overall. The classifiers performed worst for darker females. Further work is needed to see if the substantial error rate gaps on the basis of gender, skin type and intersectional subgroup revealed in this study of gender classification persist in other human-based computer vision tasks. Future work should explore intersectional error analysis of facial detection, identification and verification. Intersectional phenotypic and demographic error analysis can help inform methods to improve dataset composition, feature selection, and neural network architectures.

Because algorithmic fairness is based on different contextual assumptions and optimizations for accuracy, this work aimed to show why we need rigorous reporting on the performance metrics on which algorithmic fairness debates center. The work focuses on increasing phenotypic and demographic representation in face datasets and algorithmic evaluation. Inclusive benchmark datasets and subgroup accuracy reports will be necessary to increase transparency and accountability in artificial intelligence. For human-centered computer vision, we define transparency as providing information on the demographic and phenotypic composition of training and benchmark datasets. We define accountability as reporting algorithmic performance on demographic and phenotypic subgroups and actively working to close performance gaps where they arise. Algorithmic transparency and accountability reach beyond technical reports and should include mechanisms for consent and redress which we do not focus on here. Nonetheless, the findings from this work concerning benchmark representation and intersectional auditing provide empirical support for increased demographic and phenotypic transparency and accountability in artificial intelligence.

Acknowledgments

We thank board-certified surgical dermatologist Dr. Helen Raynham for providing the official Fitzpatrick annotations for the Pilot Parliaments Benchmark.

References

Face++ API. http://old.faceplusplus.com/demo-detect/. Accessed: 2017-10-06.

Face, Google APIs for Android, Google Developers. https://developers.google.com/android/reference/com/google/android/gms/vision/face/Face. Accessed: 2017-10-06.

Watson Visual Recognition. https://www.ibm.com/watson/services/visual-recognition/. Accessed: 2017-10-06.

Microsoft Face API. https://www.microsoft.com/cognitive-services/en-us/faceapi. Accessed: 2017-10-06.

Affectiva Emotion Recognition Software and Analysis. https://www.affectiva.com/. Accessed: 2017-10-06.

Physiognomys New Clothes. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a. Accessed: 2017-10-06.

Face++ Terms of Use. a. Accessed: 2018-12-13.

Faception, Facial Personality Analytics. https://www.faception.com/, b. Accessed: 2017-10-06.

Visual Recognition API Reference. Accessed: 2018-12-13.

How to Detect Faces in Image. Accessed: 2018-12-13.

Proportion of seats held by women in national parliaments. https://data.worldbank.org/indicator/SG.GEN.PARL.ZS?year_high_desc=true. Accessed: 2017-10-06.

Syafeeza Ahmad Radzi, Khalil-Hani Mohamad, Shan Sung Liew, and Rabia Bakhteri. Convolutional neural network for face recognition with pose and illumination variation. *International Journal of Engineering and Technology (IJET)*, 6(1):44–57, 2014.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: Theres software used across the country to predict future criminals. and its biased against blacks. *ProPublica, May*, 23, 2016.

Yancheng Bai and Bernard Ghanem. Multi-scale fully convolutional network for face detection in the wild. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 2078–2087. IEEE, 2017.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf.

Encyclopedia Britannica. Skin distribution map. https://media1.britannica.com/eb-media/59/61759-004-9A507F1C.gif, 2012. Accessed: 2017-12-17.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Danielle Keats Citron and Frank A Pasquale. The scored society: due process for automated predictions. 2014.

Afshin Dehghan, Enrique G Ortiz, Guang Shu, and Syed Zain Masood. Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280*, 2017.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

Giovanna Farinella and Jean-Luc Dugelay. Demographic classification: Do gender and ethnicity affect each other? In *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*, pages 383–390. IEEE, 2012.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Suranjan Ganguly, Debotosh Bhattacharjee, and Mita Nasipuri. Illumination, pose and occlusion invariant face recognition from range images using erfi model. *International Journal of System Dynamics Applications (IJSDA)*, 4(2): 1–20, 2015.

Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. *The Perpetual Line-Up: Unregulated Police Face Recognition in America.* Georgetown Law, Center on Privacy & Technology, 2016.

Hu Han and Anil K Jain. Age, gender and race estimation from unconstrained face images. *Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5)*, 2014.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016a.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016b.

Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

Ioannis A Kakadiaris, George Toderici, Georgios Evangelopoulos, Georgios Passalis, Dat Chu, Xi Zhao, Shishir K Shah, and Theoharis Theoharis. 3d-2d face recognition with pose and illumination normalization. *Computer Vision and Image Understanding*, 154:137–151, 2017.

Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, 2015.

Michal Kosinski and Yilun Wang. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. 2017.

Marco Leo, Marco Del Coco, Pierluigi Carcagni, Cosimo Distante, Massimo Bernava, Giovanni Pioggia, and Giuseppe Palestra. Automatic emotion recognition in robot-children interaction for asd treatment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 145–153, 2015.

Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015a.

Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015b.

Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014.

Chiara Melloni, Jeffrey S Berger, Tracy Y Wang, Funda Gunes, Amanda Stebbins, Karen S Pieper, Rowena J Dolor, Pamela S Douglas, Daniel B Mark, and L Kristin Newby. Representation of women in randomized clinical trials of cardiovascular disease prevention. *Circu-*

lation: Cardiovascular Quality and Outcomes, 3(2):135–142, 2010.

Mei Ngan, Mei Ngan, and Patrick Grother. *Face recognition vendor test (FRVT) performance of automated gender classification algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2015.

Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

Giuseppe Palestra, Giovanna Varni, Mohamed Chetouani, and Floriana Esposito. A multimodal and multilevel system for robotics treatment of autism in children. In *Proceedings of the International Workshop on Social Learning and Multimodal Interaction for Designing Artificial Agents*, page 3. ACM, 2016.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):14, 2011.

Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature*, 538 (7624):161, 2016.

Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 17–24. IEEE, 2017.

Daniel Reid, Sina Samangooei, Cunjian Chen, Mark Nixon, and Arun Ross. Soft biometrics for surveillance: an overview. *Machine learning: theory and applications. Elsevier*, pages 327–352, 2013.

Lorna Roth. Looking at shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity. *Canadian Journal of Communication*, 34(1):111, 2009.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, pages 1–14, 2016.

Ramprakash Srinivasan, Julie D Golomb, and Aleix M Martinez. A neural basis of facial action recognition in humans. *Journal of Neuroscience*, 36(16):4434–4442, 2016.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

Fitzpatrick TB. The validity and practicality of sun-reactive skin types i through vi. *Archives of Dermatology*, 124(6):869–871, 1988. doi: 10.1001 / archderm.1988.01670060015008. URL +http : / / dx.doi.org / 10.1001 / archderm.1988.01670060015008.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 1(8), 2015.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, 2016.

Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015.

GARBAGE IN, GARBAGE OUT

FACE RECOGNITION ON FLAWED DATA

Clare Garvie

May 16, 2019

INTRODUCTION

On April 28, 2017, a suspect was caught on camera reportedly stealing beer from a CVS in New York City. The store surveillance camera that recorded the incident captured the suspect's face, but it was partially obscured and highly pixelated. When the investigating detectives submitted the photo to the New York Police Department's (NYPD) facial recognition system, it returned no useful matches.[1]

Rather than concluding that the suspect could not be identified using face recognition, however, the detectives got creative.

One detective from the Facial Identification Section (FIS), responsible for conducting face recognition searches for the NYPD, noted that the suspect looked like the actor Woody Harrelson, known for his performances in *Cheers*, *Natural Born Killers*, *True Detective*, and other television shows and movies. A Google image search for the actor predictably returned high-quality images, which detectives then submitted to the face recognition algorithm in place of the suspect's photo. In the resulting list of possible candidates, the detectives identified someone they believed was a match—not to Harrelson but to the suspect whose photo had produced no possible hits.[2]

This celebrity "match" was sent back to the investigating officers, and someone who was not Woody Harrelson was eventually arrested for petit larceny.

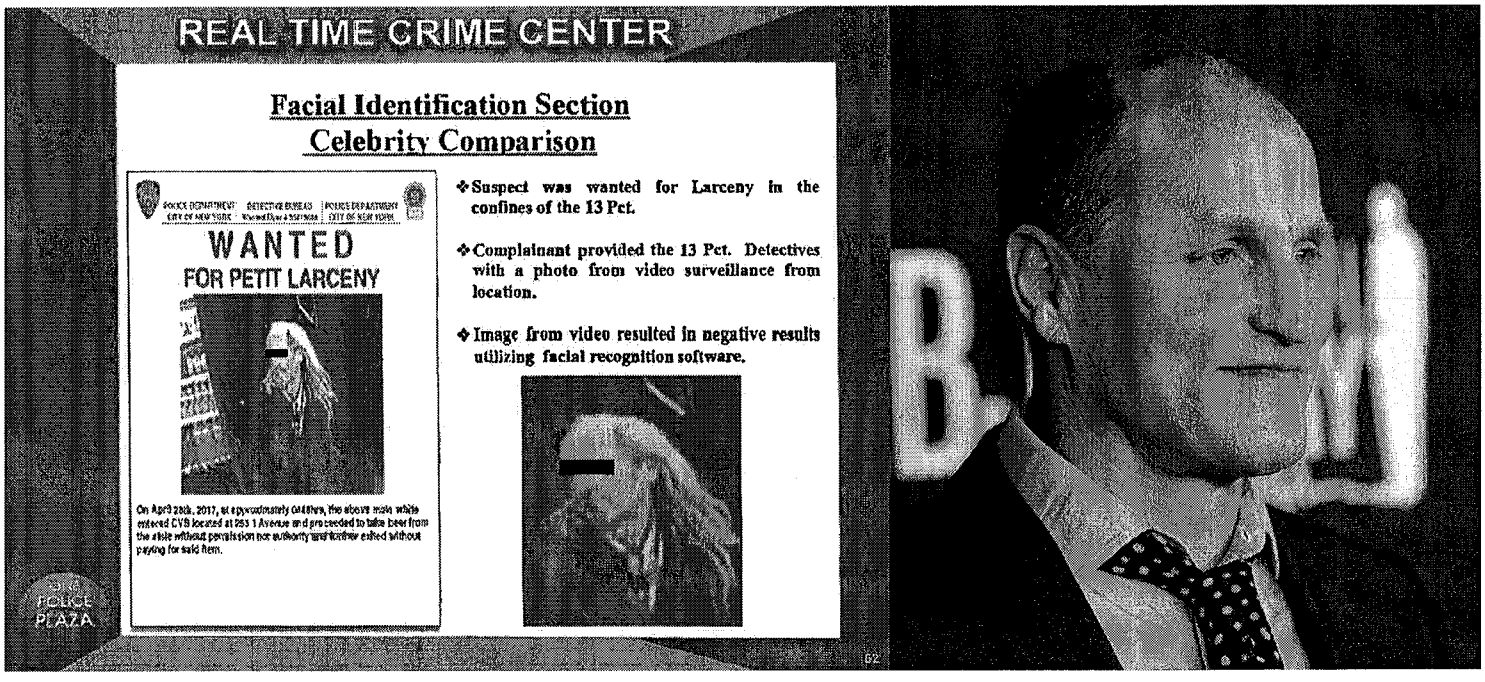

Figure 1: On the left: a slide from the NYPD FIS describing its "celebrity comparison" technique. On the right, a photo of Woody Harrelson. (Source: left, NYPD; right, Gabriel Cristóver Pérez/LBJ Presidential Library.)

