Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers

Author: Joy Adowaa Buolamwini

B.S. in Computer Science, Georgia Institute of Technology (2012)
M.Sc. in Learning and Technology, University of Oxford (2014)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning, in partial fulfillment of the requirements of the degree of Master of Science at the Massachusetts Institute of Technology

September 2017

Signature of Author Signature redacted

Program in Media Arts and Sciences
August 10, 2017

Certified by Signature redacted

Ethan Zuckerman
Associate Professor of the Practice in Media Arts and Sciences
Director, Center for Civic Media
Massachusetts Institute of Technology

Accepted by Signature redacted

Prof Patricia Maes
Academic Head
Program in Media Arts and Sciences
Massachusetts Institute of Technology

Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers

Author: Joy Adowaa Buolamwini

Submitted to the Program in Media Arts and Sciences,School of Architecture and Planning, on August 10, 2017 in partial fulfillment of the requirements of the degree of Master of Science at the Massachusetts Institute of Technology

ABSTRACT

This thesis (1) characterizes the gender and skin type distribution of IJB-A, a government facial recognition benchmark, and Adience, a gender classification benchmark, (2) outlines an approach for capturing images with more diverse skin types which is then applied to develop the Pilot Parliaments Benchmark (PPB), and (3) uses PPB to assess the classification accuracy of Adience, IBM, Microsoft, and Face++ gender classifiers with respect to gender, skin type, and the intersection of skin type and gender.

The datasets evaluated are overwhelming lighter skinned: 79.6% - 86.24%. IJB-A includes only 24.6% female and 4.4% darker female, and features 59.4% lighter males. By construction, Adience achieves rough gender parity at 52.0% female but has only 13.76% darker skin. The Parliaments method for creating a more skin-type-balanced benchmark resulted in a dataset that is 44.39% female and 47% darker skin. An evaluation of four gender classifiers revealed a significant gap exists when comparing gender classification accuracies of females vs males (9 - 20%) and darker skin vs lighter skin (10 - 21%). Lighter males were in general the best classified group, and darker females were the worst classified group. 37% - 83% of classification errors resulted from the misclassification of darker females. Lighter males contributed the least to overall classification error (.4% - 3%).

For the best performing classifier, darker females were 32 times more likely to be misclassified than lighter males. To increase the accuracy of these systems, more phenotypically diverse datasets need to be developed. Benchmark performance metrics need to be disaggregated not just by gender or skin type but by the intersection of gender and skin type. At a minimum, human-focused computer vision models should report accuracy on four subgroups: darker females, lighter females, darker males, and lighter males.

The thesis concludes with a discussion of the implications of misclassification and the importance of building inclusive training sets and benchmarks.

Thesis supervisor: Ethan Zuckerman
Title: Associate Professor of the Practice in Media Arts and Sciences
Director, Center for Civic Media

Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers

Author: Joy Adowaa Buolamwini

Thesis Reader Signature redacted

Hal Abelson
Class of 1922 Professor of Computer Science and Engineering
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers

Author: Joy Adowaa Buolamwini

Thesis Reader

Signature redacted

Mitchel Resnick
LEGO Papert Professor of Learning Research
Associate Academic Head, Program in Media Arts and Sciences
Massachusetts Institute of Technology

ACKNOWLEDGMENTS

To my parents, Frema and John Buolamwini, thank you for instilling in me an appreciation for the search of truth through art and science.

To my thesis committee, Ethan Zuckerman, Hal Abelson, and Mitch Resnick, thank you for setting high expectations, sharpening my thinking, and sharing your knowledge, time, and insights. Gender Shades is a reflection of your high-impact approach to academic mentorship and your appreciation of artistic freedom.

To Timnit Gebru, thank you for bringing me into the world of computer vision and sharing generously of your technical expertise and tenacious spirit.

To Cathy O'Neil, thank you for writing *Weapons of Math Destruction.* Hearing you speak at the Harvard Book Store was the nudge I needed to develop the Algorithmic Justice League.

To Boaz Sender, thank you and Bocoup for your faith in the seeds of the Algorithmic Justice League and partnering with us to build a platform to enable more inclusive technology.

To Karen Adams, thank you for copy editing yet another thesis. Your suggestions always add clarity to my writing. But more importantly, thank you for your instistence that I apply to MIT and start the journey towards a terminal degree.

To Aubry Threlkeld, thank you for introducing me to the world of gender studies, sharing your library, and encouraging me to persist through the exploration of intersectionaility and machine learning.

To Christopher Morris-King, thank you for inspiring the idea for the Parliaments Benchmark through countless conversations and brainstorm sessions.

To Nicole Hughes, thank you for helping me find time to write, speak, and recharge throughout this process. From Memphis to Barcelona to Edinburgh and now Boston, may our orbits continue to intersect.

To Deborah Raji, thank you for helping with reference verification. Your enthusiasm in creating more inclusive computer vision models gives me hope for the future.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

Whoever codes the system, embeds her views.

1.1 Unmasking Bias

Artificial Intelligence - which is infiltrating society, helping determine who is hired, fired, granted a loan, or even how long someone spends in prison - has a bias problem. The scope and nature of this problem is largely hidden. Selecting training data to fine-tune artificial intelligence systems is a pivotal part of developing robust predictive models. However, bias reflecting social inequities in training data can embed unintended bias in the models that are created. Furthermore, benchmark datasets are used to assess progress on specific tasks like machine translation and pedestrian detection. Unrepresentative benchmark datasets and aggregate accuracy metrics can provide a false sense of universal progress on these tasks.

To ensure artificial intelligence works well for a given target population, we need to evaluate the composition of training and benchmark datasets while making subgroup performance checks part of common practice. Unaddressed, bias in training data can result in algorithms that perform poorly on underrepresented groups. Unaltered, skewed benchmarks can mask performance difference between genders, ethnicities, and other demographic categories. In the case of computer vision powered by artificial intelligence, skewed benchmarks and aggregate metrics can mask performance disparities between individuals with different phenotypic features like skin type and facial geometry. This thesis focuses exclusively on facial analysis in computer vision to demonstrate the more general need for inclusive benchmark datasets and disaggregated accuracy metrics across a range of human-focused automated tasks.

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

Automated facial recognition is not the only area where artificial intelligence is being applied to computer vision. Self-driving cars are equipped with pedestrian tracking systems that partially rely on computer vision to prevent automotive fatalities. Ensuring pedestrian tracking works well on all subjects is not just a technical challenge but also a public safety imperative. Beyond the automotive industry, artificial intelligence is increasingly used in healthcare, employment decision-making, and government surveillance – parallel problems may exist in these fields where judgments made about a subgroup of the population are less reliable than judgments made about a majority group.

Given the high stakes, companies, research communities, and foundations are dedicating substantial resources to explore algorithmic fairness. Algorithmic fairness is an intersectional domain that spans many areas including computer science, ethics, policy, and civics. Institutions like Data & Society[1] and AI Now[2] are investigating the social issues stemming from the rise of data-centric technological advances. The Fairness Accountability and Transparency in Machine Learning (FATML) research community[3] supports computer scientists and social scientists who are grappling with the social impact of machine learning, the leading approach being used in artificial intelligence. The John S. and James L. Knight Foundation, Omidyar Network, LinkedIn founder Reid Hoffman, and others announced a $27 million Ethics and Governance of Artificial Intelligence Fund focused on artificial intelligence research for the public interest.[4] The Partnership for AI is bringing together industry, research, and civic actors to provide peer support in creating ethical artificial intelligence.[5] The unifying goal of these initiatives is to build theoretical frameworks and best practices for developing technical systems that minimize disparate impacts or inadvertent discrimination.

Because algorithmic fairness is based on different contextual assumptions and optimizations for accuracy, this thesis aims to show why we need rigorous reporting on the accuracy rates on which algorithmic fairness debates center. The work focuses on increasing phenotypic and demographic representation in datasets and algorithmic evaluation. Inclusive benchmark datasets and subgroup accuracy reports will be necessary to increase transparency and accountability in artificial intelligence. For human-focused computer vision, I define transparency as providing information on the demographic and phenotypic composition of training and benchmark datasets. I 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 I do not focus on in this thesis. Nonetheless, the findings from this work concerning dataset representation and algorithmic evaluation provide empirical support for increased demographic and phenotypic transparency and accountability in artificial intelligence.

[1] https://datasociety.net/

[2] https://artificialintelligencenow.com/

[2] https://artificialintelligencenow.com/

[3] http://www.fatml.org

[3] http://www.fatml.org

[4] https://knightfoundation.org/press/releases/knight-foundation-omidyar-network-and-linkedin-founder-

[4] https://knightfoundation.org/press/releases/knight-foundation-omidyar-network-and-linkedin-founder-reid-hoffman-create-27-million-fund-to-research-artificial-intelligence-for-the-public-interest

[5] https://www.partnershiponai.org/

1.2 Full-Spectrum Inclusion: Curation, Testing, and Reporting

Left unchecked, algorithms can reproduce structural bias and manifest discrimination. To as suitability for real world use, high-impact algorithms will require ongoing monitoring throughout their lifecycles. Checking for structural bias should be part of the design, development, deployment, testing, and maintenance of any systems that have profound consequences for the general public.

For machine learning algorithms that rely on training examples, data is destiny. Biased data that is not repaired can lead to biased results. Full-spectrum curation involves collecting diverse data that is inclusive of a target population for a task at hand. Full-spectrum curation also necessitates continual interrogation of assumptions about representation. We cannot assume data collected from one demographic group can be extrapolated to other groups. Even within a demographic group, we need to account for intragroup variation. For instance, race is a common demographic category in the United States. However, people who self-identify with the same race can exhibit a range of phenotypic features such as skin color and facial geometry. In computer vision, full-spectrum curation for facial analysis tasks must rigorously attend to intragroup variation as well as intergroup differences.

Despite our best efforts, there will always be unanticipated failure cases. The complement to full-spectrum curation is full-spectrum testing. Full-spectrum testing provides an opportunity to catch errors or inaccuracies that were not previously considered. These failures can provide valuable insights into ways to improve outcomes. More inclusive benchmark datasets will facilitate full-spectrum testing.

Ideally, full-spectrum curation and testing explicitly attend to individual differences to create technology that works well for all of humanity. In reality, reflecting the full breadth of human variation in a single training set or benchmark is challenging and not always desirable. Creating group-specific datasets can at times yield more accurate results. Sparse and coarse data for historically excluded groups can also pose a challenge for full-spectrum curation and testing. Because of challenges with data collection and representation, transparency about the composition of training sets and the nature of benchmarks need to be reported for high-impact algorithms that effect areas like employment, health, and security.

While full-spectrum curation and testing are guiding ideals, in practice, full-spectrum reporting is a process that can help mark progress towards the ideal of full-spectrum inclusion. Generalizability is a common goal for artificial intelligence. Artificial intelligence practitioners aim to create robust systems that can work well on a given task. When a face detection algorithm is created, the goal is for the algorithm to work well on any human face. Even if the goal of generalizability is not met or demographic performance varies, technology such face detection software is presented as general purpose. This assumption of generalizability risks masking important limitations that can hide inadvertent discrimination. In this thesis, I understand "discrimination" in terms of disparate impact. Intent is not essential for a system to discriminate – it merely has to produce disparate impacts for different social groups to be discriminatory. Echoing previous calls for transparency in the use of algorithms[6] (Crawford et al., 2016; O'Neil, 2016), high-impact algorithms need to have publicly available reports that document potential for disparate impact and provide test results on group specific benchmarks (Garvie et al., 2016).[7]

1.3 Intersectional Benchmarking

One way to bring increase transparency in artificial intelligence is to examine datasets for bias. For automated facial analysis tasks like gender classification, benchmark datasets exist to assess performance. However, as this thesis will highlight, existing facial analysis benchmarks are not always reflective of target populations and do not explicitly attend to demographic or phenotypic performance. We need to develop more robust benchmarks that better reflect the true performance of human-focused computer vision algorithms. In this thesis, I present the need for subgroup performance reports that can illuminate disparities hidden in aggregate performance metrics. The goal of full-spectrum testing is to assess bias that can lead to social identity based discrimination. Prior work has shown that if a facial analysis algorithm works well on male faces one cannot assume it works as well on female faces (Ngan & Grother, 2015).

Still, since individuals embody multiple social identities, subgroup analysis should not stop at evaluating one social identity in isolation. Instead, the intersection of identities should also be examined. Kimberlé Crenshaw introduced the term "intersectionality" to describe how intersecting identities like gender and race interrelate and can result in unique dimensions of discrimination not fully captured by gender or race alone (Collins, 2015). For example, roughly half of adults have the social identity of being a woman in the United States. While all women may experience some form of sexism –subpar treatment based on perceived gender– the experience of being an Asian-American woman, an African-American woman, or a White woman is not the same. Thus, the experience of one should not be substituted for another, nor should the experience of one group be used to reflect the experiences of all women or all people.

In machine learning, training data provides an algorithm with experience. Benchmark data is used to validate an algorithm's suitability for use in the real world. Since an algorithm learns to represent the world based on the data it is trained on and benchmarks are meant to reflect real-world scenarios, I apply the social science lens of intersectionality to computer vision to increase rigor in the evaluation of algorithmic performance. Crenshaw's argument for intersectional analysis as it relates to cases of discrimination in a legal context (1989) informs the curation of data and the approach to algorithmic evaluation used in this work. I examine the intersection of the social

[6] In her book *Weapons of Math Destruction*, Cathy O'Neil talks about how WMDs –widespread, mysterious, and destructive- algorithms are intentionally obfuscated by industry players and the need to increase public awareness about their use and impact

[7] In the "Perpetual Lineup Report" , devoloping demographic specific acuracy benchmarks is among the salient recommendations to provide oversight to the use of automated facial analysis by law enforcement.

identities of gender and nationality, and I also explore the intersection of the social demographic of gender and the phenotypic attribute of skin type. Because computer vision algorithms analyze images and for a given nationality, race, or ethnicity there can be many visual differences within the social group, I explore skin type as a physical feature that can signal social identity[8].