There are no rules when it comes to what images police can submit to face recognition algorithms to generate investigative leads. As a consequence, agencies across the country can—and do—submit all manner of "probe photos," photos of unknown individuals submitted for search against a police or driver license database. These images may be low-quality surveillance camera stills, social media photos with filters, and scanned photo album pictures.[3] Records from police departments show they may also include computer-generated facial features, or composite or artist sketches.[4]

Or the probe photo may be a suspect's celebrity doppelgänger. Woody Harrelson is not the only celebrity to stand in for a suspect wanted by the NYPD. FIS has also used a photo of a New York Knicks player to search its face recognition database for a man wanted for assault in Brooklyn.[5]

The stakes are too high in criminal investigations to rely on unreliable—or wrong—inputs. It is one thing for a company to build a face recognition system designed to help individuals find their celebrity doppelgänger[6] or painting lookalike[7] for entertainment purposes. It's quite another to use these techniques to identify criminal suspects, who may be deprived of their liberty and ultimately prosecuted based on the match. Unfortunately, police departments' reliance on questionable probe photos appears all too common.

GARBAGE IN, GARBAGE OUT

"Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?"

—Charles Babbage[8]

"Garbage in, garbage out" is a phrase used to express the idea that inputting low-quality or nonsensical data into a system will produce low-quality or nonsensical results. It doesn't matter how powerful or cleverly-designed a system is, it can only operate on the information it is provided—if data is missing, the system cannot operate on it. Any attempt to reconstruct or approximate missing data will necessarily be a "guess" as to what information that data contained.

Worse, if data is wrong—like a photo of someone other than the suspect—the system has no way to correct it. It has literally no information about the suspect, and can't make it up.

Photos that are pixelated, distorted, or of partial faces provide less data for a face recognition system to analyze than high-quality, passport-style photos, increasing room for error.[9]

Face recognition technology has improved immensely in the past two years alone, enabling rapid searches of larger databases and more reliable pairings in testing environments.[10] But it doesn't matter how good the machine is if it is still being fed the wrong figures—the wrong answers are still likely to come out.

1. COMPOSITE SKETCHES AS PROBE IMAGES

"Composite art is an unusual marriage of two unlikely disciplines: police investigative work and art .... It is essential to realize that a composite sketch is a drawing of a victim's or witness's perception of a perpetrator at the time he or she was observed. It is not meant to be an exact portrait of the suspect. Keep the two words 'likeness' and 'similarity' in mind at all times. This is the best a composite sketch can achieve."

—*The Police Composite Sketch*[11]

In early 2018, Google rolled out "Art Selfie" — an app designed to match a user's photo to a famous painting lookalike using face recognition.[12] The result is an often-humorous photo pairing and an opportunity to learn more about art.

Less humorous is the fact that some police departments do the same thing when looking for criminal suspects, just in reverse—submitting art in an attempt to identify real people.

At least half a dozen police departments across the country permit, if not encourage, the use of face recognition searches on forensic sketches.

At least half a dozen police departments across the country permit, if not encourage, the use of face recognition searches on forensic sketches—hand drawn or computer generated composite faces based on descriptions that a witness has offered. In a brochure informing its officers about the acquisition of face recognition, the Maricopa

County Sheriff's Office in Arizona states: "[T]he image can be from a variety of sources including police artist renderings," and that the technology "can be used effectively in suspect identifications using photographs, surveillance still and video, suspect sketches and even forensic busts."[13] A presentation about the face recognition system that the Washington County Sheriff's Department in Oregon operates includes a "Real World Example" of the technology being used to identify an artist's drawing of a face.[14]

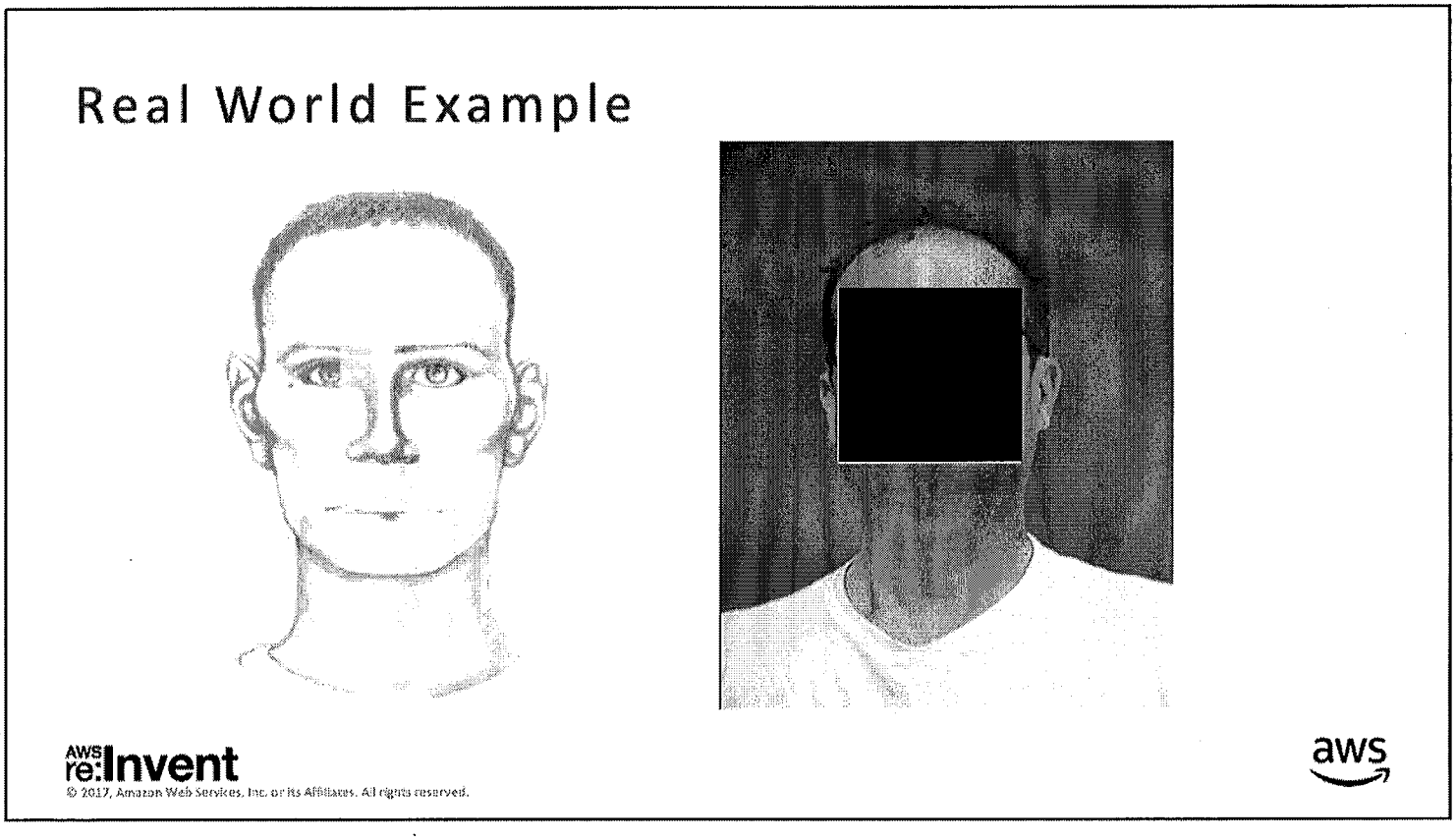

Figure 2: Slide from an AWS presentation titled "Washington County Sheriff's Office Rekognition Case Study." (Source: Public records obtained by ACLU Oregon & Northern California.)

A face recognition Privacy Impact Assessment that a working group of 15 state and federal agencies authored in 2011 states that it should be permissible to use face recognition to "...identify suspects based upon artist's sketches."[15] Information about the Maryland Department of Public Safety and Correctional Services, the Northern Virginia Regional Information System, and the Pinellas County Sheriff's Office in Florida suggest that sketches could be submitted to these agencies' face recognition systems as well.[16]

This practice is endorsed by some of the companies providing these face recognition systems to police departments. The example from the Washington County in Figure 2 is part of a case study that Amazon Web Services highlighted in a presentation about the capabilities of its face recognition software, Rekognition. Cognitec, one of the leading providers of face recognition algorithms to U.S. law enforcement, promotes the use of its software to "identify individuals in crime scene photos, video stills and sketches."[17] Vigilant Solutions markets tools specifically for "creating a proxy image from a sketch artist or artist rendering" to be submitted to its face recognition system.[18]

A. SCIENTIFIC REVIEW OF COMPOSITE IMAGE FACE RECOGNITION

Even the most detailed sketches make poor face recognition probe images. The Los Angeles County Sheriff's Department face recognition user guide summarizes this well:

> "A photograph taken of a real person should be used. Composite drawing will have marginal success because they are rendered pictures and do not accurately detail precise features."[19]

Studies that have analyzed the performance of face recognition systems on composite sketches conclude the same. A 2011 Michigan State University study noted that "[c]ommercial face recognition systems are not designed to match forensic sketches against face photographs."[20] In 2013, researchers studying this question ran sketches against a face recognition database using a commercially-available algorithm from Cognitec—one of the companies that advertises this as a feature of its system. The algorithm was programmed to return a list of 200 possible matches searching a database of 10,000 images. For sketches, it retrieved the correct match between 4.1 and 6.7 percent of the time.[21] Put another way, in only about 1 of every 20 searches would the correct match show up in the top 200 possible matches that the algorithm produced.[22]

In 2014, the National Institute of Standards and Technology (NIST) found similarly poor results, concluding that "[s]ketch searches mostly fail."[23] The NYPD has separately concluded the same thing from their own experience. According to NYPD detective Tom Markiewicz, FIS has tried running face recognition on sketches in the past and found that "sketches do not work."[24] So did the Pinellas County Sheriff's Office, concluding that the practice "is doubtful on yielding successful results with the current [system]" —yet it still permits the practice nonetheless.[25]

B. FORENSIC SKETCHES AND MISIDENTIFICATION

The most likely outcome of using a forensic sketch as a probe photo is that the system fails to find a match—even when the suspect is in the photo database available to law enforcement. With this outcome, the system produces no useful leads, and investigating officers must go back to the drawing board.

But this practice also introduces the possibility of misidentification. The process of generating a forensic sketch is inherently subjective. Sketches typically rely on:

a. An eyewitness's memory of what the subject looked like;

b. The eyewitness's ability to communicate the memory of the subject to a sketch artist;

c. The artist's ability to translate that description into an accurate drawing of the subject's face, someone whom the artist has never seen in person.[26]

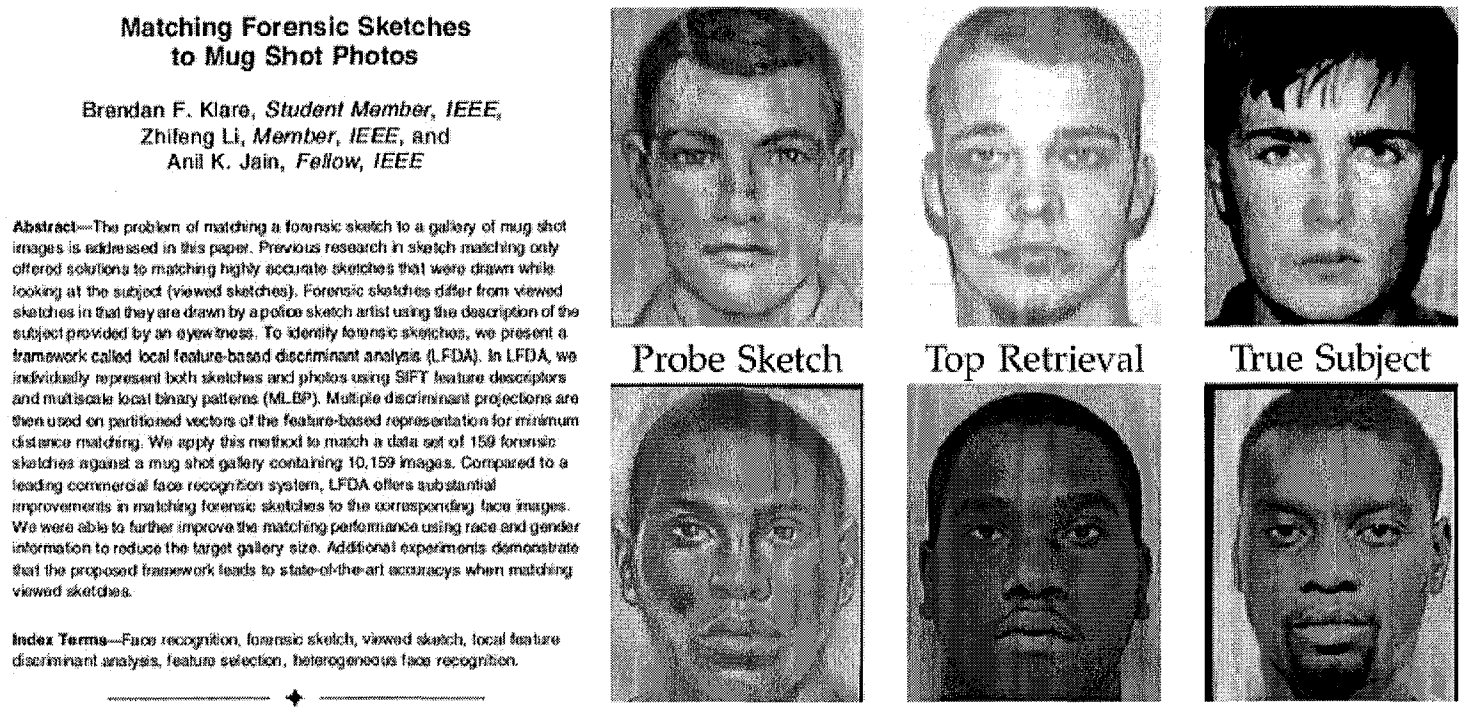

Figure 3: Examples where an imposter, not the subject of the forensic sketch, is returned as the highest ranking face recognition match. (Source: Klare, Li, & Jain (2010), all rights reserved.)

Each of these steps introduces elements of subjective interpretation and room for error.[27] For example, an eyewitness may not remember the shape of the subject's jaw, yet the resulting sketch will necessarily include one. Or the witness may remember the suspect had "bug eyes," something the artist would need to interpret figuratively rather than literally.[28] As a consequence, the resulting sketch may actually look more like someone in the face recognition database other than the subject being searched for, as illustrated in Figure 3.

In this scenario, human review of the face recognition matches will not be able to remove the risk of error. When examining the face recognition results for a possible match, the analyst will have only the sketch to refer back to. The analyst will have no basis to evaluate whether the image accurately represents the subject being searched for. This compounds the risk that the face recognition search will lead to an investigation, if not an arrest, of the wrong person.

2. AN ART OR A SCIENCE? COMPUTER-GENERATED FACIAL FEATURES

A white paper titled "Facial Recognition: Art or Science?" published by the company Vigilant Solutions posits that face recognition systems—even without considering composite sketches—are "[p]art science and part art."[29] The "art" aspect is the process of modifying poor quality images before submitting them to a recognition algorithm to increase the likelihood that the system returns possible matches.[30]

Editing photos before submitting them for search is common practice, as suggested by responses to records requests and a review of the software packages that face recognition vendor companies offer. These documents also illustrate that the edits often go well beyond minor lighting adjustments and color correction, and often

amount to fabricating completely new identity points not present in the original photo.

One technique that the NYPD uses involves replacing facial features or expressions in a probe photo with ones that more closely resemble those in mugshots—collected from photos of other people. Presentations and interviews about FIS include the following examples:

- "Removal of Facial Expression"—such as replacing an open mouth with a closed mouth. In one example provided in a NYPD presentation, detectives conducted "...a Google search for Black Male Model" whose lips were then pasted into the probe image over the suspect's mouth.[31]
- "Insertion of Eyes"—the practice of "graphically replacing closed eyes with a set of open eyes in a probe image," generated from a Google search for a pair of open eyes.[32]
- Mirrored effect on a partial face—copying and mirroring a partial face over the Y axis to approximate the missing features, which may include adding "[e]xtra pixels … to create a natural appearance of one single face."[33]
- "Creating a virtual probe"—combining two face photographs of different people whom detectives think look similar to generate a single image to be searched, to locate a match to one of the people of the combined photograph.[34]
- Using the "Blur effect" on an overexposed or low-quality image—adding pixels to a photo that otherwise doesn't have enough detail "to render a probe that [has] a similar nose, mouth, and brow as that of the suspect in the photo."[35]
- Using the "Clone Stamp Tool" to "create a left cheek and the entire chin area" of a suspect whose face was obscured in the original image.[36]

Another technique that the NYPD and other agencies employ involves using 3D modeling software to complete partial faces and to "normalize" or rotate faces that are turned away from the camera. After generating a 3D model, the software will fill in the missing facial data with an approximation of what it should look like, based on the visible part of what the subject's face looks like as well as the measurements of an "average" face.[37] According to the NYPD, the software creates "a virtual appearance of the suspect looking straight ahead, replicating a pose of a standard mugshot."[38]