An intersectional approach to algorithmic performance evaluation leads to more nuanced questions. Instead of merely asking "if a facial analysis algorithm works well on male faces, does it work as well on female faces?", we move to questions like "if a facial analysis algorithm works well on lighter female faces, does it work as well on all female faces or darker female faces?" Benchmarks that lack demographic and phenotypic diversity can make these questions difficult to answer due to a lack of representation. Without knowing if our facial analysis algorithms achieve an agreed upon threshold of performance for intersectional subgroups, we risk creating systems that are only optimized for individuals who are best represented in existing training and benchmark datasets. Intersectional benchmarking as presented in this thesis will enable researchers and practitioners to better assess our progress on creating facial analysis algorithms that work well on a larger portion of humanity.

1.4 Algorithmic Justice League : Fighting the Coded Gaze

Alongside the research efforts presented in this thesis, I founded the Algorithmic Justice League (AJL) to help increase transparency and accountability in artificial intelligence. The goal of AJL is to create a world with more inclusive technology by fighting "the coded gaze", my term for bias in artificial intelligence that can lead to exclusionary experiences or discriminatory practices. The coded gaze is a view that posits any technology created by humans will reflect individual or collective values, priorities and if unchecked, prejudices. To address bias, the coded gaze must be acknowledged. Exploring the coded gaze can inform ways to make artificial intelligence more inclusive.

AJL fights the coded gaze through a bias-busting strategy that (1) highlights bias by raising public awareness on the shortcomings of artificial intelligence through media production, public talks, and exhibitions, (2) identifies bias by conducting research and building tools that practitioners and researchers can use to check datasets and algorithms for demographic and phenotypic bias, and (3) mitigates bias by providing inclusive benchmarks and best practices to create more inclusive artificial intelligence.

Highlighting Bias
To highlight bias, AJL engages in media and advocacy work concerning the need for algorithmic fairness, accountability, and transparency. I first articulated the term "the coded gaze" in a May 2016 Medium article that called for an Inclusive Code movement that would eventually become the Algorithmic Justice League. Since November of 2016, I have created explainer videos, presented an art exhibition, and given numerous talks.

[8] Skin type alone does not necessarily signify membership to a specific nationality, race, or ethnicity.

The most notable of these public outreach initiatives is a TED.com talk that to date has received over 750,000 views and spurred public discussion about algorithmic bias.

Identifying and Mitigating Bias
Alongside consciousness raising initiatives, AJL is working on building tools to help identify and mitigate bias with a focus on facial analysis technology. We start with facial analysis technology because of its widespread use by law enforcement and in consumer products. This thesis focuses primarily on identifying demographic and phenotypic bias in face datasets. The research provides an in-depth approach for evaluating algorithmic performance to inform strategies for mitigating bias.

1.5 Research Questions

This thesis focuses on the intersectional evaluation of face datasets and gender classifiers with respect to gender and skin type. Because benchmarks are used to assess the state-of-the-art for a given computer vision task, I examine the phenotypic and demographic composition of existing benchmarks. What demographic and phenotypic groups are well represented in these benchmark face datasets, and which groups are underrepresented? I then focus on evaluating the performance of four gender classification algorithms (gender classifiers) on the novel Pilot Parliaments Benchmark (PPB) created for this thesis to explore intersectional benchmarking. Even though there are a number of automated facial analysis tasks, for efficiency, I focus on binary gender classification, which by definition reduces gender identity to a binary construct. The algorithms evaluated classify faces as either essentially male or female (Fuss, 1989). In this work, the questions I address about dataset representation and algorithmic performance are as follows:

Dataset Representation: For the IJB-A, Adience, and PPB Benchmarks

- What is the gender composition of the unique subjects in the dataset?
- What is the skin type composition of the unique subjects in the dataset?
- What is the intersectional gender and skin type composition of unique subjects in the dataset?
- How do the intersectional gender and skin type compositions of the datasets compare?

Algorithmic Performance: For Microsoft, IBM, Adience and Face++

- What are the classification accuracy rates by gender on PPB in aggregate? By region? By nationality?
- What are the classification accuracy rates by skin type on PPB in aggregate? By region? By nationality?
- What are the intersectional classification accuracy rates by gender and skin type?

- How much does each intersectional subgroup (darker-skinned females, darker-skinned males, lighter skinned-females, and lighter-skinned males) contribute to the error rates of each classifier?

1.6 Thesis Contributions

1.6.1 Intersectional Demographic and Phenotypic Performance Evaluation

To my knowledge, this is the first investigation of the intersectional performance of gender classification in relation to both gender and skin type. Studies that look at subgroup differences in automated facial analysis focus on demographic factors like race or nationality without attending to interclass or intraclass phenotypic differences like skin type.

1.6.2 Fitzpatrick Skin Type Labeled Datasets

To evaluate skin type representation in the selected benchmarks and assess phenotypic algorithmic performance, I first provide new labels to existing datasets that do not have skin type labels. I use the Fitzpatrick Scale developed to classify skin responses to UV radiation as a skin type labeling scheme and assess its advantages and limitations. I use the scale along with gender annotations to produce gender (female, male) and skin type (I -VI) labels for 1270 unique subjects in the Pilot Parliaments Benchmark dataset constructed for this thesis. I also produce new skin type annotations for the 500 unique subjects in the government IJB-A benchmark and 2194 unique subjects in the research Adience benchmark.

1.6.3 Phenotypically Diverse Face Curation Methodology

Collecting images of diverse faces is a nontrivial endeavor that is complicated by licensing considerations, underrepresentation in publicly available face images, and limited discoverability of alternative image sources. This thesis defines a diverse face curation methodology used to create the intersectional Pilot Parliaments Benchmark that can be expanded to develop more inclusive training and benchmark datasets for facial analysis algorithms.

1.6.4 Complementary Contributions

In working to highlight algorithmic bias, I produced a number of media artifacts to spur public discourse about the need for inclusive artificial intelligence. These artifacts include "The Coded Gaze: Unmasking Bias" mini-documentary[9], which debuted at the Museum of Fine Arts Boston and a widely viewed TED Talk. In partnership with Bocoup

[9] "The Coded Gaze" is available at https://www.youtube.com/watch?v=162VzSzzoPs

Foundation, I have also established AJL.ai to gather crowd annotations on existing facial datasets. AJL.ai currently focuses on demographic labels. The lessons derived from phenotypic skin type labeling of the Pilot Parliaments Benchmark introduced in this thesis will be used to develop a system for crowd-sourced phenotypic labels.

1.7 Overview of Thesis

Debates around the need for fairness in artificial intelligence have largely focused on predictive models for constructs like recidivism risk (Angwin et al., 2016), instead of the accuracy of models that assess verifiable traits like gender. However, disparities in accuracy between different demographic and phenotypic groups can also be defined as unfair. For this thesis, fairness is defined as having comparable classification accuracy rates across intersectional subgroups. In exploring gender classification in particular, I argue that to adequately assess fairness we need disaggregated and intersectional accuracy metrics for human-focused computer vision models.

To begin, Chapter 2 examines common facial analysis tasks, how structural bias can lead to subpar performance for underrepresented groups, the potential impacts of gender misclassification, and ways in which gender classification can be abused. Chapter 3 provides a literature review on breakthroughs in automated facial analysis that influence gender classification, the evolution of benchmark face datasets, related work on demographic performance evaluation of facial recognition algorithms, and efforts to curate more inclusive face training data. Chapter 4 situates the evaluation of facial analysis algorithms in evolving discussions concerning fairness, accountability, and transparency in machine learning. The chapter also presents existing measures for discrimination that can be used to assess the demographic and phenotypic performance of facial analysis algorithms. Chapter 5 outlines the rationale for creating an alternative gender classification benchmark, a curation methodology for creating an intersectional gender and skin type benchmark, the selection of demographic and phenotypic labels for the benchmark, and the limitations of the benchmark. The chapter concludes with a comparison of the gender and skin type distribution between the new benchmark and existing ones. Chapter 6 justifies the selection of four gender classification algorithms for evaluation and details their performance results in regards to gender, skin type, nationality, and region. Chapter 7 redefines the conceptual task of gender classification, discusses how curation bias and sensor readings can impact training data, and assesses algorithmic performance of four gender-classifiers using measures of discrimination as defined in antidiscrimination legal literature. Chapter 8 closes the thesis with a discussion of future work and action steps for practitioners committed to full-spectrum inclusion.

2. Background

Boxes hold judgments, and labels have consequences.

2.1 Common Applications of Automated Facial Image Analysis

What is at stake?
Equipping machines with the ability to evaluate faces holds the promise of developing more empathetic human-machine interactions, monitoring health, and locating missing persons or dangerous criminals. But automated facial analysis also poses risks to civil liberties and privacy. Automated facial analysis technology can be deployed in a clandestine manner that limits public scrutiny due to the often invisible nature of the technology. In this chapter, I explicitly delineate common applications and implications of automated facial analysis to illuminate the pressing need for increased transparency in how these systems are deployed and how well they work on different groups. The history, implementation, and technical challenges of automated facial analysis algorithms are addressed in Chapter 3.

Automated facial image analysis entails the use of computer vision to evaluate images that contain faces to accomplish a range of perceptual tasks. Common facial image analysis tasks include face detection, face classification, face verification, and face identification. The domain of facial video analysis is beyond the scope of this thesis, though many of the tasks and limitations discussed here are applicable. For example, face detection done over a series of frames in a video stream can be used to accomplish the task of face tracking, but this thesis focuses on still images and not video.

2.2 Face Detection - Fundamental Task

Face detection is the fundamental task of automated facial image analysis. The task involves detecting the presence of one or more faces in an image. Face detection is a popular application of the computer vision task of object detection. By providing training data with many examples of faces, a face detection algorithm can learn to find locations in images containing human faces (Viola & Jones, 2001). Though face detection is an essential precursor to other tasks like face classification and face identification, it can also be applied in isolation. One of the most visible applications of face detection is on Facebook, where faces in uploaded images are surrounded with bounding boxes generated from a facial detection algorithm (see Figure 1). For image editing, face detection can enable automatic photo cropping that retains the majority of a face within the resulting photo (Suh et al., 2003; Yamamoto et al., 2016).

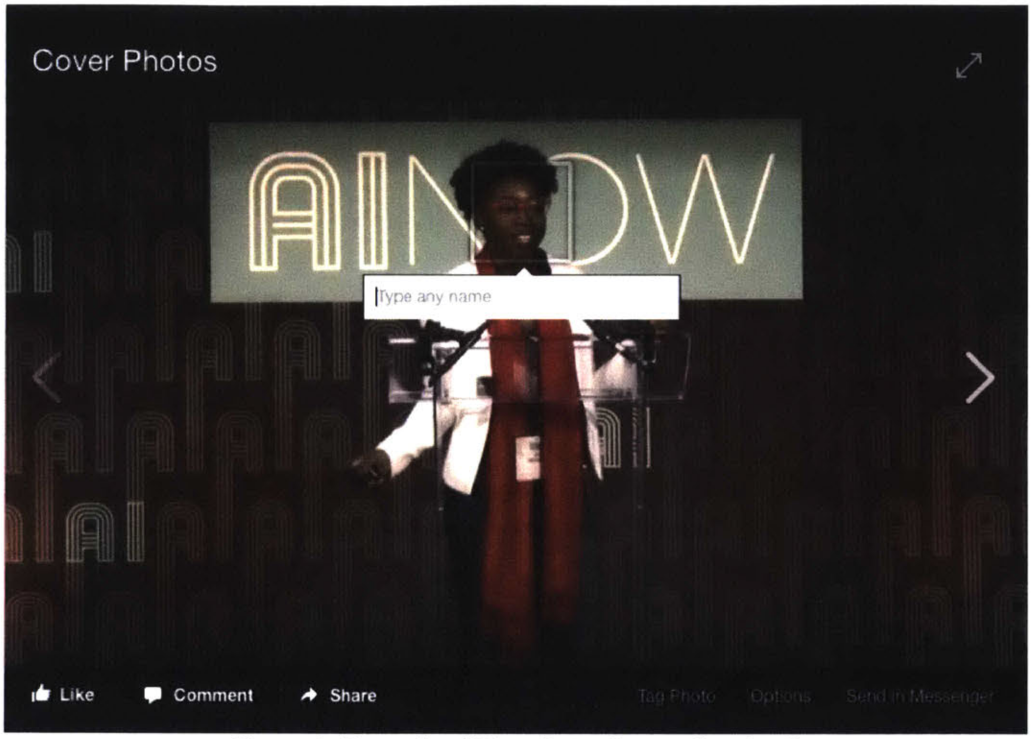


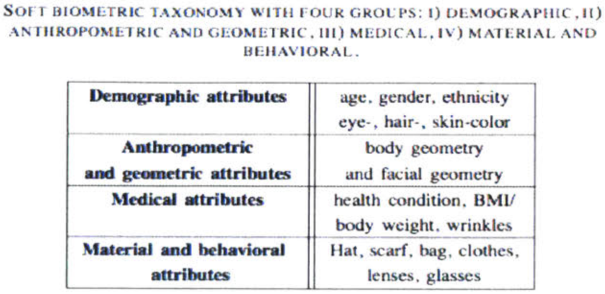
Figure 1. Example of Face Detection on Facebook

Implications of Face Detection for Algorithmic Performance Evaluation

Because face detection is the fundamental facial image analysis task, it can be a hidden source of bias. If widely adopted face detection methods are less effective for specific groups, tacit bias will impact dependent facial image analysis tasks. Demographic evaluations of the performance of these subsequent tasks must factor in the demographic performance of underlying face detection algorithms on which the tasks depends. In Chapter 3, I provide technical detail on how face detection can influence the accuracy of gender classifiers and performance on existing benchmarks.

2.3 Face Classification - Type of Face

Once a face is detected, additional analysis can be done to classify soft biometric information including demographic, anthropomorphic, medical, and material attributes (see Figure 2).



SOFT BIOMETRIC TAXONOMY WITH FOUR GROUPS: I) DEMOGRAPHIC, II) ANTHROPOMETRIC AND GEOMETRIC, III) MEDICAL, IV) MATERIAL AND BEHAVIORAL.

Demographic attributes	age, gender, ethnicity eye-, hair-, skin-color
Anthropometric and geometric attributes	body geometry and facial geometry
Medical attributes	health condition, BMI/ body weight, wrinkles
Material and behavioral attributes	Hat, scarf, bag, clothes, lenses, glasses

Figure 2. Soft Biometric Taxonomy

Like a face detection algorithm, a face classification algorithm can learn to apply class labels to faces by being trained on datasets that provide many examples of a class of interest. Gender classification can be learned by providing example data of male and female faces. Likewise, age classification and body type classification can be learned through many training examples that typify the classes of interest.