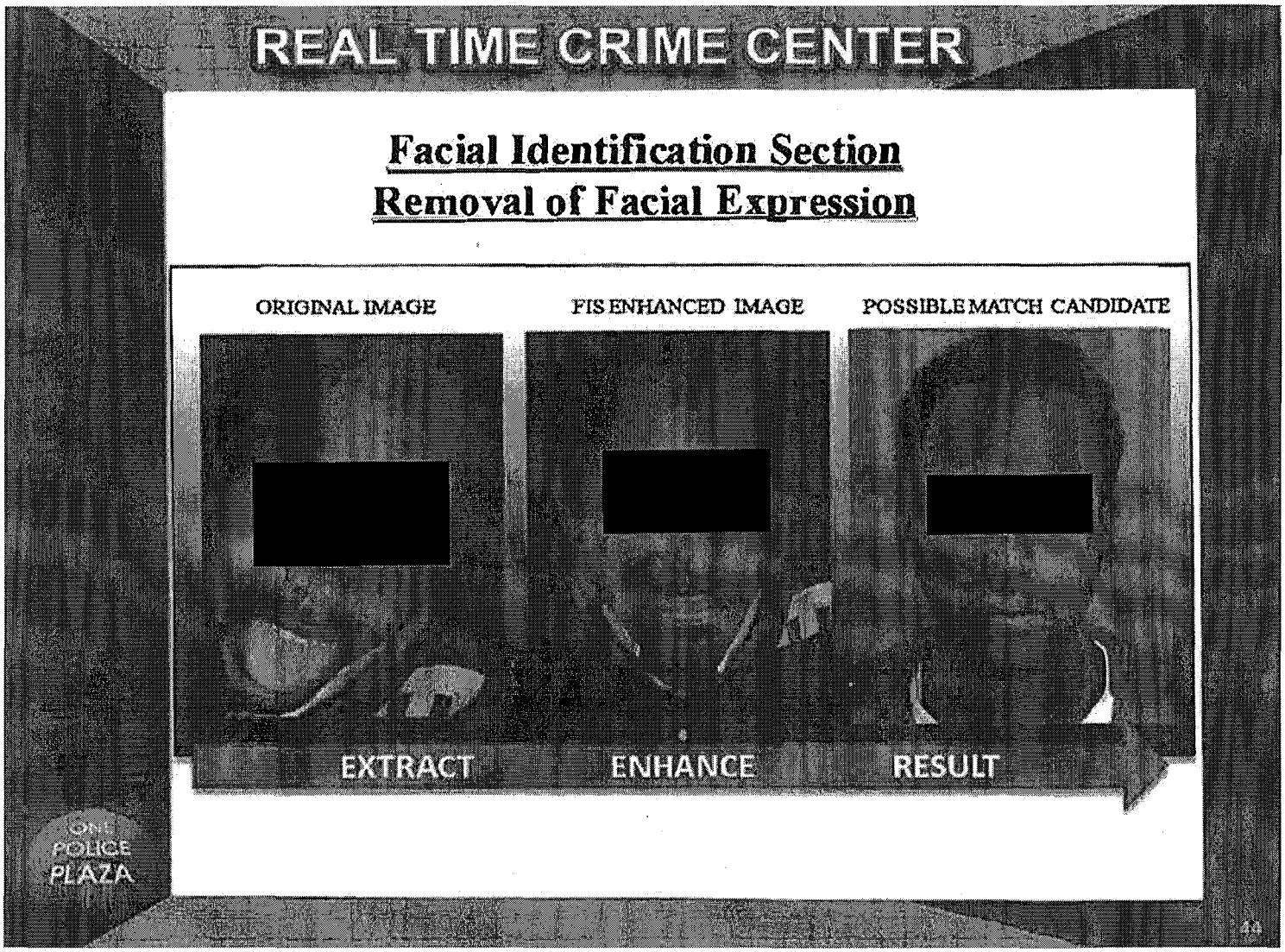

Figure 4: A slide from NYPD FIS describing "Removal of Facial Expression" technique. (Source: NYPD.)

These techniques amount to the fabrication of facial identity points: at best an attempt to create information that isn't there in the first place and at worst introducing evidence that matches someone other than the person being searched for. During a face recognition search on an edited photo, the algorithm doesn't distinguish between the parts of the face that were in the original evidence—the probe photo—and the parts that were either computer generated or added in by a detective, often from photos of different people unrelated to the crime.[39] This means that the original photo could represent 60 percent of a suspect's face, and yet the algorithm could return a possible match assigned a 95 percent confidence rating, suggesting a high probability of a match to the detective running the search.[40]

If it were discovered that a forensic fingerprint expert was graphically replacing missing or blurry portions of a latent print with computer-generated—or manually drawn—lines, or mirroring over a partial print to complete the finger, it would be a scandal.[41] The revelation could lead to thousands of cases being reviewed, possibly even convictions overturned.[42]

3. RESULTS AS "INVESTIGATIVE LEADS ONLY..."

Most agencies do not yet consider face recognition is not yet considered a positive identification. Many law enforcement agencies, the NYPD included, state that the results of a face recognition search are possible matches only and must not be used as positive identification.[43]

In theory, this is a valuable check against possible misidentifications, including those introduced into the system by inputting celebrity comparisons, composite sketches, or other computer-altered photographs that don't accurately represent the person being searched for.

However, in most jurisdictions, officers do not appear to receive clear guidance about what additional evidence is needed to corroborate a possible face recognition match. The NYPD guide states: "Additional investigative steps must be performed in order to establish probable cause to arrest the Subject [sic]" of the face recognition search.[44] But what or how many additional steps are needed, and how independent they must be from the face recognition process, is left undefined.

Absent this guidance, the reality is that suspects are being apprehended almost entirely on the basis of face recognition "possible matches." For example:

- In a recent case, NYPD officers apprehended a suspect and placed him in a lineup solely on the basis of a face recognition search result.[45] The ultimate arrest was made on the basis of the resulting witness identification, but the suspect was only in the lineup because of the face recognition process.
- NYPD officers made an arrest after texting a witness a single face recognition "possible match" photograph with accompanying text: "Is this the guy…?" The witness' affirmative response to viewing the single photo and accompanying text, with no live lineup or photo array ever conducted, was the only confirmation of the possible match prior to officers making an arrest.[46]
- Sheriffs in Jacksonville, Florida, who were part of an an undercover drug sale arrested a suspect on the basis of the face recognition search. The only corroboration was the officers' review of the photograph, presented as the "most likely" possible match from the face recognition system.[47]
- A Metro Police Department officer in Washington, D.C., similarly printed out a "possible match" photograph from MPD's face recognition system and presented that single photograph to a witness for confirmation. The resulting arrest warrant application for the person in the photograph used the face recognition match, the witness confirmation, and a social media post about a possible birth date (month and day only) as the only sources of identification evidence.[48]

There are probably many more examples that we don't know about. These represent a fraction of the cases that have used face recognition to assist in making an identification. The NYPD made 2,878 arrests pursuant to face recognition searches in the first 5.5 years of using the technology.[49] Florida law enforcement agencies, including the Jacksonville Sheriff's Office, run on average 8,000 searches per month of the Pinellas County Sheriff's Office face recognition system, which has been in operation since 2001.[50] Many other agencies do not keep close track of how many times their officers run face recognition searches and whether these searches result in an arrest.

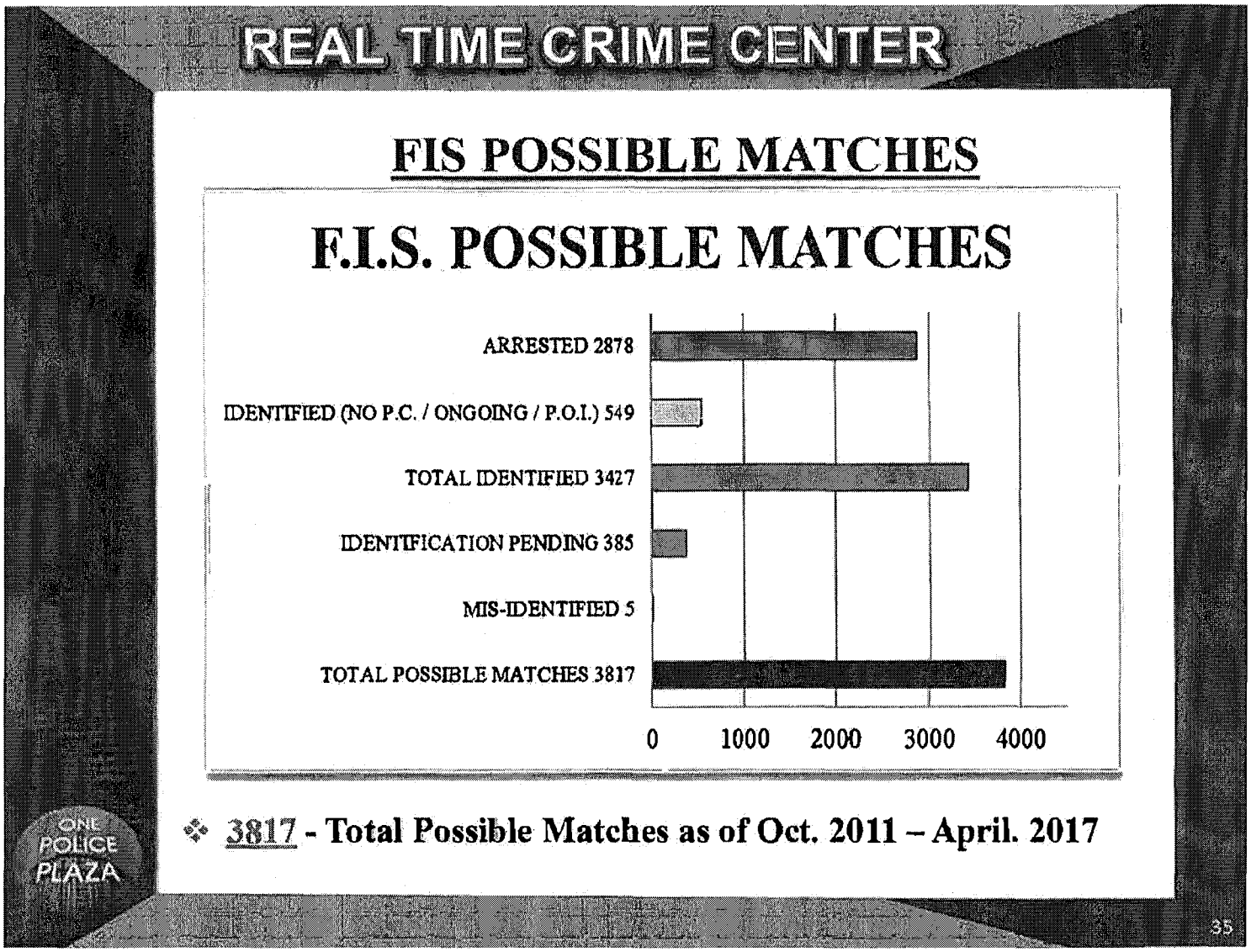

Figure 5: In the first 5.5 years of operation, the NYPD's face recognition system led to 2,878 arrests. NYPD Det. Markiewicz estimates that 8,000 cases will have used a face recognition search in 2018 alone. (Source: NYPD.)

Another valuable check against mistaken identification—and unreliable investigative leads—would be to allow defendants access to the inputs and outputs of a face recognition search that resulted in their arrest. But this does not happen. Even though prosecutors are required under federal law to disclose any evidence that may exonerate the accused, defense attorneys are not typically provided with information about "virtual probes," celebrity doppelgängers, or really any information about the role face recognition played in identifying their client.[51] This is a failure of the criminal justice system to protect defendants' due process.[52]

It may be that many of those arrested on the basis of questionable face recognition searches did in fact commit the crime of which they were accused. But the possibility that they didn't—that the face recognition system identified the wrong person—looms large in the absence of additional, independent police investigation and sufficient access to the evidence by the defense. This is risky, and the consequences will be borne by people investigated, arrested, and charged for crimes they didn't commit.

4. CONCLUSION AND RECOMMENDATIONS

There is no easy way to discover just how broad of a trend this represents—and just how many arrests have been made in large part on the basis of celebrity lookalikes, artist sketches, or graphically altered faces submitted to face recognition systems.[53]

But we can anticipate that the problem will get a lot bigger. Police departments across the country are increasingly relying on face recognition systems to assist their investigations. In addition, an official for the Federal Bureau of Investigation (FBI), which runs its own face recognition system, has indicated that the agency plans to do away with the "investigative lead only" limitation altogether. At a conference in 2018, FBI Section Chief for Biometric Services Bill McKinsey said of the FBI: "We're pretty confident we're going to have face [recognition] at positive ID in two to three years."[54]

In setting this goal, the FBI has assumed that the results of face recognition systems will become more accurate as the algorithms improve. But these improvements won't matter much if there are no standards governing what police departments can feed into these systems. In the absence of those rules, we believe that a moratorium on local, state, and federal law enforcement use of face recognition is appropriate and necessary.

The stakes are too high in criminal investigations to rely on unreliable—or wrong—inputs.

Law enforcement agencies that persist in using face recognition in their investigations should at a minimum take steps to reduce the risk of misidentification and mistake on the basis of unreliable evidence. These steps include:

- Stop using celebrity look-alike probe images. Face recognition is generally considered to be a biometric, albeit an imperfect one. Police cannot substitute one person's biometrics for another's, regardless of whatever passing resemblance they may have.
- Stop submitting artist or composite sketches to face recognition systems not expressly designed for this purpose. Sketches are highly unlikely to result in a correct match—and carry a real risk of resulting in a misidentification that a human review of the possible matches cannot correct.
- Establish and follow minimum photo quality standards, such as pixel density and the percent of the face that must be visible in the original photo, and prohibit the practice of pasting other people's facial features into a probe. Any photo not meeting these minimum standards should be discarded—not enhanced through the addition of new identity points like another person's mouth or eyes.
- If edits to probe images are made, carefully document these edits and their results. Retain all versions of the probe image submitted to the face recognition system for production to the defense.
- Require that any subsequent human review of the face recognition possible match be conducted against the original photo, not a photo that has undergone any enhancements, including color and pose correction.

- As is the practice in some police departments, require double-blind confirmation. The face recognition system should produce an investigative lead only if two analysts independently conclude that the same photo is a possible match.

- Provide concrete guidance to investigating officers about what constitutes sufficient corroboration of a possible match generated by a face recognition system before law enforcement action is taken against a suspect. This should include: mandatory photo arrays; a prohibition on informing witnesses that face recognition was used; and a concrete nexus between the suspect and the crime in addition to the identification, such as a shared address.

- Make available to the defense any information about the use of face recognition, including the original probe photo, any edits that were made to that photo prior to search, the resulting candidate list and the defendant's rank within that list, and the human review that corroborated the possible match.

- Prohibit the use of face recognition as a positive identification under any circumstance.

These recommendations should be considered as minimum requirements, and are made in addition to the broader recommendations the Center on Privacy & Technology made in its 2016 report, *The Perpetual Line-up: Unregulated Police Face Recognition in America* (https://www.perpetuallineup.org/).[55]

As the technology behind these face recognition systems continues to improve, it is natural to assume that the investigative leads become more accurate. Yet without rules governing what can—and cannot—be submitted as a probe photo, this is far from a guarantee. Garbage in will still lead to garbage out.

5. ACKNOWLEDGEMENTS

This report would not be possible without the tireless advocacy of Professor David Vladeck, Stephanie Glaberson, and numerous student lawyers of the Georgetown Law Civil Litigation Clinic, which represents the Center on Privacy & Technology in our public records lawsuit against the New York City Police Department. Only with the assistance of the clinic have we been able to recover thousands of pages of documents regarding use of face recognition technology by the NYPD, even though the agency itself has tried hard to keep its use of this technology hidden from public view.

Critical guidance and close reading were provided by our team of outside reviewers, who will remain anonymous, but who lent us their expertise on New York City policing, criminal litigation, and the technical functioning of face recognition systems. This report would not be possible without the entire team at the Center, who helped in countless ways: Alvaro Bedoya, Laura Moy, Katie Evans, Harrison Rudolph, Jameson Spivack, Gabrielle Rejouis, and Julia Chrusciel. We are also grateful to the Center's research assistants and summer fellows; our copy editor, Joy Metcalf; our design and web development firm, Rootid; and our cover designer, Eve Tyler.

We also acknowledge, with gratitude, the work of our friends and allies at other organizations also striving to shed light on how face recognition technology is used and to prevent powerful police tools from being used in ways that are harmful to individuals and communities. In particular, perhaps no one has done more to address

and expose harmful, secret, and unfair uses of police technology than criminal defense attorneys, many of whom continue to provide us with invaluable guidance.

The Center on Privacy & Technology at Georgetown Law is supported by the Ford Foundation, the Open Society Foundations, the MacArthur Foundation, Luminate, the Media Democracy Fund, and Georgetown University Law Center.

1. NYPD, Real Time Crime Center Facial Identification Section (FIS), presentation by Detective Markiewicz (Sept. 17, 2018) (notes on file with author).
2. *Id.*
3. *See, e.g.*, Eric Sofge, *The End of Anonymity*, Popular Science (Jan. 15, 2014), https://www.popsci.com/article/technology/end-anonymity (https://www.popsci.com/article/technology/end-anonymity) (describing the Pennsylvania system as used in Cheltenham Township, Pa.).
4. *See, e.g.*, Washington County Sheriff's Office, *PSWeb Facial Recognition Training Guide*, 47, *available at* https://www.aclunc.org/docs/20180522_ARD.pdf#page=47 (https://www.aclunc.org/docs/20180522_ARD.pdf#page=47).
5. NYPD, *Facial Identification Section Case #8: Celebrity Comparison*, Document p. 025428 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing). The name and image of the New York Knicks player has been redacted in the files provided to the Center by the NYPD.
6. *See* Phoebe Weston, *Who is YOUR celebrity lookalike? Find out with this online AI tool that reveals your famous doppelganger*, Daily Mail (Mar. 30, 2017) https://www.dailymail.co.uk/sciencetech/article-4363640/Who-celebrity-lookalike-online-tool.html (https://www.dailymail.co.uk/sciencetech/article-4363640/Who-celebrity-lookalike-online-tool.html).
7. Hamza Shaban, *A Google app that matches your face to artwork is wildly popular. It's also raising privacy concerns.*, Washington Post (Jan. 17, 2018), https://www.washingtonpost.com/news/the-switch/wp/2018/01/16/google-app-that-matches-your-face-to-artwork-is-wildly-popular-its-also-raising-privacy-concerns/ (https://www.washingtonpost.com/news/the-switch/wp/2018/01/16/google-app-that-matches-your-face-to-artwork-is-wildly-popular-its-also-raising-privacy-concerns/).
8. Charles Babbage, *Passages from the Life of a Philosopher* 67 (Longman, Green, Longman, Roberts, & Green ed. 1864).
9. *See* Zhifei Wang et al., *Low-resolution face recognition: a review*, 30 The Visual Computer 359, 359–360 (April 2014), *available at* https://link.springer.com/article/10.1007/s00371-013-0861-x (https://link.springer.com/article/10.1007/s00371-013-0861-x).
10. Patrick Grother et al., National Institute of Standards and Technology, *Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification* 2 (Nov. 2018), https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8238.pdf (https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8238.pdf) ("The major result of the evaluation is that massive gains in accuracy have been achieved in the last five years (2013–2018).").
11. Stephen Manusci, The Police Composite Sketch 6–7 (Humana Press 2010).
12. Art Selfie, Google Arts & Culture, https://artsandculture.google.com/camera/selfie (https://artsandculture.google.com/camera/selfie) (last accessed Jan. 28, 2019).
13. Maricopa County Sheriff's Office (MCSO), *Counter-Terrorism Information Center Facial Recognition*, Document p. 014951 (https://drive.google.com/drive/folders/0B-MxWJP0ZmePR2xTYzl4ZWZ0Wkk?usp=sharing); MCSO, *Homeland Security & National Facial Recognition Network Briefing Paper* (Oct. 6, 2008), Document p. 014952 (https://drive.google.com/drive/folders/0B-MxWJP0ZmePR2xTYzl4ZWZ0Wkk?usp=sharing); MCSO, MCSO/ACTIC *Facial Recognition Procedures: Image Records Request*, Document p. 014962 (https://drive.google.com/drive/folders/0B-MxWJP0ZmePR2xTYzl4ZWZ0Wkk?usp=sharing).
14. Washington County Sheriff's Office, *PSWeb Facial Recognition Training Guide*, 47, *available at* https://www.aclunc.org/docs/20180522_ARD.pdf#page=47 (https://www.aclunc.org/docs/20180522_ARD.pdf#page=47).
15. Nlets, *Privacy Impact Assessment Report for the Utilization of Facial Recognition Technologies to Identify Subjects in the Field* (2011), Document p. 016668 (https://drive.google.com/drive/folders/0B-MxWJP0ZmePSi04Wkd5OG1vanc?usp=sharing).
16. Baltimore Police Dep't, *Governor's Office of Crime Control & Prevention (GOCCP) Fact Sheet: Facial Recognition* (Apr. 2015), Document p. 010954 (https://drive.google.com/drive/folders/0B-MxWJP0ZmePbVh2U2tKcmhaRWs?usp=sharing); Maryland

Dep't of Public Safety & Correctional Services, *GOCCP Fact Sheet: Criminal Justice Dashboard* (Apr. 2015), Document p. 011104 (https://drive.google.com/drive/folders/0B-MxWJP0ZmePZndGQmUtVmNDWEU?usp=sharing); Northern Virginia Regional Information System, *LOB #207: NOVARIS* (2016), Document p. 015231 (https://drive.google.com/drive/folders/0B-MxWJP0ZmePU0MwbXhZY0lickE?usp=sharing); Pinellas County Sheriff's Office, *Interagency Use of Facial Recognition...Does it work?*, 43–52, *available at* https://www.aamva.org/uploadedFiles/MainSite/Content/EventsEducation/Event_Materials/2013/2013_Region_II_Conference/061013_10_30_FR_Complete.pdf (https://www.aamva.org/uploadedFiles/MainSite/Content/EventsEducation/Event_Materials/2013/2013_Region_II_Conference/061013_10_30_FR_Complete.pdf).