Unlike hard biometrics like a fingerprint, soft biometrics do not necessarily reveal an individual's unique identity on their own. Facial classification algorithms which assess soft biometric data like gender, age, and ethnicity are used for security, image-tagging, surveillance, age-specific access control, human-computer interaction, and marketing (Dantcheva, Elia, & Ross, 2016). Facial expression classification has found use in affective computing where the emotional states of individuals are the attributes of interest (Picard, 1997; Poria et al., 2017). Facial expression classification can be used to infer the Eckman emotions of joy, surprise, disgust, sadness, anger and fear (Hammal et al., 2007). These inferences can be applied to applications that factor in emotions for decision-making. In security, the emotion of a person of interest can be used to determine if an event is flagged as suspicious. Human-computer interaction (HCI) can be augmented using the perceived emotional state of a user (Abdat, Maaoui, & Pruski, 2011). One example of facial expression for HCI is the use of smile detection to trigger image capture on digital cameras (Shan, 2012).

Reductive, Unrepresentative Classifications

By definition, face classification algorithms label faces with categories. Demographic categories like gender, race, and ethnicity are linked to a wide range of overlapping physical attributes and are also historically and socially constructed. As a result, face classification algorithms focused on identity attributes categorize faces using simplifications of complex constructs. Many of these categories are reductive and contested. Though arguably fluid in construction, gender in particular is often reduced to a binary construct in the coded data structures for identity. Nevertheless, the reductive categories that are used for classification are mutable. They can be reviewed and expanded to become more inclusive. Chapter 7 provides an in-depth discussion on how gender classification in computer vision can be redefined to become less reductive.

2.4 Implications of Face Classification: Discrimination, Consent and Profiling

2.4.1 Class- based descrimination

Even when classifications are deemed accurate, their use can perpetuate discrimination and exclusion. Race or ethnic classification can be used by advertisers to exclude showing housing listings to a protected class[10] like African-Americans. Individuals classified as female based on their facial appearance may be subjected to higher prices as has been reported in instances where vendors use gender information to set prices. Classifiers that can assess body mass index (BMI) from faces can be used to infer body type for discriminatory purposes. For example, dating applications that include a prescreening could use automated body type classification to exclude individuals over a certain BMI treshhold. Classification algorithms that label accessories could use particular body adornment or coverings to infer information about ethnicity or religion. This data could also be used to unfairly exclude or stereotype individuals who exhibit culutral characteristics deemed unfavorable by the creators of a product or service. Assessing how facial classification is used is just as important as evaluating the accuracy of classifications.

2.4.2 Unattainable Consent

Regardless of the labels used for classification, a major question remains: "Can citizens opt out of being boxed in?"

Biometric data like identified fingerprints generally require cooperation to obtain. Soft biometric data that do not reveal a unique identity, but instead estimate a demographic attribute like age, and can be obtained without cooperation using automated facial image analysis. The 2016 AI Now report emphasizes the need for citizens to have the ability to know when they have been impacted by automated decision making and to have the ability to opt out (Crawford et al). The report focuses on automated decision making for high stake domains like employment, credit, and insurance decisions. Yet predictive modeling is not the only place automation is being used surreptitiously with large-scale social impact. The use of automated facial analysis similarly lacks consent and needs an "invest[ment] in research and technical prototyping that will ensure that basic rights and liberties are respected in contexts where AI systems are increasingly used to make important decisions" (p. 23). How can the general public be notified of mass surveillance that uses soft demographics, which does not on the surface invade privacy directly? Can individuals opt out? In theory, if the demographic classification data are not stored and the classification decision made by the system cannot be intercepted by an attacker, then

[10] Protected classes are social groups in the United States that have been histroically dicriminated against and have legal protections against discrimination.

ad hoc gender classification could pose minimum threat to privacy. Nonetheless, even if the collection of data is argued to have minimum privacy harms the decisions made based on protected attributes could still be used for illegal discrimination under the laws of the United States and the European Union.

2.4.3 Do soft biometrics protect privacy?

Even if gender, age, or race is used to discriminate against a collective group, one might argue that used alone these soft biometrics can preserve individual privacy. However, one data point rarely sits in insolation. Used in concert with biometric data, soft biometrics can be used to enhance surveillance making it easier to identify an individual by improving and expediting the searching process. In the case of gender-based biometric indexing, the gender attribute could significantly reduce the search space, the number of images that needed to be checked to look up an identity (Dey & Samanta, 2014). Due to the potential for soft biometrics to undermine privacy when used with other information, soft biometric data should be handled securely.

2.4.4 Stereotype Propagation and Profiling

Beyond creating classifiers for widely accepted demographic categories that can be ill defined, practitioners continue to develop new face classification algorithms that are arguably ill advised. Faception, an Israeli startup company uses computer vision to conduct personality profiling based on facial features. Their personality profile categories include terrorist, ace-poker player, and pedophile (Lubin, 2016). These categories risk placing people of certain ethnicities or facial geometries into stereotypical categories that do not reflect their actual behavior. While a person who is subjected to in-person profiling has a starting point to initiate a complaint, a person subjected to algorithmic profiling enabled by face classification is unable to affirm rights for fair and equal treatment or individual protections granted by law. Given this imbalance of power and anonymity, regulations that govern disclosure of surveillance should be expanded to include the use of automated collection of soft biometric data. Citizens can also be proactive in protecting their soft identities through use of adversarial tactics that obfuscate demographic markers. Wearing sunglasses and hats can provide some protection, but ultimately greater transparency in the use of facial classification technology will be needed to keep vendors accountable and the public informed.

2.5 Gender Classification

As shown in the previous section, face classification spans a variety of domains. For each application of face classification, vendors should explore the potential for exclusionary experiences and discriminatory practices. Since evaluation of the performance of select gender classification algorithms is a major focus of this thesis, I will now look at some domain-specific concerns of using automated gender classification.

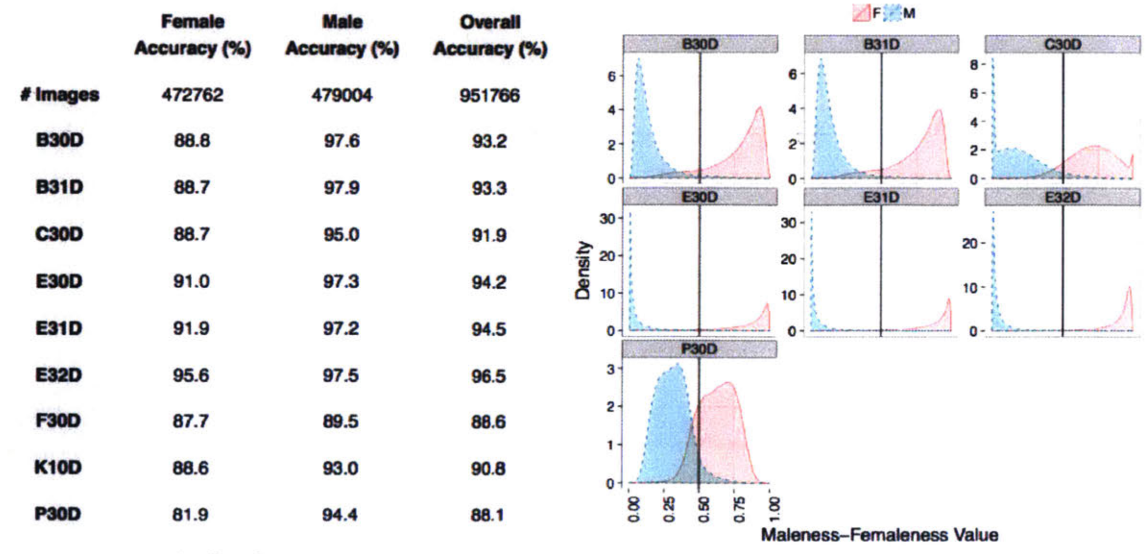
2.5.1 Reduced Privacy for Underrepresented Groups

Though a soft biometric like gender can arguably provide more privacy than a hard biometric like a fingerprint, the level of anonymity is based on the demographic distribution of the population under surveillance. In a male-dominated workplace environment that is under soft biometric surveillance, identifying an individual as female might be equivalent to revealing her individual identity. I choose a male-dominated workplace as a motivating example since the tech industry and research institutions that develop gender classification algorithms tend to be overwhelmingly male-identified. If there are only a few female-identifying workers who are classified as female by the system, their unique identities can be more readily deduced than their-male identifying counterparts. In addition, expanding the representations of gender used by these algorithms can introduce additional harms. If gender classification algorithms that attempt to categorize trans-identifying individuals are used, they can elevate the risk of outing an individual in a group who already faces discrimination and hate crimes (Stotzer, 2009). Gender classification algorithms used in tandem with other demographic classification algorithms can further reduce privacy and compromise security.

2.5.2 Gender Misclassification

Age & Ethnicity Interactions

Gender classification algorithms are fallible. The National Institute for Standards and Technology (NIST), a government institution tasked with benchmarking the accuracy of facial analysis algorithms evaluates voluntarily submitted gender classification algorithms. The latest gender classification report shows that algorithms NIST evaluated performed worse for female-labeled faces than on male-labeled faces (see Figure 3).



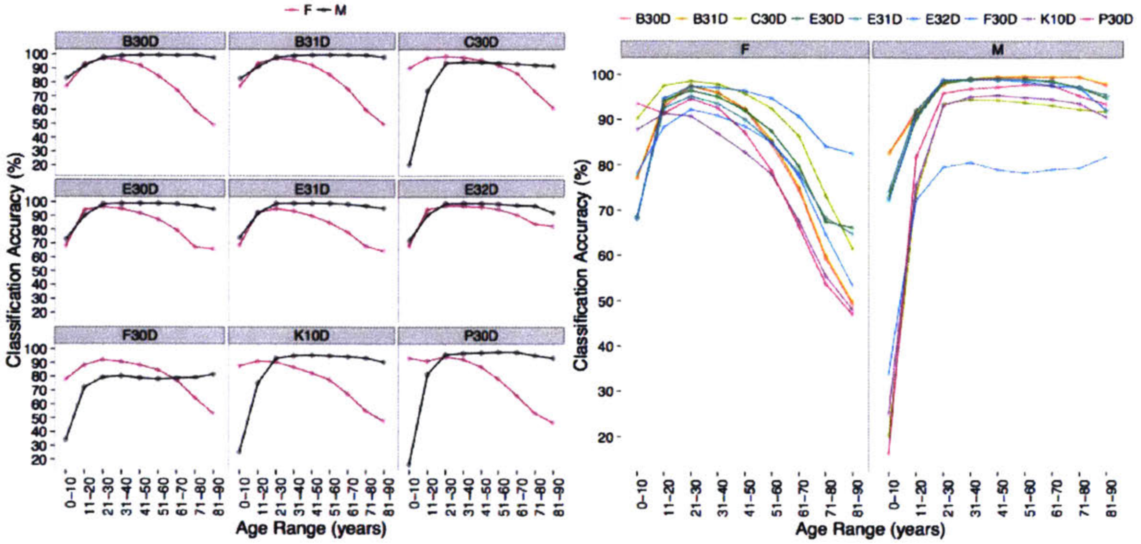
	Female Accuracy (%)	Male Accuracy (%)	Overall Accuracy (%)
# Images	472762	479004	951766
B30D	88.8	97.6	93.2
B31D	88.7	97.9	93.3
C30D	88.7	95.0	91.9
E30D	91.0	97.3	94.2
E31D	91.9	97.2	94.5
E32D	95.6	97.5	96.5
F30D	87.7	89.5	88.6
K10D	88.6	93.0	90.8
P30D	81.9	94.4	88.1

(a) Classification accuracy

(b) Distribution of maleness-femaleness value

Figure 3. NIST Gender Classification Evaluation Results

The likelihood of misclassifying a female-labeled face increased as the age of female subjects increased beyond thirty (see Figure 4) (Ngan & Grother, 2015).



(a) *By algorithm* (b) *By gender*

Figure 4. NIST Line Plots Showing Classification Accuracy Over Age Ranges

In addition to age, researchers have explored the impact of ethnicity on gender classification accuracy. However, study designs typically assess a limited set of ethnicities that hinders the generalizability of the results. Farinella and Dugelay claimed that ethnicity has no effect on gender classification, but they used a binary ethnic categorization scheme: Caucasian and non-Caucasian (2012). An experiment with a richer representation of ethnicity could show an interaction between gender and ethnicity as it relates to classification accuracy or might increase the validity of the claim that there is no interaction. In part, one goal of this thesis is to examine if intersectional effects occur.

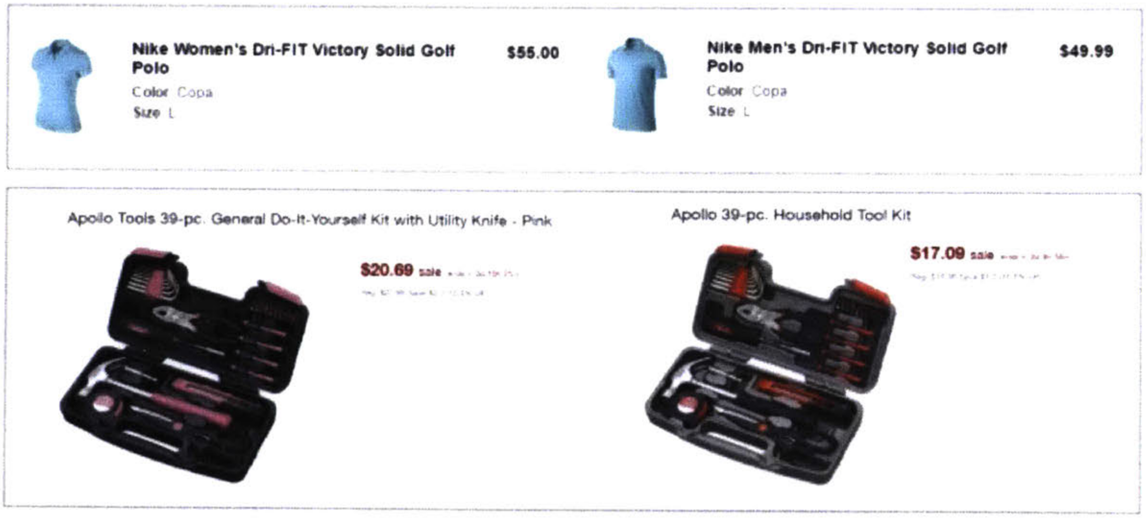
The NIST gender report explored the impact of ethnicity on gender through the use of an ethnic proxy. Acknowledging the challenges of defining ethnicity, NIST researchers used country of origin as an ethnic proxy to evaluate the ethnic performance of gender classification algorithms. None of the 10 locations (Argentina, Brazil, China, Colombia, Mexico City, India, Israel, Japan, Korea, Peru, Philippines, Poland, Russia, Taiwan) were in Africa or the Caribbean where there are significant Black populations. Even if it were argued that Brazil has a significant Black population, we must consider that the NIST evaluation used VISA images collected between 1996 and 2010. Due to the legacy of colonization, the portion of the population most likely to obtain a visa would be "*brasileiros brancos*", White Brazilians. To address the underrepresentation of people of African-descent in previous studies, this work explores gender classification on African faces to further scholarship on the impact of phenotype on gender classification.

False Negatives and Exclusion

Not only is more work needed on the ethnic dimensions of gender classification, but the impact of misclassification also needs to be further evaluated. Systems that are used for gender-specific access control could potentially deny access to individuals who should be allowed access based on the gender rules of the system. These denials could potentially have a higher impact on older women or any other group where there are systematically higher misclassification rates as compared to the overall population. Beyond potentially offending the misclassified, misclassification undermines the reliability of soft biometric systems that assess gender. In the case where gender is used to reduce the search space of a set of possible identities, gender misclassification can result in false negatives that may have been avoided if the correct gender labels were assigned.

2.5.3 Gender-Discrimination

Gender classification derived from facial image analysis introduces the possibility of gender-based discrimination that can be exploitive if permitted or outright illegal.
In December 2016, the Joint Economic Committee of the United States Congress released a report exploring the 'pink tax' - the charge that women pay more for similar products and services than men.



Nike Women's Dri-FIT Victory Solid Golf Polo **$55.00**
Color Copa
Size L

Nike Men's Dri-FIT Victory Solid Golf Polo **$49.99**
Color Copa
Size L

Apollo Tools 39-pc. General Do-It-Yourself Kit with Utility Knife - Pink
$20.69 sale

Apollo 39-pc. Household Tool Kit
$17.09 sale

Figure 5. Gendered Pricing Differences

According to a New York City Department of Consumer Affairs report exploring the price of 400 pairs of consumer goods. Goods differentiated for women had a 7% mark up on average (2015). Though differentiated prices based on gender-based presentation is not expressly illegal, the practice leaves women who already face a gender-pay gap further disadvantaged. The use of gender classification for interactive display ads can perpetuate the pink tax by using perceived gender to change the price of goods or to show higher priced goods. Online platforms like Amazon have received criticism for using demographic information like zip code to change available services and prices (Ingold & Soper, 2016). The demographic information derived from gender classification algorithms could be used in a similar manner. Furthermore, the price differentiation can become even more fine-grained and support the use of predatory marketing tactics. Some gender classification algorithms provide scores for the maleness or femaleness of a face. Instead of using gender to produce binary prices, the maleness or femaleness rating of a face can be used to further set the price of a product and associated marketing images and languages used with the product. For some products and service, federal regulations already exist to prevent discrimination. The Fair Housing Act Amendment of 1974 (42 U.S. Code 3601-3619) prohibits sex-based discrimination to renting and selling houses. Any advertising system employing gender classification needs to ensure methods for displaying ads for housing do not violate existing laws. Automated facial analysis as applied to advertising can make it easier for advertisers to mask intentional sex-based discrimination, which is in violation of the federal laws of the United States. Greater oversight is needed.

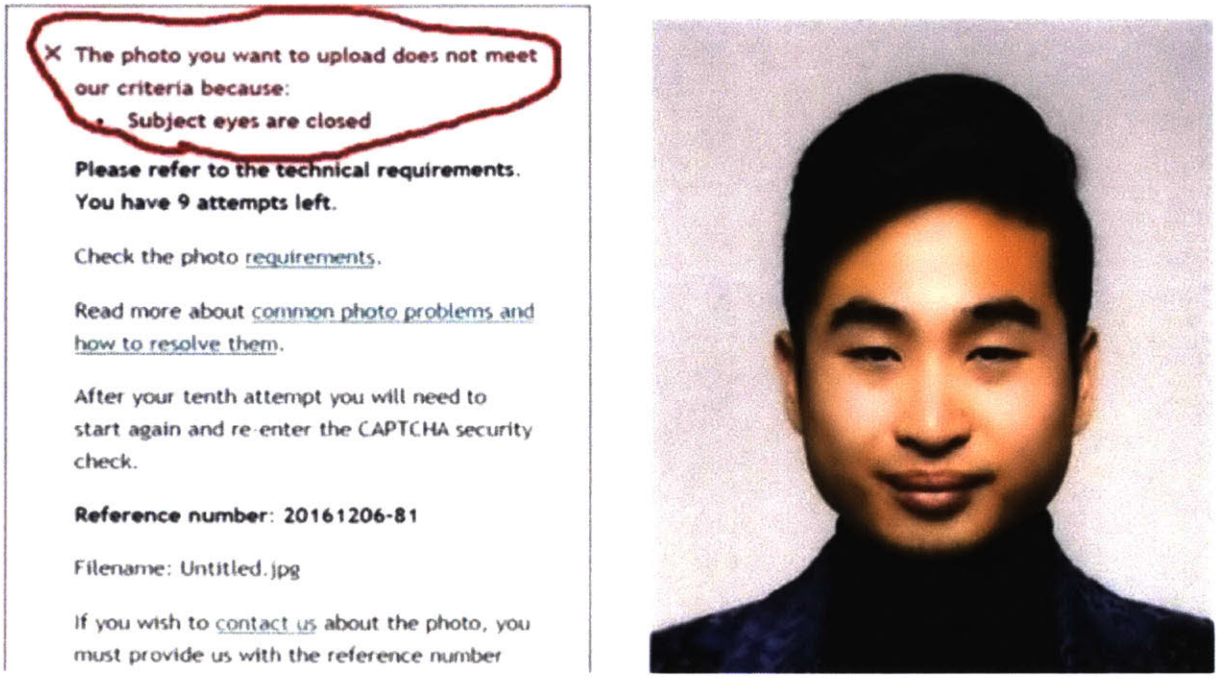
2.6 Facial Recognition (Verification and Identification)

This thesis is predominantly focused on gender classification, and in the previous sections I have raised concerns about the use and associated risks with the task of gender classification and the task of face detection that classification relies on. Here I turn my attention to facial recognition. While not the core focus of this thesis, facial recognition is important for two reasons. First, facial recognition is linked to high stakes domains like national-security and biometric authentication where accuracy is crucial. Unsurprisingly, facial recognition is a predominant research area in computer vision. Second, given that facial recognition and face classification performance have both been enhanced by advancements in the use of convolutional neural networks for deep learning, lessons learned from exploring facial recognition can inform future directions in face classification and vice versa.

Even though the term "facial recognition" or "face recognition" is often used colloquially to refer to a range of automated facial analysis tasks including face detection, biometric facial recognition is specifically focused on verifying or identifying a unique individual from an image. The task of face verification involves determining if two images contain the same face or not. Face verification can be used for security applications like biometric authentication used by banking institutions. For consumers, Samsung's Galaxy S8 and iPhone 8 come with the ability to unlock the phone through face verification. The task of face identification attempts to determine if a probe face is contained in an existing gallery of faces. For example, a police officer can take an image of a citizen and search it against a database of wanted criminals to see if there is a match. These algorithms attempt to return a set of faces from the gallery that most closely matches the probe image.

2.6.1 Enrollment Exclusion- Phenotype-Based Failures

To identify or verify a face, an existing record of the face of interest must be obtained and stored. The process of capturing a face to be used for a future verification or identification is called enrollment. The enrollment image, which provides the ground truth for the facial identity of an individual, is ideally captured in a controlled setting. The ISO provides a standard for front facing image capture including criteria such as open-eyes. While this criterion is seemingly innocuous, an infamous visa-enrollment incident shows the importance of checking not only for demographic bias that can lead to exclusionary experiences but phenotypic bias as well.



✕ The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

Please refer to the technical requirements.
You have 9 attempts left.

Check the photo requirements.

Read more about common photo problems and how to resolve them.

After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.

Reference number: **20161206-81**

Filename: Untitled.jpg

If you wish to contact us about the photo, you must provide us with the reference number

Figure 6. Phenotypic-Based Automated Facial Analysis Failure
Credit: Reuters/Richard Lee

As reported by Reuters, an automated passport system in New Zealand failed on a man of Asian descent deeming the subject's eyes to be closed though they were not. In this case, a demographic test may or may not have revealed the problem before deploying the system. A more appropriate test would forego coarse demographic demarcations and focus on specific phenotypic attributes of interest. A phenotypic test would specifically focus on a range of eye shapes to see if the automated checks put in place to obtain an ideal enrollment image could inadvertently exclude particular phenotypic groups. This border-patrol instance illuminates the importance of checking not only the accuracy of facial recognitions systems but also the enrollment processes these systems use to obtain ideal images. As automated facial analysis systems become increasingly used in society, phenotypic aware algorithms and phenotypic inclusive datasets should be further developed to minimize exclusionary experiences and offensive outcomes.

2.6.2 Differential Demographic Accuracy in Facial Recognition

There are growing concerns about the use of facial recognition technology by law enforcement. While it is true facial recognition can be used by law enforcement to combat identity fraud, identify missing children, and locate criminals, there are few regulations in place to safe-guard its use in the United States and protect privacy. Past research has shown that the accuracies of facial 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), yet there are no general accuracy standards and no demographic-specific standards in place for the procurement of facial recognition technology used by most law enforcement departments.

As new algorithmic methods are developed to improve facial recognition, ongoing accuracy checks that attend to demographic performance are needed. Garvie and colleagues provide an in-depth analysis of the unregulated police use of face recognition in 'The Perpetual Lineup'. Of the recommendations made in the report, the following support the creation of more rigorous standards for the use of automated facial analysis, racial accuracy testing, and processes for regularly informing the public about the use of facial recognition.

- **Recommendation 7. LEGISLATION Use of face recognition to track people on the basis of their race, ethnicity, religious or political views should be prohibited.**
- **Recommendation 8. LEGISLATION All law enforcement use of face recognition should be subject to public reporting requirements and internal audits**
- **Recommendation 9. LEGISLATION Congress should provide funding to increase the frequency and scope of accuracy tests and create more diverse photo datasets for training.**
- **Recommendation 14: FBI & DOJ The FBI should test its face recognition system for accuracy and racially biased error rates and make the results public.**
- **Recommendation 22: LAW Implement internal audits, tests for accuracy and racial bias, and the use of trained face examiners.**
- **Recommendation 26: NIST[11] Develop tests that closely mirror law enforcement workflows and issue best practices for accuracy testing**
- **Recommendation 27: NIST Develop and distribute diverse datasets of photos.**

A largely policy oriented report, 'The Perpetual Lineup' broadly proposes ongoing accuracy audits and more diverse datasets without explicitly delineating the scope of these audits or what constitutes diversity. What accuracy criteria should be put into place to deem if a facial recognition system is permissible for wide scale deployment on diverse populations? What constitutes diverse datasets? Building on recommendation 27, the Pilot Parliaments Benchmark developed in this thesis provides a concrete example of how a phenotypically diverse dataset of photos can be constructed. In addition to assessing demographic representation performance, this thesis argues for the inclusion of phenotypic performance evaluation. Recommendation 22 should include phenotypic bias and gender bias in addition to racial bias. This thesis seeks in part to expand NIST facial recognition standards and argues that intersectional benchmarking should be part of the best practices for accuracy testing and reporting that are recommended (#8, #9, #14, #26) in the 'Perpetual Lineup Report'.

[11] NIST – National Institute for Standards and Technology

2.7 Synthesis

This review of common applications of face detection, face classification, face verification, and face identification technologies highlights key limitations in regards to transparency, demographic performance, and potential abuses. The surreptitious way in which automated facial image analysis can be deployed requires developing methods to keep the public informed of the growing use of this technology. It also demands greater accountability from vendors who deploy these systems as they are positioned to exploit demographic data obtained from automated facial analysis for discriminatory practices like gender-based price gouging. Demographic performance for various facial analysis tasks differ and leave specific groups at higher risk for adverse effects that can result from differential accuracy. For example African-Americans in the US are subject to more interactions with law enforcement and also more likely to be misidentified by facial recognition technology. Automated facial analysis will become more pervasive and largely hidden from the public. Past research shows this technology performs worse on people of color than on Whites. In the case of gender classification, males are more accurately classified than females. Given the pending dangers, it is critical we better assess the performance of these algorithms. As will be further discussed in subsequent chapters, more work is needed to compose inclusive benchmarks that provide reality checks on the advances that have been made in the domain of automated facial analysis.

3. Related Work

The past dwells within our algorithms.

3.1 Overview

This thesis examines the evaluation of gender classification algorithms in relation to demographic and phenotypic differences. Of primary concern are influential automated facial analysis algorithms, means for evaluating algorithmic performance, the face datasets used to train and benchmark algorithms, and the establishment of fair accuracy standards. I review the evolution of facial analysis algorithms to highlight key breakthroughs and their implications for the task of gender classification. Finally, I present current practices for evaluating automated facial analysis performance along with the current limitations of existing approaches.

3.2 Breakthroughs in Automated Facial Analysis

Automated facial image analysis as described in *Chapter 2* describes a range of face perception tasks including, but not limited to, face detection, face classification, and face recognition. These tasks rely on machine learning, an approach to artificial intelligence that uses training data to enable an algorithm to learn a task instead of explicitly codifying rules. For example, instead of attempting to program rules to detect a face in an image, supervised learning algorithms accomplish the task by training on data that has images and bounding boxes around faces in the images used. The algorithm can then learn patterns associated with a human face without being explicitly programmed. Advances in machine learning techniques, increased computational capacity and greater availability of facial images as well as public and private investment have galvanized breakthroughs in automated facial image analysis. This chapter presents a brief survey of seminal and influential algorithmic developments along with their impact on automated gender classification. (Zafeiriou, Zhang, & Zhang, 2015), (Chihaoui et al., 2016) and (Reid et al., 2013) offer a more in-depth survey of advances in face detection, facial recognition, and gender classification respectively.