17. Center for Advancing Retail and Technology, *Cognitec: FaceVACS-VideoScan*, https://www.advancingretail.org/solutions/cognitec (https://www.advancingretail.org/solutions/cognitec). ("Law enforcement professionals can identify individuals in crime scene photos, videos stills and sketches by matching facial images against the agency's mugshot repository"). *See also:* Cognitec, FaceVACS-DBScan LE: Face Recognition Technology for for image and video investigations, and database matching, https://www.cognitec.com/files/layout/downloads/FaceVACS-DBScan-LE-1-1-flyer.pdf (https://www.cognitec.com/files/layout/downloads/FaceVACS-DBScan-LE-1-1-flyer.pdf) ("supports investigation of faces in video footage, still images and sketches").

18. Vigilant Solutions, *FaceSearch,* https://www.vigilantsolutions.com/products/facial-recognition/ (https://www.vigilantsolutions.com/products/facial-recognition/)(last viewed May 13, 2019). Vigilant Solutions is now part of Motorola Solutions. *See* Susan Crandall, *Motorola Solutions Acquires VaaS Holdings, Leader in Data and Image Analytics for Vehicle Location*, Vigilant Solutions (Jan. 7, 2019), https://www.vigilantsolutions.com/motorola-solutions-acquires-vaas-international-holdings-leader-data-image-analytics-vehicle-location/ (https://www.vigilantsolutions.com/motorola-solutions-acquires-vaas-international-holdings-leader-data-image-analytics-vehicle-location/). In a 2008 contract to provide a face recognition solution to Utah's Department of Public Safety, Hummingbird Communications also indicated that its solution can "identify individuals from ... Police Artist Sketches ... or any image from any number or variety of sources." Utah State Analysis and Information Center, *State of Utah Contract with Hummingbird Garden Ranch LLC (*Dec. 22, 2008), Document p. 108705 (https://drive.google.com/drive/folders/0B-MxWJP0ZmePc1QxZGtuNXVaOFU?usp=sharing).

19. Los Angeles County Sheriff's Office, *Facial Recognition & Comparison: Create a Good Source Image*, Document p. 000681 (https://drive.google.com/drive/folders/0B-MxWJP0ZmePOGVuSE1qRFVaM2c?usp=sharing).

20. Anil Jain et al, *Face Recognition: Some Challenges in Forensics*, IEEE Int'l Conference on Automatic Face and Gesture Recognition (Mar. 2011), *available at* https://ieeexplore.ieee.org/document/5771338 (https://ieeexplore.ieee.org/document/5771338).

21. Scott Klum, Hu Han, Anil Jain, & Brendan Klare, *Sketch Based Face Recognition: Forensic vs. Composite Sketches* (2013), *available at* https://openbiometrics.org/publications/klum2013sketch.pdf (https://openbiometrics.org/publications/klum2013sketch.pdf) ("In forensic and biometrics scenarios involving facial sketch to mugshot matching, the standard procedure involves law enforcement officers looking through top-N matches (rather than only considering rank-one retrieval rates). In our experiments, N = 200. We also used the performance of a commercial-off-the-shelf face matcher, FaceVACS v8.2 as a baseline. As shown in Fig. 5, FaceVACS achieves rank-200 retrieval rates of 4.1% and 6.7% for forensic and composite sketches, respectively.")

22. *Id.*

23. Patrick Grother & Mei Ngan, *Face Recognition Vendor Test (FRVT): Performance of Face Identification Algorithms, NIST Interagency Report 8009*, 4 (May 26, 2014) https://nvlpubs.nist.gov/nistpubs/ir/2014/NIST.IR.8009.pdf (https://nvlpubs.nist.gov/nistpubs/ir/2014/NIST.IR.8009.pdf) ("By searching a non-operational set of sketch images against photographs seeded into a population of 640,000 nonmated mugshots, the most accurate algorithms produce the mated photograph only infrequently: The mate is not among the top 50 candidates at the following rates: 73.3% (3M/Cogent), 73.8% (NEC), 78.5% (Toshiba), 80.3% (Morpho), and 81.5% (Neurotechnology).") Note these accuracy results appear much higher than those in the Michigan State University study, likely because NIST used sketches created by an artist viewing the mugshot, not sketches created based on an eyewitness description of the suspect, which is more akin to real-world scenarios. *Id.* at 39–40 ("the fact that the sketches were prepared by an artist viewing the exemplar photograph probably means that the accuracy measurements here represent a "best case" upper bound on accuracy.").

24. *FIS Presentation* (Sept. 17, 2018) (on file with author).

25. Lance Taylor, Ga. Dep't Driver Serv., Moderation of Interagency Use of Facial Recognition... Does it Work? at the 2013 AAMVA Region II Conference 43–53 (June 10, 2013), https://www.aamva.org/uploadedFiles/MainSite/Content/EventsEducation/Event_Materials/2013/2013_Region_II_Conference/061013_10_30_FR_Complete.pdf

(https://www.aamva.org/uploadedFiles/MainSite/Content/EventsEducation/Event_Materials/2013/2013_Region_II_Conference/061013_10_30_FR_Complete.pdf).

26. *See* Anil Jain et. al., *Face Recognition: Some Challenges in Forensics*, IEEE Int'l Conference on Automatic Face and Gesture Recognition (Mar. 19, 2011), https://ieeexplore.ieee.org/document/5771338 (https://ieeexplore.ieee.org/document/5771338).

27. *See* Stephen Manusci, The Police Composite Sketch 22–23 (Humana Press 2010). ("It is essential to realize that a composite sketch is a drawing of a victim's or witness's perception of a perpetrator at the time he or she was observed. It is not meant to be an exact portrait of the suspect. Keep the two words "likeness" and "similarity" in mind at all times … Unfortunately, the composite artist does not have an image of the subject in front of him or her while working. The composite artist needs to rely on the verbal description supplied by the witness. Thus, the look of a composite sketch will range from a portrait-type drawing to a caricature-type sketch, unfortunately never achieving either."), 70 ("How the forensic artist applies these witness and victim impressions and presumptions is certainly subjective.")

28. Forensic sketch artists report that eyewitnesses are likely to use analogies, such as a "horse face" or "bug eyes" when describing subjects. *See, e.g.* Manusci, The Police Composite Sketch, at 73, 86.

29. Rodger Rodriguez, *Facial Recognition: Art or Science?*, Vigilant Solutions (Apr. 4, 2016), http://www2.vigilantsolutions.com/facial-recognition-art-or-science-whitepaper (http://www2.vigilantsolutions.com/facial-recognition-art-or-science-whitepaper). Note that Roger Rodriguez is a former detective with the NYPD, credited for helping implement the NYPD's face recognition program.

30. *Id.*

31. *FIS Presentation* (Sept. 17, 2018) (on file with author); Document pp. 020423–24 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing), 025457 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing).

32. Michelle Taylor, *The Art of Facial Recognition*, Forensic Mag. (Mar. 13, 2017), https://www.forensicmag.com/article/2017/03/art-facial-recognition (https://www.forensicmag.com/article/2017/03/art-facial-recognition). This was corroborated by Detective Tom Markiewicz in a presentation on NYPD FIS September 17, 2018. Det. Markiewicz provided the example where a photo of a suspect whose eyes were turned to the side returned no possible leads. Replacing them with eyes facing towards the camera yielded a possible match. *FIS Presentation* (Sept. 17, 2018) (on file with author) and Document p. 025463 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing).

33. *FIS Presentation* (Sept. 17, 2018) (on file with author), NYPD, *Real Time Crime Center FIS Presentation: Partial Face* (Sept. 17, 2018), Document pp. 020421–22 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing).

34. NYPD, *Real Time Crime Center FIS Presentation: Partial Face* (Sept. 17, 2018, Document pp. 025423, 025466 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing) ("The goal was to create an image which highlighted the pronounced facial features of the suspect in this image. (Hairline, Forehead, Brows, and Nose). The FIS Investigator utilized the head of [redacted] in the previous case mentioned because of the similarities to the hairline and forehead. Both photos were combined within the Photoshop software and a Virtual Probe was created.").

35. NYPD, *Real Time Crime Center FIS Presentation: Partial Face* (date unknown), Document pp. 025469–70 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing).

36. NYPD, *Real Time Crime Center FIS Presentation: Partial Face* (date unknown), Document p. 025458 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing).

37. For a detailed description of 3D modeling software, see NYPD, *Animetrics User Guide* (May 6, 2017), Document pp. 018287–95 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing) and NYPD, *DataWorks Plus FACE Plus Case Management User Guide*, Document p. 018235–39 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing).

38. NYPD, *Sample case 3 of 4 – 3-Dimensional Enhancement* (date unknown), Document p. 025558 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing).

39. *See, e.g.* Felix Juefei-Xu et al., *A Preliminary Investigation on the Sensitivity of COTS Face Recognition Systems to Forensic Analyst-style Face Processing for Occlusions*, IEEE Conf. on Computer Vision and Pattern Recognition Workshop 25, 31 (2015), http://openaccess.thecvf.com/content_cvpr_workshops_2015/W02/papers/Juefei-Xu_A_Preliminary_Investigation_2015_CVPR_paper.pdf (http://openaccess.thecvf.com/content_cvpr_workshops_2015/W02/papers/Juefei-Xu_A_Preliminary_Investigation_2015_CVPR_paper.pdf). (Analysis of the results on edited faces "...questions the credibility of the

FRS since the swapped in part contains biometric information of an other subject. It is questionable and surprising that the FRS uses some other biometric information to its benefit.").

40. Not all face recognition systems present the confidence scores of the photos in the candidate list; and of those that do, some are presented as a percentage and some are on a logarithmic or other scale. Percentages are being used here for illustrative purposes.

41. Latent fingerprints, fingerprints left unintentionally on surfaces and lifted for investigative purposes, may be subject to "preprocessing," editing. However, the goal of this editing is to "improve the retrievable information in a latent image while avoiding any edits that alter critical aspects of this [biometric] information." Paul Lee et al., *Forensic Latent Fingerprint Preprocessing Assessment, NISTIR 8215*, NIST, 5 (June 2018), https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8215.pdf (https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8215.pdf). Improper or overuse of editing tools leads to "accidentally darkened valleys that blend together with nearby ridges, or adding false minutiae or obscuring potentially usable minutiae." *Id.*

42. For a discussion of the potential consequences of misconduct or error by fingerprint examiners, *see* Tom Jackman, *Orlando Fingerprint Examiner Suspended, 2,600 cases possibly affected in latest police lab scandal*, Washington Post, Feb. 27, 2017, https://www.washingtonpost.com/news/true-crime/wp/2017/02/27/orlando-fingerprint-examiner-suspended-2600-cases-possibly-affected-in-latest-police-lab-scandal/ (https://www.washingtonpost.com/news/true-crime/wp/2017/02/27/orlando-fingerprint-examiner-suspended-2600-cases-possibly-affected-in-latest-police-lab-scandal/); Simon A. Cole, *Scandal, Fraud, and the Reform of Forensic Science: The Case of Fingerprint Analysis*, Cole-Monteleone (Proof), Jan. 21, 2017, available at https://wvlawreview.wvu.edu/files/d/94befc60-12bc-47d5-9e72-c8249a566415/cole-monteleone-post-page-proof.pdf (https://wvlawreview.wvu.edu/files/d/94befc60-12bc-47d5-9e72-c8249a566415/cole-monteleone-post-page-proof.pdf).

43. NYPD, *Real Time Crime Center Facial Identification Section (FIS) Notifications, Chief of Detectives Memo No. 3* (Mar. 27 2012), Document pp. 017349–52 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing). ("Real Time Crime Center Facial Identification Section (FIS) analyst determines that Subject is POSSIBLY the suspect whose image is depicted in the video and / or photograph regarding a crime. A FIS Possible Match does NOT constitute a positive identification and does NOT establish probable cause to arrest the Subject. Additional investigative steps MUST be performed in order to establish probable cause to arrest the Subject." (emphasis in original)).

44. NYPD, *Real Time Crime Center Facial Identification Section (FIS) Notifications, Chief of Detectives Memo No. 3* (Mar. 27, 2012), Document pp. 017349–52 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing).

45. Specifics withheld given the ongoing nature of this case.

46. Notice of Motion to Suppress Identification Testimony filed before the Supreme Court of the State of New York, Index number withheld, on file with author. Case specifics are not provided given the ongoing nature of the case.

47. Willie Allen Lynch v. State of Florida, 1D16-3290.

48. Superior Court of the District of Columbia Criminal Division, Affidavit in Support of an Arrest Warrant, on file with author. Specifics withheld given the ongoing nature of the case.

49. NYPD, *Real Time Crime Center, FIS Possible Matches as of Oct. 2011–April 2017*, Document no. 018587 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing) (2878 arrested, 549 additionally identified, 3427 total identified, 385 identification pending, 5 mis-identified, 3817 total possible matches).

50. Pinellas County Sheriff's Office, *Florida's Facial Recognition Network, FACES Training* (2015), Document p. 014396 (https://drive.google.com/drive/folders/0B-MxWJP0ZmePQ2kyMm1LVFVnOTg?usp=sharing).

51. Interviews with public defenders in New York, Washington, D.C, San Francisco, Orlando, Pinellas County, and Baltimore (on file with author). *See generally* Brady v. Maryland, 373 U.S. 83 (1963). *See* Clare Garvie, Alvaro Bedoya & Jonathan Frankle, *The Perpetual Line-Up: Unregulated Police Face Recognition in America* (Oct. 18, 2016), https://www.perpetuallineup.org/findings/transparency-accountability (https://www.perpetuallineup.org/findings/transparency-accountability) (discussing the fact that in the 15 years the Pinellas County Sheriff's Office system has been using face recognition technology, the Public Defenders Office has never received face recognition information as part of *Brady* disclosure).

52. *See* Lynch v. Florida Amici Curiae brief of American Civil Liberties Union, Electronic Frontier Foundation, Georgetown Law's Center on Privacy & Technology, and Innocence Project in support of petitioner, No. SC2019-0298 (2019), *available at* https://efactssc-public.flcourts.org/casedocuments/2019/298/2019-298_notice_86166_notice2dappendix2fattachment20to20notice.pdf (https://efactssc-public.flcourts.org/casedocuments/2019/298/2019-298_notice_86166_notice2dappendix2fattachment20to20notice.pdf).

53. Based on records provided to us from the NYPD, we have an approximate number of the arrests made that involved some face recognition search total, but this is not disaggregated by photo editing or probe photo format. Between October 2011 and April

2017, NYPD arrested 2,878 individuals based in part on a face recognition possible match, and ran a total of 3,817 searches. *See* NYPD, *Real Time Crime Center FIS Possible Matches* (Feb. 9, 2018), Document p. 018587 (https://drive.google.com/drive/folders/1OxzGtFuWBU9PecG2cmpE8QfVwZm9kr22?usp=sharing). In September 2018, FIS Detective Markiewicz anticipated a total of 8,000 NYPD cases to have involved a face recognition search by the end of the year. *FIS Presentation* (Sept. 17, 2018) (on file with author).

54. IJIS Institute National Symposium (Feb. 7, 2018) (on file with author).

55. Clare Garvie, Alvaro Bedoya & Jonathan Frankle, *The Perpetual Line-Up: Unregulated Police Face Recognition in America* (Oct. 18, 2016), https://www.perpetuallineup.org/recommendations (https://www.perpetuallineup.org/recommendations).

APPROVED AS TO FORM AND LEGALITY

CITY ATTORNEY'S OFFICE

OAKLAND CITY COUNCIL

ORDINANCE NO. _____ C.M.S.

INTRODUCED BY COUNCIL PRESIDENT KAPLAN

ORDINANCE AMENDING OAKLAND MUNICIPAL CODE CHAPTER 9.64 TO PROHIBIT THE CITY OF OAKLAND FROM ACQUIRING AND/OR USING FACE RECOGNITION TECHNOLOGY

WHEREAS, according to the American Civil Liberties Union (ACLU), "facial recognition systems are built on computer programs that analyze images of human faces for the purpose of identifying them"; and

WHEREAS, Georgetown Law's Center on Privacy and Technology (CPT) issued a report "Garbage in and Garbage Out" in May 2019, detailing how law enforcement agencies across the country are feeding facial recognition software flawed data stating "when blurry or flawed photos of suspects have failed to turn up good leads, analysts have instead picked a celebrity they thought looked like the suspect, then run the celebrity's photo through their automated face recognition system looking for a lead" and that there are "no rules when it comes to what images police can submit to face recognition algorithms to generate investigative leads"; and

WHEREAS, in a 2018 report by the MIT Lab, "Gender Shades: Intersection Accuracy Disparities in Commercial Gender Classification," the study concluded, using a data set of 1,270 people, that facial recognitions systems worked best on white males and failed most often with the combination of female and dark-skin individuals with error rates of up to 34.7%; and

WHEREAS, the ACLU in 2018, tested a face recognition tool, called "Rekognition," and the software incorrectly matched 28 members of Congress, identifying them as people who had been arrested for a crime; and

WHEREAS, at May 2019 World Economic Forum, George Soros warned of the Chinese government's use of artificial intelligence as an "unprecedented danger" in their monitoring and targeting members of the Uighurs, a Muslim minority group in China; and

WHEREAS, a Stanford study used face recognition technology to see if it could determine sexual orientation of participants and this raises ethical concerns on the use of this technology as a tool for persecution of historically disenfranchised groups; and

WHEREAS, in 2018, the South Wales Police used face recognition software on 170,000 people at a Real Madrid versus Juventus football game and out of 2,470 potential matches with possible criminals, 92% or 2,297 were incorrect; and

WHEREAS, in Baltimore, Maryland, police agencies used face recognition technology to target activists in the aftermath of Freddie Gray's death by law enforcement; and

WHEREAS, in Sri Lanka, authorities using face recognition technology misidentified an American student as a terrorist responsible for killing 300 people in April 2019, widely circulating her image before having to issue an apology; and

WHEREAS, an 18-year-old college student Ousmane Bah, is suing Apple and its contractor, Security Industry Specialists, for allegedly relying on facial recognition systems that misidentified him as a serial shoplifter; and

WHEREAS, police forces in Great Britain are using facial recognition software at festivals and in malls and public spaces and are currently facing legal challenges; and

WHEREAS, the New York City Police Department is currently facing a lawsuit on their use of face recognition technology; and

WHEREAS, United States Representative Alexandria Ocasio-Cortez expressed concerns at a May 2019 House Oversight Committee hearing on facial recognition technology about "the harvesting of facial recognition data without the consent or knowledge of individuals amid the rise of fascism and authoritarianism"; and

WHEREAS, in adopting the City of Oakland's Surveillance and Community Safety Ordinance (Ordinance No. 13489 CMS, codified as Chapter 9.64 of the Oakland Municipal Code), the Oakland City Council (City Council) found that "strong consideration" is required on behalf of the City Council on the" impact such technologies may have on civil rights and civil liberties"; and

WHEREAS, on May 2, 2019, the City of Oakland's Privacy Advisory Commission voted unanimously to support a proposal that would ban the City of Oakland's use of face recognition technology based on empirical evidence on misidentification, concerns around privacy, and studies of misuse by police departments; and

WHEREAS, the City Council finds that ethical dilemmas exist around privacy and the intrusiveness of face recognition technology, the lack of parameters set for the use of this technology by police departments, and that a multitude of studies show that algorithms have gender and race bias; and

NOW, THEREFORE, THE CITY COUNCIL OF THE CITY OF OAKLAND DOES ORDAIN AS FOLLOWS:

SECTION 1. Recitals. The City Council finds and determines the foregoing recitals to be true and correct and hereby adopts and incorporates them into this Ordinance.

SECTION 2. Purpose and Intent. It is the purpose and intent of this Ordinance to prohibit the City's acquisition and/or use of any Face Recognition Technology.

SECTION 3. Amendments to Chapter 9.64 of the Oakland Municipal Code. Oakland Municipal Code Chapter 9.64, is hereby amended as set forth below. Chapter and section numbers and titles are indicated in bold type. Additions are indicated in underline and deletions are shown as ~~strikethrough~~. Provisions of Chapter 9.64 not included herein or not shown in underline or strikethrough type are unchanged.