3.2.1 Face Detection

Face detection is the fundamental automated facial analysis task that enables other face perceptual tasks to be performed. Over the years, two dominant approaches have emerged: rigid-template and deformable parts-based model face detection. Rigid-template approaches include boosting algorithms and deep convolutional neural networks that will be further explained in this chapter. Deformable parts-based models use variations of general object detection methodology in computer vision to detect a face by relying on its composite components

Early work in face detection focused on finding faces in a constrained environment where head pose, scene illumination, and facial expression were relatively fixed and were commonly referred to as "constrained" (Zafeiriou et al., 2015). These early attempts performed poorly in real-world conditions where variations in pose, illumination, and expression along with the introduction of occlusions complicate the task. The major breakthrough that made unconstrained or "in-the-wild" face detection viable came from Viola and Jones in 2001. They made real-time face detection possible by combining the key ideas of the integral image, classifier learning using AdaBoost, and the attentional cascade structure. The integral image increased efficiency in face detection by enabling rapid constant time calculation of Haar-like features, which are derived using simple rectangular templates to find likely face patterns in an image (see Figure 7).

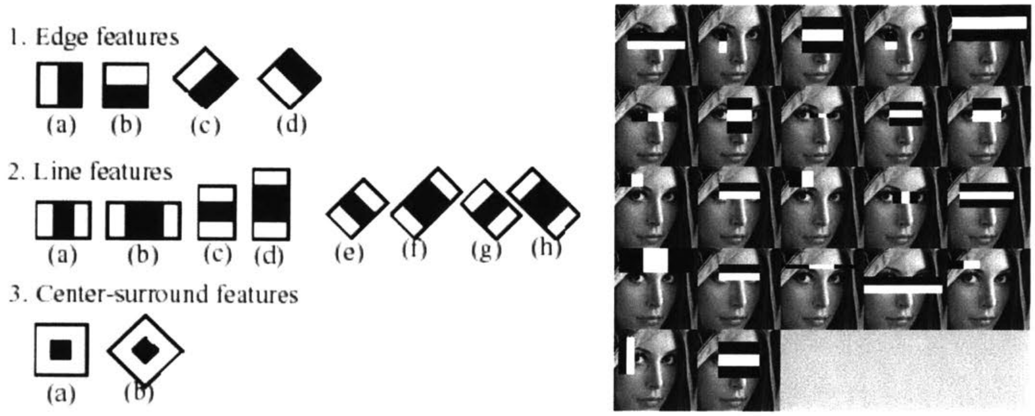


Figure 7. Examples of Haar-like Features

In this context, features are characteristics of an image. Features used for automated facial analysis include histograms of oriented gradients (HoGS), local binary patterns (LBPs), and scale-invariant feature transforms (SIFTs) that can be used to classify the contents of an image.

Choosing features and a suitable learning algorithm to classify an image based on those features are pivotal decisions for computer vision tasks. For face detection, Viola and Jones used the AdaBoost learning algorithm to both choose suitable Haar-like features and learn classification.

Earlier I introduced the term classification in the context of "face classification" to determine attributes about a face such as gender. Here I use classification as a machine learning term, which means to label a collection of features as an exemplar of a specific category or "class". In this case, the learning algorithm is used to classify regions in an image as either having a "face" or "no face" to perform face detection. To make the process of face detection even faster, Viola and Jones introduced the attentional cascade structure. This approach quickly rules out background regions that are unlikely to contain a face to optimize computation on regions in an image more likely to contain a face. Their ideas continue to influence derivative rigid-template face detection techniques like the popular Head Hunter (Mathias et al., 2014) detector. Derivative techniques explore combinations of alternative features, boosting algorithms, training methods, and regularization steps that mitigate overfitting.

3.2.2 Neural Networks

In recent years, increased parallel computational power and data abundance have led to renewed interests in neural networks (NN) for the field of artificial intelligence with applications in natural language processing as well as computer vision. Convolutional neural networks (CNNS) are a specific type of NN created for image-related tasks. Deep neural networks (DNNS) are NNs with many layers of connection. Each additional layer adds more depth. Deep Convolutional Neural Networks have achieved promising results for rigid-template face detection (Zhang & Zhang, 2014). Unlike the previously described methods, features are not explicitly selected. Instead an NN, artificial brain-like structure can be used to determine the presence, location, and pose of a face. This structure is composed of interconnected layers of perceptrons that behave like rudimentary neurons.

The DCNN is trained on a large dataset that is used to establish weights and hyperparameters that tune the DCNN to the task of face detection.

Moving past rigid-template face detection, deformable parts-based models offer another approach for face detection that has achieved state-of-the-art results on recent benchmarks. DPMs for face detection build on pictorial structures that represent an object of interest in a graph of related components and their connections to one another that define a given object. Figure 8 below shows a representation of a pictorial tree representation face.

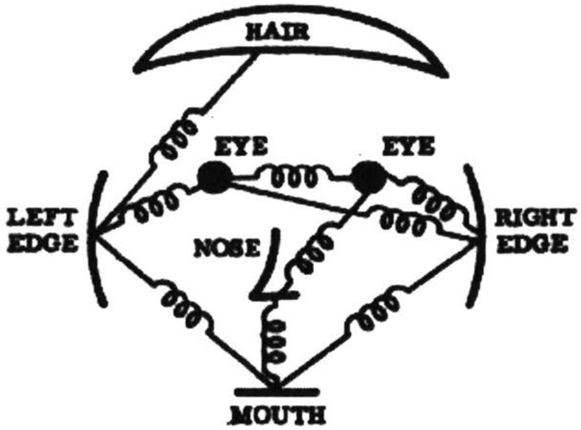


Figure 8. Pictorial Tree Representation of Face

Face detection is achieved by finding areas in an image that have a probability above a certain threshold of containing components of interests with connections that approximate the expected spatial/graph relationship that define a face. In general DPMs are computationally expensive given that the time required to perform feature extraction using HoGs, filter features, and score the correlations between features.

Even though there have been further advances optimizing object detection using DPMs (Yan et al., 2013, Yan et al., 2014), large scale attempts to collect face image datasets still rely on rigid-template approaches that can be deployed quickly. 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).

As face detection becomes commodified, any demographic or phenotypic limitations present in widely adopted pre-trained face detection algorithms can perpetuate bias - i.e., any biases Head Hunter is subject to may be perpetuated through dissemination of Megaface, and future algorithms tuned on Megaface will likely inherit these biases. Benchmark and training data that are collected using commodity algorithms with tacit bias can limit the diversity and difficulty of newly collected datasets. The intersectional benchmarking and curation schemes proposed in this thesis can be applied to face detection to make measurements of accuracy more robust across a range of facial geometries and skin types.

3.2.3 Face Recognition

Early work in automated facial recognition began in the 1970s. Goldstein, Harmon, and Lesk employed 22 manually derived, subjective markers including hair color and lip size to achieve automatic facial recognition (1971). By the late 1980s Kirby and Sirovich overcame manual coding of specific features by using the linear algebra technique of principal components analysis (PCA) (1987). This work undergirded the rise of appearance-based approaches for facial recognition. Appearance-based approaches, also called global approaches, treat face images globally instead of focusing on specific facial regions like the mouth or eyes. The face is represented by a matrix of pixels. For efficiency, dimension reduction is used to project the discriminative parts of a face into a lower dimensional sub-space also known as a face space. In 1991, Turk and Pentland famously extended the use of PCA for facial recognition with the introduction of eigenfaces. With eigenfaces, linear combinations of the eigenvectors derived from the covariance matrix of sample faces are used to reconstruct individual faces. The facial recognition task of identification is achieved by finding a face in a gallery of images that has the most similar linear combination of eigenvectors as the probe image.

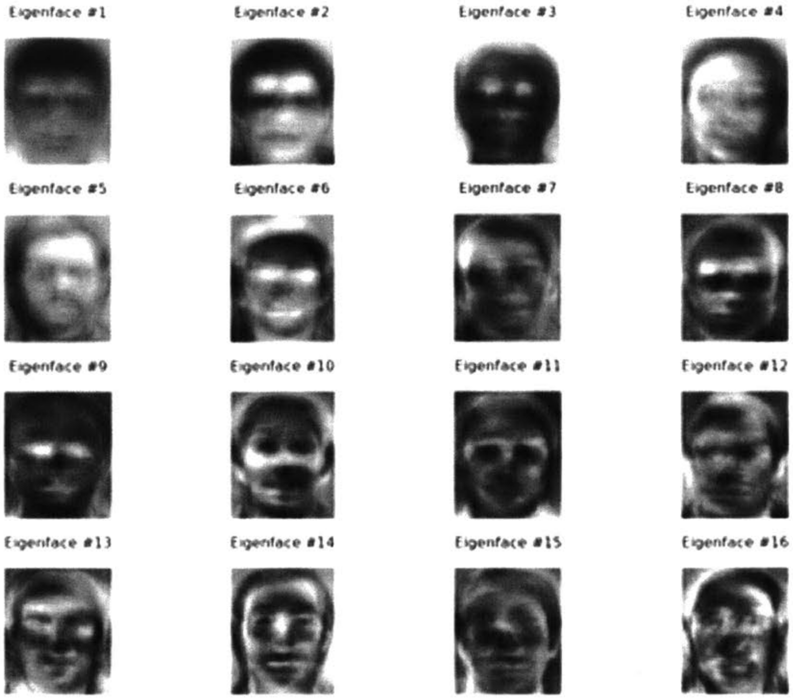


Figure 9. Eigenfaces

Linear Discriminant Analysis (LDA) for facial recognition also known as Fisher Linear Discriminant Analysis is another well known appearance-based approach for facial recognition (Belhumeur, Hespanha, & Kriegman, 1997). The "Fisher faces" method creates a subspace of a face image that optimally distinguishes the face of different people. Independent Component Analysis (ICA), Gabor Wavelets, and their derivatives have all been employed for linear appearance-based face recognition. The "kernel trick" has been used in combination with the previously mentioned techniques for non-linear appearance-based face recognition along with support vector machines (SVMs) (Chihaoui et al., 016).

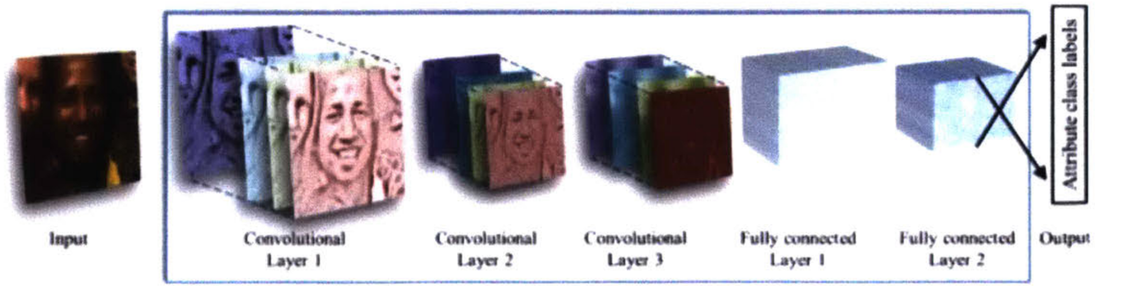
Feature-based approaches, sometimes referred to as local approaches, for facial recognition are also used. Here specific features of the face are characterized based on well-defined statistics. Interest-point-based methods like Dynamic Link Architecture (DLA) and its derivative Elastic Bunch Graph Matching (EBGM) can be effective in cases where only a single reference image is available, but they are limited by the effectiveness of face localization algorithms that find the position of facial landmarks like the center of the pupils. As with the case of face detection algorithms, feature-point localization algorithms that can encode tacit phenotypic bias based on original training data could limit demographic performance of subsequent algorithms that build upon their results. Local appearance-based approaches use region localization and then proceed to determine the best representation for each region by using characteristics such as Gabor Coefficients, Haar wavelets, and Local Binary Patterns. Each region can be classified by using the most appropriate method. To perform recognition, graph matching or score fusion techniques are used. In the former the spatial relationships of regions are represented by the edges between regions that represent nodes in a face graph. In the latter, separate classifiers calculate a score for each local characteristic, which is combined into a global score to determine a match.

Deep learning, machine learning that utilizes many layers of abstractions to map input data to desired output results, outperforms the local and global approaches described previously. As with face detection, performance in facial recognition has increased with the resurgence of neural network based machine learning pioneered by Geoffrey Hinton. Hinton integrated the backpropagation learning algorithm for training multilayer neural nets in the 1970s to lay the foundation for the deep learning techniques to come. In 2014, Facebook researchers published DeepFace, which demonstrated facial recognition could benefit from the promise of deep learning. Through the application of deep learning on 4.4 million labeled images of 4030 unique subjects, DeepFace achieved 97.35% accuracy on the Labeled Faces in the Wild (LFW) face recognition benchmark database that was considered one of the most challenging at the time (Taigman et al., 2014). These results demonstrated an impressive 27% improvement over the previous state-of-the-art (Sun, Yang, & Tang, 2014). Google researchers followed with another convolutional neural net (CNN) trained on over 100 million images of roughly 8,000 subjects and achieved 99.63% Accuracy on LFW (Schroff, Kalenichenko, & Philbin, 2015). These promising results led to growing research exploring the application of deep learning to face recognition (Parkhi, Vedaldi, & Zisserman, 2015). Nonetheless, improved performance on LFW should be celebrated cautiously. As will be further discussed in Section 3.3.4, the LFW benchmark suffers from significant gender and ethnic imbalances.

3.2.4 Gender Classification & Convolutional Neural Networks

Given the success of convolutional neural networks in computer vision tasks applicable to automated facial analysis such as object recognition, image classification, and pose estimation (Gu, 2017), CNNs are also being used for face gender classification. Introduced in 1991 SEXNET, one of the first gender classification approaches, relied on a three-layer fully connected neural network (Golomb, Lawrence, Sejnowski). For efficiency, convolutional neural network architectures eschew fully connected layers throughout all levels and instead use strategic subsets of artificial neurons to perform convolutions between layers in a network. More recently, influenced by the gains made in facial recognition made with CNNs, Israeli researchers improved the state-of-art on gender classification by using a standard CNN structure and introduced a new gender and age benchmark dataset called Adience

Network Architecture



All 3 RGB channels First, resized to 256 x 256, then cropped to 227 x 227

96 filters size 3x7x7

256 filters size 96x5x5

384 filters size 256x3x3

Each convolutional layer is followed by rectified linear operator (ReLU), max pooling layer of 3x3 regions with 2-pixel strides and a local normalization layer

Both fully connected layers contain 512 neurons followed by ReLU and dropout layer

Output to class labels (age / gender)

Figure 10. CNN Architecture for Face Classification

The Adience benchmark was introduced because less work has been done to benchmark performance on gender classification than on facial recognition (Levi & Hassner, 2015). However 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 (see section 2.5.2 for performance results).