9.64.010 Definitions. The following definitions apply to this Chapter.

1. "Annual Surveillance Report" means a written report concerning a specific surveillance technology that includes all the following:
    a. description of how the surveillance technology was used, including the type and quantity of data gathered or analyzed by the technology;
    b. Whether and how often data acquired through the use of the surveillance technology was shared with outside entities, the name of any recipient entity, the type(s) of data disclosed, under what legal standard(s) the information was disclosed, and the justification for the disclosure(s);
    c. Where applicable, a breakdown of what physical objects the surveillance technology hardware was installed upon; using general descriptive terms so as not to reveal the specific location of such hardware; for surveillance technology software, a breakdown of what data sources the surveillance technology was applied to;
    d. Where applicable, a breakdown of where the surveillance technology was deployed geographically, by each police area in the relevant year;
    e. A summary of community complaints or concerns about the surveillance technology, and an analysis of the technology's adopted use policy and whether it is adequate in protecting civil rights and civil liberties;
    f. The results of any internal audits, any information about violations or potential violations of the Surveillance Use Policy, and any actions taken in response unless the release of such information is prohibited by law, including but not limited to confidential personnel file information;

g. Information about any data breaches or other unauthorized access to the data collected by the surveillance technology, including information about the scope of the breach and the actions taken in response;
h. Information, including crime statistics, that helps the community assess whether the surveillance technology has been effective at achieving its identified purposes;
i. Statistics and information about public records act requests regarding the relevant subject surveillance technology, including response rates;
j. Total annual costs for the surveillance technology, including personnel and other ongoing costs, and what source of funding will fund the technology in the coming year; and
k. Any requested modifications to the Surveillance Use Policy and a detailed basis for the request.

2. "City" means any department, agency, bureau, and/or subordinate division of the City of Oakland as provided by Chapter 2.29 of the Oakland Municipal Code.

3. "City Staff" means City personnel authorized by the City Administrator or designee to seek City Council approval of surveillance technology in conformance with this Chapter.

4. "Continuing Agreement" means an agreement that automatically renews unless terminated by one (1) party.

5. "Exigent Circumstances" means a law enforcement agency's good faith belief that an emergency involving danger of, or imminent threat of the destruction of evidence regarding, death or serious physical injury to any person requires the use of surveillance technology or the information it provides.

6. <u>"Face Recognition Technology" means an automated or semi-automated process that assists in identifying or verifying an individual based on an individual's face.</u>

7. "Large-Scale Event" means an event attracting ten thousand (10,000) or more people with the potential to attract national media attention that provides a reasonable basis to anticipate that exigent circumstances may occur.

8. "Personal Communication Device" means a mobile telephone, a personal digital assistant, a wireless capable tablet and a similar wireless two-way communications and/or portable internet accessing devices, whether procured or subsidized by a city entity or personally owned, that is used in the regular course of city business.

9. "Police Area" refers to each of the geographic districts assigned to a police commander and as such districts are amended from time to time.

10. "Surveillance" or "Surveil" means to observe or analyze the movements, behavior, data, or actions of individuals. Individuals include those whose identity can be revealed by license plate data when combined with any other record.

11. "Surveillance Technology" means any software, electronic device, system utilizing an electronic device, or similar used, designed, or primarily intended to collect, retain, analyze, process, or share audio, electronic, visual, location, thermal, olfactory, biometric, or similar information specifically associated with, or capable of being associated with, any individual or group. Examples of surveillance technology include, but is not limited to the following: cell site simulators (Stingrays); automatic license plate readers; gunshot detectors (ShotSpotter); facial recognition software; thermal imaging systems; body-worn cameras; social media analytics software; gait analysis software; video cameras that record audio or video, and transmit or can be remotely accessed. It also includes software designed to monitor social media services or forecast criminal activity or criminality, biometric identification hardware or software.

    A. "Surveillance technology" does not include the following devices or hardware, unless they have been equipped with, or are modified to become or include, a surveillance technology as defined above:
        1. Routine office hardware, such as televisions, computers, credit card machines, badge readers, copy machines, and printers, that is in widespread use and will not be used for any surveillance or law enforcement functions;
        2. Parking Ticket Devices (PTDs);
        3. Manually-operated, non-wearable, handheld digital cameras, audio recorders, and video recorders that are not designed to be used surreptitiously and whose functionality is limited to manually capturing and manually downloading video and/or audio recordings;
        4. Surveillance devices that cannot record or transmit audio or video or be remotely accessed, such as image stabilizing binoculars or night vision goggles;
        5. Manually-operated technological devices used primarily for internal municipal entity communications and are not designed to surreptitiously collect surveillance data, such as radios and email systems;
        6. City databases that do not contain any data or other information collected, captured, recorded, retained, processed, intercepted, or analyzed by surveillance technology, including payroll, accounting, or other fiscal databases.
        7. Medical equipment used to diagnose, treat, or prevent disease or injury.
        8. Police department interview room cameras.
        9. Police department case management systems.
        10. Police department early warning systems.
        11. Personal communication devices that have not been modified beyond stock manufacturer capabilities in a manner described above.

12. "Surveillance Impact Report" means a publicly-released written report including at a minimum the following:
    a. Description: information describing the surveillance technology and how it works, including product descriptions from manufacturers;

b. Purpose: information on the proposed purposes(s) for the surveillance technology;
c. Location: the location(s) it may be deployed, using general descriptive terms, and crime statistics for any location(s);
d. Impact: an assessment of the technology's adopted use policy and whether it is adequate in protecting civil rights and liberties and whether the surveillance technology was used or deployed, intentionally or inadvertently, in a manner that is discriminatory, viewpoint-based, or biased via algorithm;
e. Mitigations: identify specific, affirmative technical and procedural measures that will be implemented to safeguard the public from each such impacts;
f. Data Types and Sources: a list of all types and sources of data to be collected, analyzed, or processed by the surveillance technology, including "open source" data, scores, reports, logic or algorithm used, and any additional information derived therefrom;
g. Data Security: information about the steps that will be taken to ensure that adequate security measures are used to safeguard the data collected or generated by the technology from unauthorized access or disclosure;
h. Fiscal Cost: the fiscal costs for the surveillance technology, including initial purchase, personnel and other ongoing costs, and any current or potential sources of funding;
i. Third Party Dependence: whether use or maintenance of the technology will require data gathered by the technology to be handled or stored by a third-party vendor on an ongoing basis;
j. Alternatives: a summary of all alternative methods (whether involving the use of a new technology or not) considered before deciding to use the proposed surveillance technology, including the costs and benefits associated with each alternative and an explanation of the reasons why each alternative is inadequate; and
k. Track Record: a summary of the experience (if any) other entities, especially government entities, have had with the proposed technology, including, if available, quantitative information about the effectiveness of the proposed technology in achieving its stated purpose in other jurisdictions, and any known adverse information about the technology (such as unanticipated costs, failures, or civil rights and civil liberties abuses).

13. "Surveillance Use Policy" means a publicly-released and legally enforceable policy for use of the surveillance technology that at a minimum specifies the following:
    a. Purpose: the specific purpose(s) that the surveillance technology is intended to advance;
    b. Authorized Use: the specific uses that are authorized, and the rules and processes required prior to such use;
    c. Data Collection: the information that can be collected by the surveillance technology. Where applicable, list any data sources the technology will rely upon, including "open source" data;
    d. Data Access: the category of individuals who can access or use the collected information, and the rules and processes required prior to access or use of the information;

e. Data Protection: the safeguards that protect information from unauthorized access, including encryption and access control mechanisms;
f. Data Retention: the time period, if any, for which information collected by the surveillance technology will be routinely retained, the reason such retention period is appropriate to further the purpose(s), the process by which the information is regularly deleted after that period lapses, and the specific conditions that must be met to retain information beyond that period;
g. Public Access: how collected information can be accessed or used by members of the public, including criminal defendants;
h. Third Party Data Sharing: if and how other city departments, bureaus, divisions, or non-city entities can access or use the information, including any required justification or legal standard necessary to do so and any obligations imposed on the recipient of the information;
i. Training: the training required for any individual authorized to use the surveillance technology or to access information collected by the surveillance technology;
j. Auditing and Oversight: the mechanisms to ensure that the Surveillance Use Policy is followed, including internal personnel assigned to ensure compliance with the policy, internal recordkeeping of the use of the technology or access to information collected by the technology, technical measures to monitor for misuse, any independent person or entity with oversight authority, and the legally enforceable sanctions for violations of the policy; and
k. Maintenance: The mechanisms and procedures to ensure that the security and integrity of the surveillance technology and collected information will be maintained.

9.64.045 Prohibition on City's Acquisition and/or Use of Face Recognition Technology

A. Notwithstanding any other provision of this Chapter (9.64), it shall be unlawful for the City or any City staff to obtain, retain, request, access, or use:
1. Face Recognition Technology; or
2. Information obtained from Face Recognition Technology.

B. City staff's inadvertent or unintentional receipt, access of, or use of any information obtained from Face Recognition Technology shall not be a violation of this Section 9.64.045 provided that:
1. City staff did not request or solicit the receipt, access of, or use of such information; and
2. City staff logs such receipt, access, or use in its Annual Surveillance Report as referenced by Section 9.64.040. Such report shall not include any personally identifiable information or other information the release of which is prohibited by law.

SECTION 4. Severability. If any section, subsection, sentence, clause or phrase of this Ordinance is for any reason held to be invalid or unconstitutional by decision of any court of competent jurisdiction, such decision shall not affect the validity of the remaining portions of the Chapter. The City Council hereby declares that it would have passed this Ordinance and each section, subsection, clause or phrase thereof irrespective of the fact that one or more other sections, subsections, clauses or phrases may be declared invalid or unconstitutional.

SECTION 5. Effective Date. This ordinance shall become effective immediately on final adoption if it receives six or more affirmative votes; otherwise it shall become effective upon the seventh day after final adoption. effective immediately upon final adoption.

IN COUNCIL, OAKLAND, CALIFORNIA,

PASSED BY THE FOLLOWING VOTE:

AYES - FORTUNATO BAS, GALLO, GIBSON MCELHANEY, KALB, REID, TAYLOR, THAO AND PRESIDENT KAPLAN

NOES –

ABSENT –

ABSTENTION –

ATTEST:_____
LATONDA SIMMONS
City Clerk and Clerk of the Council of the City of Oakland, California

Date of Attestation: _____

NOTICE AND DIGEST

ORDINANCE AMENDING OAKLAND MUNICIPAL CODE CHAPTER 9.64 TO PROHIBIT THE CITY OF OAKLAND FROM ACQUIRING AND/OR USING FACE RECOGNITION TECHNOLOGY

This ordinance amends Oakland Municipal Code Chapter 9.64 to prohibit the City of Oakland from acquiring and/or using face recognition technology. The ordinance also defines the term "Face Recognition Technology."