3.3 Evaluation in Automated Facial Analysis

3.3.1 Overview

Mechanisms for tracking and improving performance on face perception tasks play a critical role in moving the state-of-the art for automated facial analysis. To galvanize research and development activity, facial analysis challenges are overseen by government agencies, conference organizations, and research institutes. Peer recognition, published rankings, and at times monetary incentives motivate participation. Standard benchmarks in the form of datasets provide a comparable means of assessing competing algorithms and evaluating the current state-of-the-art for specific tasks. These benchmarks are crucial in improving facial analysis algorithms because they set uniform standards by which researchers and companies publicly prove the effectiveness of their algorithms. Once a benchmark is saturated, that is to say numerous methods are developed that achieve perfect or near perfect performance accuracy, more challenging benchmarks are composed. Performance on a benchmark dataset is measured by using standard machine classification and accuracy metrics that are generally reported in aggregate. In addition to benchmark datasets, training datasets of face images are also made available to researchers.

This thesis makes the case for reconfiguring the way in which benchmark and training datasets are composed and how benchmark metrics are reported so that demographic and phenotypic performance become part of standard evaluation procedures. This segment will conclude with a look at seminal research demonstrating the utility of demographic-specific evaluation when investigating algorithmic performance.

3.3.2 NIST Benchmarks and Challenges

The National Institute of Standards and Technology (NIST), a United States government agency tasked with promoting innovation and advancing national competitiveness through advancing standards, continues to release a series of projects focused on improving the state of automated facial analysis. In 1993, the Department of Defense (DoD) Counterdrug Technology Development Program Office introduced the five-year Face Recognition Technology (FERET) Program. FERET focused on sponsoring research to creating the FERET dataset consisting of 14,126 facial images of 1199 individuals captured in constrained environments and to evaluate performance on that dataset (Phillips et al., 1996; Phillips et al., 2000). FERET was instrumental in moving facial recognition from research into practice and setting precedents for biometric evaluation that has influenced subsequent evaluation programs including the UK Biometrics group. As with other datasets, the FERET dataset was used not just for face recognition but also for comparing research results on face detection (Zafeiriou et al., 2016) and face gender classification (Levi & Hassner, 2015). NIST maintains the FERET database that has been distributed to over 100 external entities. After FERET, NIST began a series of Face Recognition Vendor Tests (FVRT) (2000, 2002, 2006, 2013) with each interaction introducing improved test procedures such as sequestering test images to mitigate overfitting to the benchmark.

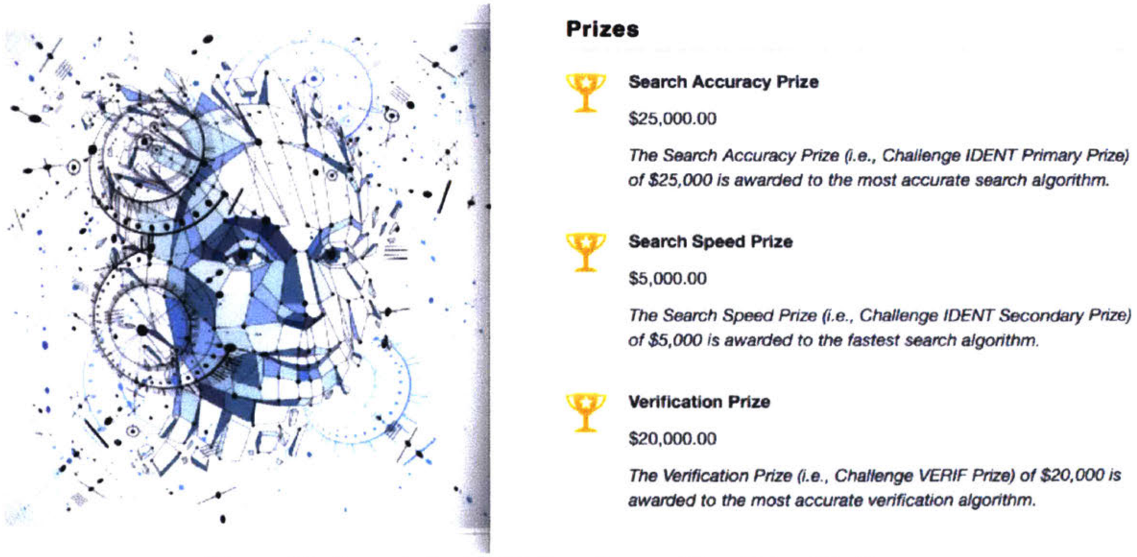
In February of 2017 in accordance with the recommendation from 'The Perpetual Lineup' (Garvie et al., 2016) report to provide continuous monitoring of facial recognition algorithms, NIST announced initiative FVRT Ongoing (FRVT-O). The first evaluation available focuses on the task of facial verification. Beyond verification, there are planned ongoing evaluations for one-to-many identification accuracy - NISTIR 8009,[12] face detection accuracy (Cheney et. al., 2015), age estimation - NISTIR 7995,[13] and gender estimation - NISTIR 8052.[14] To motivate participation in FVRT-O, NIST in partnership with the Intelligence Advanced Research Project Activity (IARPA) is offering a total prize purse of $50,000 in its first ever Facial Recognition Prize Challenge.[15] Monetary incentives along with the public prestige awarded to creators of inclusive algorithms may prove effective in driving work to mitigate demographic and phenotypic bias.

[12] https://www.nist.gov/node/558561

[13] http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=915238

[14] http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8052.pdf

[15] https://www.challenge.gov/challenge/face-recognition-prize-challenge/



Prizes

Search Accuracy Prize

$25,000.00

The Search Accuracy Prize (i.e., Challenge IDENT Primary Prize) of $25,000 is awarded to the most accurate search algorithm.

Search Speed Prize

$5,000.00

The Search Speed Prize (i.e., Challenge IDENT Secondary Prize) of $5,000 is awarded to the fastest search algorithm.

Verification Prize

$20,000.00

The Verification Prize (i.e., Challenge VERIF Prize) of $20,000 is awarded to the most accurate verification algorithm.

Figure 11. Awards for NIST Face Recognition Prize Challenge

3.3.3 Influential Benchmarks

Beyond NIST projects, other initiatives have produced influential benchmarks. In addition to the FERET dataset, XM2VTS, PIE, and FRGC were constrained face datasets used to train or test face detection. The PASCAL Visual Object Classes benchmarks and challenges (Everingham et al., 2014) along with Face Detection Dataset and Benchmark (FDDB) introduced unconstrained images to advance face detection. For facial recognition, FERET has been used along with the widely adopted Labeled Faces in the Wild (LFW) and derivative datasets. Demonstrating its prolific impact FERET, was also used for benchmarking gender estimation algorithms, though now there are datasets like Adience that provide more challenging faces for benchmarking gender estimation.

3.3.4 Dataset Challenges

Labeled face datasets have spurred the development of automated facial analysis that employs supervised learning techniques. Still, creating a labeled face dataset is often resource intensive, and current curation methods can be susceptible to bias. The process of labeled dataset creation includes defining labels, collecting face images, reliably applying labels to face images, and making the data available to the benefit of the research community. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) provides a comprehensive overview of the steps and technical challenges involved in collecting large image datasets in general (Russakovsky & Deng, 2015).

When defining demographic labels for face datasets, coarse categorization can make it difficult or impossible to assess subgroup performance due to omission or multi-group aggregation. For example, ethnic demographic labels for early face datasets often included categories for White or Asian. Faces falling into alternative groups were categorized as non-White or other due to underrepresentation in training sets. As a result, the training set is less useful for training algorithms to distinguish people of African ancestry, Indian ancestry, or other omitted groups. Next, even though collecting face images has become easier because of the large number of images available online, the images collected can have significant demographic skews. In the past, celebrity photos have been used to create face datasets due to image availability and the ability to verify identity (CelebA, FaceScrub, PubFig, IMDB-WIKI). Yet, celebrity demographics are not representative of population demographics. Furthermore, when face detection algorithms are used to scrape images online, any tacit bias present in the face detector can skew the images that are collected.

Applying demographic labels to faces can also prove challenging as social constructs like race are not well defined. When manual annotation is employed, cross-race effects may influence the effectiveness of labeling faces that are not like the face of the annotator. Though less apparent, dataset accessibility can also introduce bias. Large companies have resources to develop proprietary datasets that may have more diverse faces than the datasets available to researchers. Researchers then default to open datasets, some of which have been documented to contain significant demographic bias. LFW which has been the defacto benchmark dataset for face recognition was estimated to be 77.5% (10258/13233) male and 83.5% (11045/13233) White (Han & Jain, 2014) The dataset is composed of celebrities, and images were scraped using the OpenCV implementation of the Viola Jones face detector. In other words, it is a deeply non-representative set that is widely used primarily due to its openness.

In response to these limitations IARPA, released the IJB-A dataset as the most geographically diverse set of collected faces, and no face detector was used to select images to limit bias (Klare et al., 2015). Given the importance of benchmark datasets and the influence of NIST that maintains the IJB-A dataset, this thesis will examine the current composition of the dataset to make recommendations about how to make the benchmark and annotations more inclusive.

3.3.5 Benchmark Metrics

The need to standardize measures of accuracy and reporting requirements drove the development of challenges and benchmarks like PASCAL VOC, LFW, and FRVT. For all automated facial analysis tasks computational efficiency is of concern, thus metrics related to runtime and memory use are reported. Specific automated facial analysis tasks have associated accuracy metrics. In this thesis I adopt the approach used in the NIST Gender FVRT (NISTIR 8052) for reporting gender classification, which defines metrics for male accuracy, female accuracy, and overall accuracy as follows:

Let M and F represent the total number of male and female images. Let TM and TF represent the true males and true females classified by the algorithm.

Female Accuracy $= \frac{TF}{F}$

Male Accuracy = $\frac{TM}{M}$

Overall Accuracy = $\frac{TF + TM}{F + M}$

The NIST Gender FVRT report also explores the impact of gender and ethnicity on the classification accuracy. I follow a similar reporting pattern but with attention to the impact of skin type phenotype on classification accuracy. Disaggregating performance results can provide new insights that lead to better algorithms. In the seminal paper "Face Recognition Performance: Role of Demographic Information", FBI Expert Brendan Klare examined for the first time the impact of race, gender, and age on facial recognition performance. The study showed that the leading recognition algorithms of the time performed uniformly worse on faces labeled as female, youth (18 -30), and Black. However, for trainable algorithms, researchers were able to improve performance on these groups by training those algorithms on a dataset with uniform representation. The report did not look at the intersection of race and gender on facial recognition but called for future work in the area (2012). Now with the advent of deep learning and increased use of automated facial analysis, it is even more imperative that we examine the role of demographics in this domain. In addition, the role of phenotype on classification accuracy that may not fall along traditional demographic demarcations needs to be examined.

4. Fairness, Accountability, and Transparency in AI

The coded gaze reflects both our aspirations and our limitations.

4.1 Overview

Breakthroughs in artificial intelligence have increased the adoption of automated facial analysis technology for authentication, face-based searches, and a growing number of widely used platforms like Facebook and Snapchat. In general, artificial intelligence is increasingly employed for high stakes decision making in areas such as insurance, credit lending, employment and even criminal sentencing. As automation and the adoption of artificial intelligence rises, there is a growing consensus on the need to develop fair algorithms and to establish processes for accountability and transparency in artificial intelligence. In this chapter, I situate the establishment of subgroup accuracy standards in ongoing debates around fairness, accountability, and transparency in artificial intelligence. I end, by exploring measures for discrimination and fairness as defined by legal scholars and how they can be applied for intersectional benchmarking (See Section 1.3).

4.2 Challenges with Defining Fairness

Defining fairness presents both ethical and technical challenges. Power inequities and a lack of diversity among researchers, data scientists, artificial intelligence practitioners, and policy makers also means those who are at risk for the adverse impacts of artificial intelligence currently have little voice and influence in shaping future technology. Algorithmic fairness debates have been galvanized as of late by the 2016 ProPublica investigation of recidivism risk scores generated by Northpointe. Northpointe developed the COMPAS Risk Assessment using machine learning to predict the likelihood that an individual will reoffend. The risk score can then be used by judges to inform sentence length. Presumably, individuals with high risk scores are given longer sentences than people with low risk scores facing similar criminal charges.

ProPublica journalist Julia Angwin investigated the accuracy of the risk scores in regards to race. Her team's analysis revealed that false positives rates for Black individuals were nearly twice as high as false positive rates for White individuals. Black individuals with a low recidivism risk were twice as likely to be misclassified as high risk than White individuals with low risk. White individuals with a high recidivism risk were nearly twice as likely to be classified as low risk than Black individuals (48 percent vs. 28 percent) (2016). Thus, Angwin concluded the COMPAS Risk Assessment is racially biased. From an ethical perspective given similar charges, it is unfair for Black individuals to face longer sentences and White individuals to face shorter sentences because of algorithmic bias. Northpointe countered Angwin's claim of bias citing that the positive predictive value was equal between Black and White individuals. The overall accuracy rate was evaluated to be 70% with minimal deviation for both Black and White individuals.

Though outside observers may question the suitability of using a risk assessment with 70% accuracy for positive prediction, the company deemed its product fair because the likelihood that someone who is likely to be arrested again is arrested is the same between the two races (Dieterich, Mendoza, & Brennan, 2016).

This debate reveals how competing perspectives on fairness prevent a singular definition of equality. Because false positive and false negative rates differ between Whites and Blacks in Angwin's analysis, the assessment is deemed unfair. From Northpointe's analysis, because the positive predicative value (people who are predicted to rescind who in fact rescind) is roughly equal between the two groups, they deem the assessment fair. By focusing on different metrics to define fairness, Angwin and Northepointe reach two different conclusions even though they analyze the same data. The COMPASS debate typifies the difficulty of defining fairness.

Furthermore, fairness definitions have technical implications. In machine learning classification, the positive predictive value is captured in a metric known as precision, a measure of the proportion of cases deemed positive examples that are true positives. Recall, also called sensitivity, is a metric that measures how many of the true positive examples were correctly identified. When a learning algorithm is trained, a cost function is defined that can optimize for precision or recall. Selecting which errors to minimize is based on the context in which an algorithm will be used. A false negative for a cancer diagnostic is arguably more costly for the individual than a false positive. Ideally, we would want to minimize all errors, yet as Kleinberg et al. (2017) prove, balancing false negative rates and false positives rates while maintaining high predictive accuracy is impossible when population priors differ. Trade-offs must be made based on subjective judgments about the perceived impact of false positive predictions versus false negative predictions.

4.3 Worldviews on Fairness

Frielder et al. (2016) define the study of algorithmic fairness as an exploration of two key transformations between three spaces. The first space is the construct space that represents an abstract concept or unobservable trait like creditworthiness or intellectual capacity. The space cannot be directly known so it is approximated by the observation space with proxy attributes related to a construct. The attributes are finally mapped to a decision space using a learning algorithm that is optimized to make a prediction related to the construct. Judgments about algorithmic fairness are influenced by worldviews regarding construct and observation spaces that are not always explicitly articulated.

Frielder et al. also highlight that definitions of algorithmic fairness have underlying worldviews. Dwork et al. (2012) make the case that with fair algorithms 'similar people should be treated similarly. Here what Friedler et al. call a "What you see is what you get" (WYSIWYG) assumption is made. A WYSIWYG worldview assumes the inputs used in the observation space can be taken at face value as good proxies for the construct of interest. For example, it may be assumed that standardized test scores are a good reflection of a student's intellectual capacity. If this is the case, using standardized tests to distribute opportunities can be determined using a WYSIWYG worldview. However, when it is suspected that attributes in the observation space reflect structural bias in society, an alternative worldview "We are all equal" (WAE) may be used. This worldview assumes that differences in observable attributes can be more a reflection of historical and social factors than an individual's inherent attributes. Thus members of groups that have been disadvantaged should still have access to opportunity. Given prior research showing standardized tests have higher socioeconomic class correlations than grade correlations (Zwick & Green, 2007), a WAE worldview would question the suitability of standardized test scores as a definitive measure of intellectual capability.

In addition, underlying worldviews made when creating predictive models have mathematical implications for individual fairness and group fairness. Techniques used to optimize for individual fairness rely on WYSIWYG worldview and break down when being used in a domain where structural bias has limited opportunities in the past. Conversely, techniques used to optimize for a WAE worldview that takes structural bias into account, do not function well when working on a construct space that can be viewed as representative of an inherent individual trait or objective performance (Friedler et al., 2016).

4.4 Accountability and Transparancy

Setting aside questions about ideal ways to create fair algorithms and more representative datasets, we must still ask fundamental normative questions. If we know an algorithm performs "poorly" on a specific demographic group, should we use that algorithm to make high-stake decisions about that demographic group? Further still, **if an unfair or harmful decision is made by an automated system, the question of accountability remains. Deciding who is accountable for an algorithmic decision is complicated by the distributed nature in which algorithms are combined and developed**. At times inherited code cannot be easily changed. In reviewing breakthroughs in automated facial analysis, we have seen the interconnected manner in which artificial intelligence evolves. Thus, if a company uses off-the-shelf commodity face detection software based on a series of algorithms refined overtime by academics and trained on data sourced from the Internet, what entities should be held responsible for adverse impacts? Arguably immediate responsibility accrues to whomever ultimately decides to incorporate the algorithm into a product. However, if a commodity face detector is made part of a software library reported to have passed a prolific benchmark with over 95% accuracy, and the developer incorporating the library is unable to retrain the face detector, should the developer, product manager, or company be held responsible if the face detector is found to systematically work poorly on female faces?

Furthermore, **demanding transparency with either algorithms or datasets is hindered by the economic imperative for companies to limit access to proprietary information**. Even when algorithms are not Black boxed, the underlying model of a neural network used for predictive modeling may be uninterpretable. Even if we have accuracy metrics from benchmark or validation data, precisely how the model makes a decision cannot be explicitly explained. There is growing interest in the realm of explainable AI, particularly as regulations increase (DARPA-BAA-16-15).[16] In 2016, the European Union passed a new General Data Protection Regulation that will take effect in 2018. The legislation restricts automated decision-making that significantly impacts individuals and calls for a right to explanation. Individuals will have the right to request information on how a high stakes automated decision was made (Goodman & Flaxman, 2016). Existing machine learning techniques like the convolutional neural networks used for automated facial analysis will need to become more interpretable or defensible to justify their use. In 2017, United States House Committee on Oversight and Government Reform held a Hearing on Law Enforcement's Use of Facial Recognition Technology. In a written testimony Senior Staff Attorney Jennifer Lynch states:

> Now, law enforcement officers can use mobile devices to capture face recognition ready photographs of people they stop on the street; surveillance cameras boast real-time face scanning and identification capabilities; and the FBI has access to hundreds of millions of face recognition images of law-abiding Americans. However the adoption of face recognition technologies like these has occurred without meaningful oversight, without proper accuracy testing of the systems as they are actually used in the field, and without the enactment of legal protections to prevent their misuse. (p.2)

The time is now to increase transparency by creating more rigorous standards for the accuracy and applicable use cases of automated facial analysis.

4.5 Measures for Discrimination

Ongoing monitoring of the performance of automated facial analysis can improve transparency. In addition to creating processes for more rigorous accuracy testing, determining acceptable accuracy rates for different demographic and phenotypic groups must be addressed. As acknowledged in the prior section, defining fairness is not simple. Even when positive outcomes are equal between groups, negative outcomes may disproportionately impact a specific group. In attempting to optimize for fairness, the utility or predictive power of an algorithm can at times be compromised (Dwork et al., 2012). When should we deem an algorithm discriminatory? What factors should be considered? This section explores existing definitions for discrimination that can frame how we interpret the fairness of measures of accuracy between different groups.

4.5.1 Disparate Treatment versus Disparate Impact

Legal literature provides a rich jurisprudence history about defining discrimination. Legally discrimination falls under two main categories: disparate treatment and disparate impact. Disparate treatment occurs when intentional acts of discrimination are

[16] DARPA Broad Agency Announcement for Explainable Artificial Intelligance https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

perpetrated. For example, a landlord may refuse to offer housing to people of a specific ethnicity. Disparate treatment assumes intent to discriminate, which can be difficult to establish in a court of law when there is plausible deniability. Proving disparate treatment in the use of automated facial analysis based on deep learning is difficult. Learning algorithms are not explicitly programmed to produce an outcome. They arrive at decisions through a learning process. Disparate impact focuses on the outcomes of processes that can lead to discrimination or unequal opportunity regardless of intention. The process can appear to be neutral and done in good faith, but if the outcome disparately impacts a protected group then there can be a legal case for discrimination. In general, protected groups are people who have been historically denied opportunities like women and African Americans. The use of biased facial recognition technology can result in a disparate proportion of individuals from a protected group being subjected to unwarranted interactions due to misclassification or false identification. However, legally proving disparate impact from algorithmic classification is challenging. Even with the risks for disparate impact posed by artificial intelligence and large stores of data, current legal frameworks lag behind rapidly evolving technology (Barocas & Selbst, 2014). Regardless of the current legal landscape, creating technology that works well for subgroups and minimizes disparate impact should be a goal for all vendors and researchers who aim to develop generalizable facial analysis algorithms.

4.5.2 4/5ths Rule and Disparate Impact Risk

What constitutes disparate impact in algorithmic classification? How might we measure differences in outcomes? Statistical analysis provides a number of measures that have been adopted to assess discrimination between groups including risk difference, risk ratio, and the odds ratio also referred to as the selection rate. Given p1 is the probability that group one does not receive a benefit and p2 is the probability that group two does not receive a benefit, the measures for discrimination are defined as follows:

Risk Difference: $(p1\ - p2)$

Risk Ratio: $\frac{p1}{p2}$

Odds Ratio/Selection Rate: $\frac{1-p1}{1-p2}$

Though legal practitioners in the UK and the EU use risk difference and risk ratio respectively to compare benefits denied between different groups, the United States uses the selection rate to compare benefits granted. (Zliobaite, 2015)

Regardless of the measure, a threshold for discrimination is needed. What constitutes unfairness? Let us look at a legal measure that is commonly referenced in computer science papers exploring fairness. When faced with establishing a guideline threshold for discrimination based on statistical measures, the Equal Employment Opportunity Commission (EEOC) recommended the 4/5ths rules for assessing disparate impact. In the EEOC Uniform Guidelines on Employment Selection Procedures, a legal case for employment discrimination could be justified when compared to a privileged group another group has less than an 80% chance of having the same opportunity with a confidence interval of 95%.[17]Looking at automated facial analysis technology, a similar principle could be applied to accuracy. If a specific group has accuracy that is less than 80% of the overall accuracy advertised by a vendor, then the system can be said to risk a disparate impact on that group. When looking at accuracy instead of employment opportunities, we can assess bias by calculating disparate impact risks for specific demographic groups. Disparate impact specifically deals with the notion of harm, which in the case of employment is defined as having less chance of obtaining a job than someone of a different demographic with comparable qualifications. A case for economic harm can be made.

While the 4/5ths threshold provides a borrowed guideline that informs the assessment of disparate impacts, aiming to pass a minimum threshold or accuracy between different subgroups should not preclude efforts to develop algorithms that work exceptionally well for all groups. Focusing on maximizing benefits, which in this case means optimal accuracy on all groups, has the side effect of reducing harms caused by misclassification. Still, we should keep in mind that even accurate algorithms, especially those that classify demographic information like gender and race, can be employed in ways that perpetuate discrimination. In Chapter 2, I explore the potential harms that can arise from gender misclassification in addition to how accurate information about gender can be used to treat one gender group worse than another.

4.5.3 Relevant Populations and Data Repair and Accuracy Reporting

With automated facial analysis, the datasets used to train algorithms can lead to varying levels of accuracy on specific subgroups due to a lack of diversity in the data. Datasets should be reflective of the populations that will be impacted by their use. Data that is representative of the relevant populations should be used to train and benchmark algorithms. However, determining a relevant population and the appropriate representation is not simple. Overrepresentation and underrepresentation is dependent on the definition of the population of interest. To deal with the issue of relevant populations, the Castanda rule was introduced. The rule states for work opportunities the number of people selected from a protected group should not be smaller than 3 standard deviations from the number expected in a random selection. Though this could be applied to training datasets, ultimately the outcomes of using the data should be the main focus. It could be the case that a biased training set can be manipulated to produce acceptable results. Skewed data is common in machine learning, and there are a number of approaches for data repair that attempt to mitigate data bias.

[17] Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.4(D) (2015) https://www.law.cornell.edu/cfr/text/29/part-1607

If automated facial analysis were only used on a specific demographic group, a homogenous dataset could be permissible if there is a low enough chance of encountering someone outside of the group. However, automated facial analysis is used on heterogeneous populations and in locations like airports where people of multiple ethnicities congregate. Studies have shown that facial recognition systems developed in Western and East Asian countries tend to perform better on their respective populations (Phillips et al., 2011). As a result, we need to test performance on the relevant populations on which the systems will be used instead of accepting performance results on benchmarks that are not reflective of target populations. Acceptable accuracy reports should be based on accuracies for subgroups of target populations.

4.5.4 Assessing Fairness Requires Better Benchmarks

The public debate over algorithmic fairness initiated by Angwin's recidivism risk score investigation and algorithmic academic discourse have focused on predictive machine learning models. These models predict an unknown future outcome on an unobservable trait. This trait -for example the likelihood to commit a crime- can be viewed as innate or the result of external factors. Less attention has been focused on the accuracy of algorithms assessing readily verifiable traits like apparent gender in a given image. Accuracy for a biometric classification does not rely on an uncertain future. The gender or age of an individual can be known in a given instance. Because fairness is based on different contextual assumptions and optimizations for accuracy, this thesis aims to show why we need more rigorous reporting on the accuracy rates on which fairness debates center. This thesis focuses on deterministic accuracy for gender classification. Here fairness is defined by equal accuracy within a threshold margin of error for gender classification. Overall, this work highlights the need for transparency and accountability for automated facial analysis. What is the composition of the data we are using to train and test systems? How are we reporting accuracy? Better processes for transparency and accountability can provide nuance in assessing fairness agnostic of the criteria being used.

.

5. Parliaments Benchmark & Dataset Diversity

A dataset of the people and for the people. No prediction without representation.

5.1 Overview

Face datasets are samples meant to represent populations. Politicians, like images in a dataset, are also representatives of populations. And like images in a dataset, they do not always neatly resemble the population. This parallel inspired the creation of Pilot Parliaments Benchmark (PPB) for intersectional benchmarking. The Pilot Parliaments Benchmark consists of 1270 individuals from three African countries (Rwanda, Senegal, South Africa) and three European countries (Iceland, Finland, Sweden) selected for gender parity and skin type. Figure 12 provides as sample of the images included in PPB.

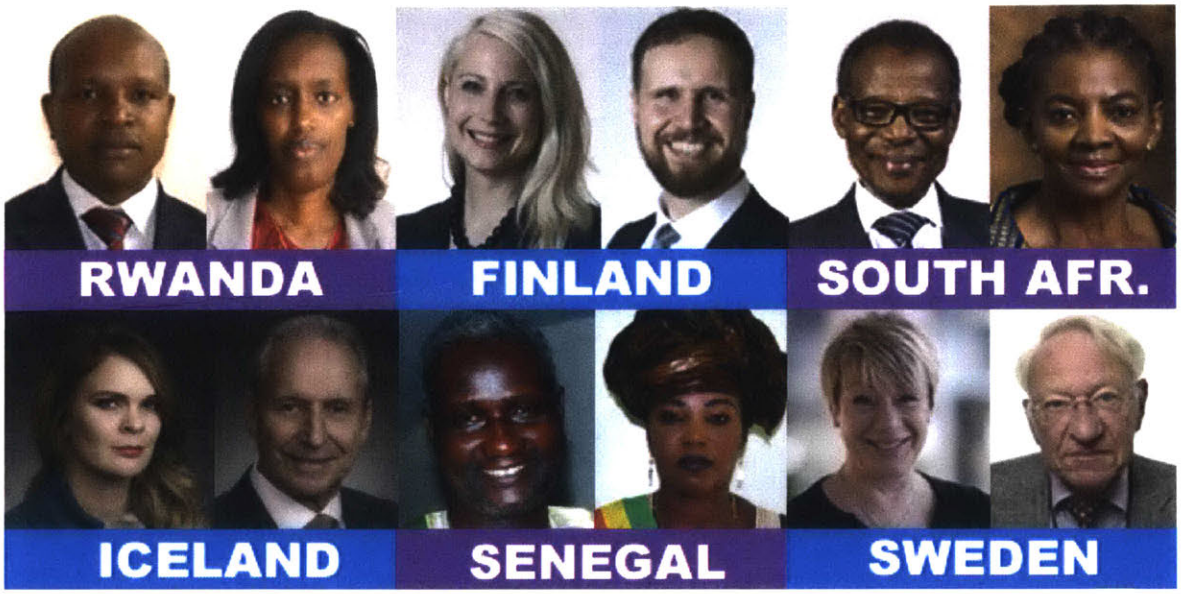


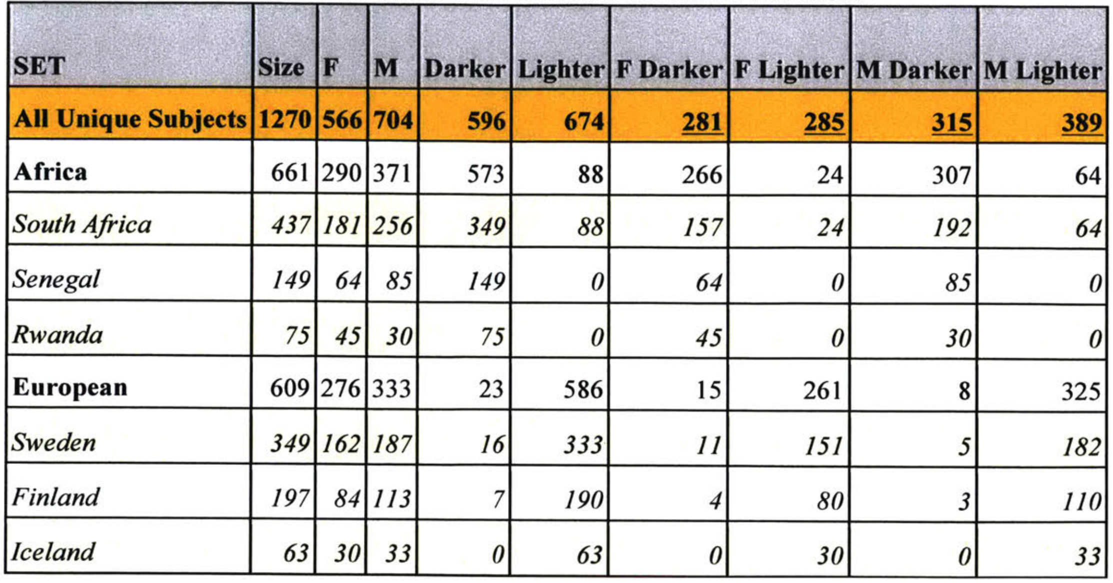
Figure 12. Sample Images from Pilot Parliaments Benchmark

This chapter presents the rationale for constructing the Pilot Parliaments Benchmark (PPB), the curation methodology for collecting a phenotype attentive dataset, the labeling scheme used to annotate images in the dataset, existing datasets that were also labeled with skin type for this research, and the limitations of PPB. The chapter concludes with a comparison of this new benchmark to existing benchmarks with respect to gender and skin type.

5.2 Benchmark Rationale

To conduct an intersectional evaluation of gender classification algorithms, I needed a new benchmark with greater phenotypic representation of skin type.

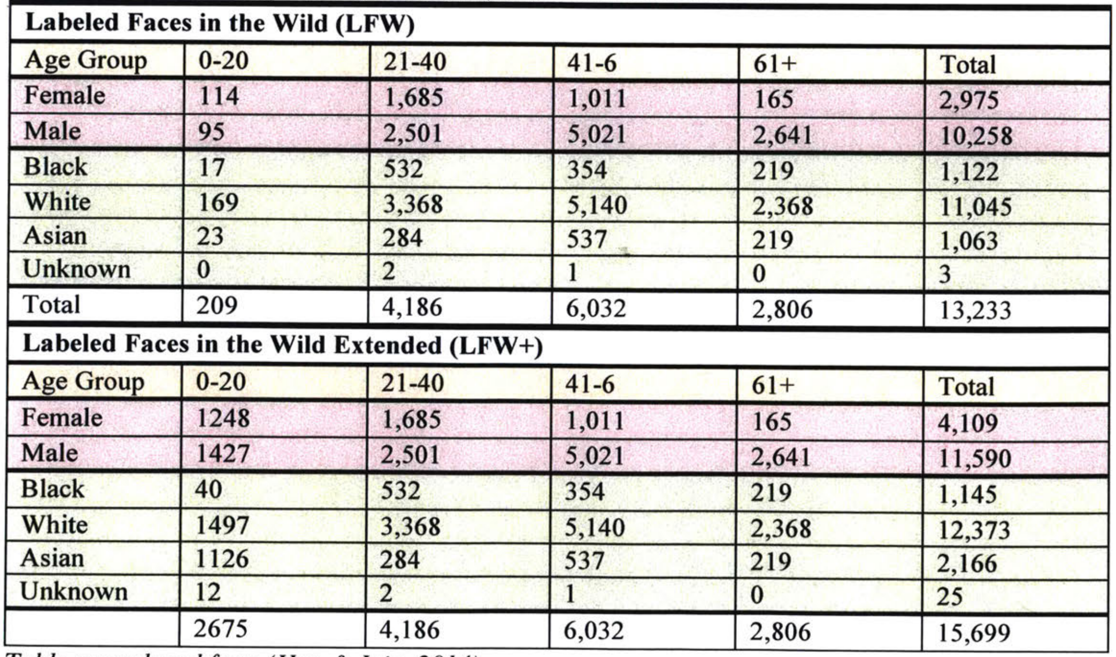
TABLE 1. PILOT PARLIAMENTS BENCHMARK DECOMPOSITION



SET	Size	F	M	Darker	Lighter	F Darker	F Lighter	M Darker	M Lighter
All Unique Subjects	**1270**	**566**	**704**	**596**	**674**	**281**	**285**	**315**	**389**
Africa	661	290	371	573	88	266	24	307	64
South Africa	*437*	*181*	*256*	*349*	*88*	*157*	*24*	*192*	*64*
Senegal	*149*	*64*	*85*	*149*	*0*	*64*	*0*	*85*	*0*
Rwanda	*75*	*45*	*30*	*75*	*0*	*45*	*0*	*30*	*0*
European	609	276	333	23	586	15	261	8	325
Sweden	*349*	*162*	*187*	*16*	*333*	*11*	*151*	*5*	*182*
Finland	*197*	*84*	*113*	*7*	*190*	*4*	*80*	*3*	*110*
Iceland	*63*	*30*	*33*	*0*	*63*	*0*	*30*	*0*	*33*

Requirement - Gender Balance and Greater Representation of People of Color
Popular public benchmarks for automated facial analysis that have been assessed for representativeness tend to exhibit demographic skews in relation to ethnicity and/or gender (see Table 2).

TABLE 2. LFW/ LFW+ AGE, GENDER, AND ETHNIC COUNTS



Labeled Faces in the Wild (LFW)					
Age Group	0-20	21-40	41-6	61+	Total
Female	114	1,685	1,011	165	2,975
Male	95	2,501	5,021	2,641	10,258
Black	17	532	354	219	1,122
White	169	3,368	5,140	2,368	11,045
Asian	23	284	537	219	1,063
Unknown	0	2	1	0	3
Total	209	4,186	6,032	2,806	13,233
Labeled Faces in the Wild Extended (LFW+)					
Age Group	0-20	21-40	41-6	61+	Total
Female	1248	1,685	1,011	165	4,109
Male	1427	2,501	5,021	2,641	11,590
Black	40	532	354	219	1,145
White	1497	3,368	5,140	2,368	12,373
Asian	1126	284	537	219	2,166
Unknown	12	2	1	0	25
	2675	4,186	6,032	2,806	15,699

Table reproduced from (Han & Jain, 2014)

Moreover, many of the publicly available datasets tend to be composed of celebrity faces, which are not representative of the general population. For example, the IMDB-WIKI[18] dataset which its creators position to be the largest public face dataset with labels for age and gender exhibits a strong age skew as shown in Figure 13. In addition, celebrities who face pressure to maintain cultural beauty standards tend to have body types that are not reflective of the general population.

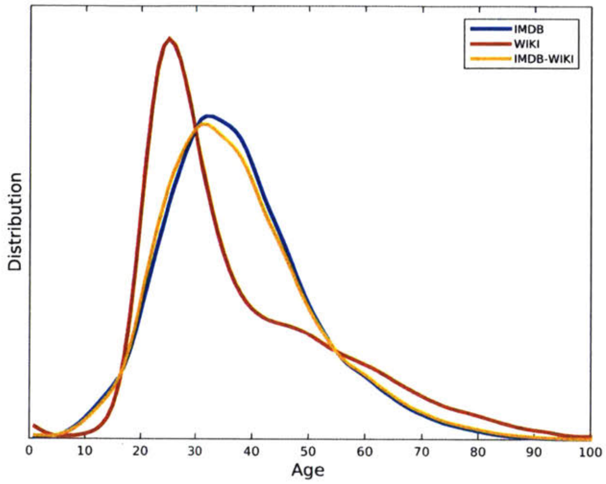


Figure 13. IMDB WIKI Age Distributions

Requirement - Faces with Minimal Privacy Concerns
Celebrity photos are often used because of availability and fewer concerns with privacy. Furthermore collecting and distributing images can be challenging given copyright laws that limit the images that can be made publicly available. As a result, many research datasets are available by request only. For example, the CASIA-WebFaces dataset, which has been used to train a number of notable face recognition algorithms is not publicly available (Yi et al., 2014).

Requirement - Publicly Available
Because this thesis is focused on increasing demographic and phenotypic transparency with automated facial analysis, constructing a diverse and publicly available dataset was a requirement for PPB. Subsequent benchmarks can be governed in a way that prevents vendors from gaming accuracy tests.

5.3 Phenotype and Gender Parity-Aware Curation Methodology

To satisfy these requirements, I 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 phenotypic (skin type) diversity to the training set, I chose to use parliamentarians from African and European countries. The map below

[18] IMDB-WIKI Dataset is available at https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/ .

shows an approximated distribution of average skin types around the world based on UV exposure.

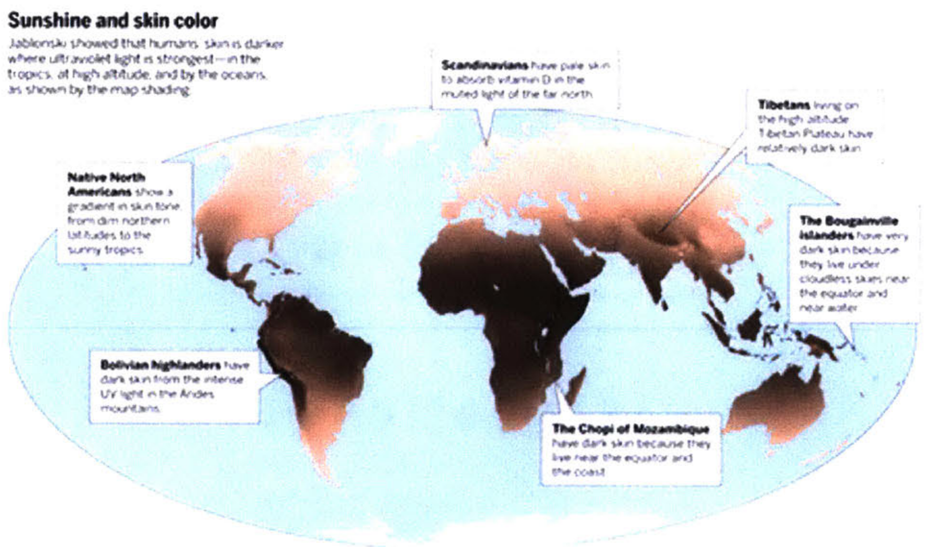

Sunshine and skin color
Jablonski showed that humans skin is darker where ultraviolet light is strongest—in the tropics, at high altitude, and by the oceans, as shown by the map shading
Scandinavians have pale skin to absorb vitamin D in the muted light of the far north
Tibetans living on the high altitude Tibetan Plateau have relatively dark skin
Native North Americans show a gradient in skin tone from dim northern latitudes to the sunny tropics
The Bougainville islanders have very dark skin because they live under cloudless skies near the equator and near water
Bolivian highlanders have dark skin from the intense UV light in the Andes mountains
The Chopi of Mozambique have dark skin because they live near the equator and the coast

Figure 14. Global Map of Skin Color
Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103266/

The African and European countries were selected based on their ranking for gender parity as assessed by the Inter Parliamentary Union. Of all the countries in the world, Rwanda has the highest proportion of women in parliament. Scandinavian countries were also well represented in the top 10 nations. Given the gender parity and the 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 added to the list.

TABLE 3. WOMEN IN PARLIAMENT WORLD RANKINGS

Rank	*Country*	*Elections*	*Seats**	*Women*	*% W*
1	***Rwanda***	*16.09.2013*	*80*	*49*	*61.30%*
2	*Bolivia*	*12.10.2014*	*130*	*69*	*53.10%*
3	*Cuba*	*03.02.2013*	*612*	*299*	*48.90%*
4	***Iceland***	*29.10.2016*	*63*	*30*	*47.60%*
5	*Nicaragua*	*06.11.2016*	*92*	*42*	*45.70%*
6	***Sweden***	***14.09.2014***	***349***	***152***	***43.60%***
7	***Senegal***	***01.07.2012***	***150***	***64***	***42.70%***
8	*Mexico*	*07.06.2015*	*500*	*213*	*42.60%*
9	***Finland***	*19.04.2015*	*200*	*84*	*42.00%*
"	***South Africa***	***07.05.2014***	***398***	***167***	***42.00%***

** Rankings as of May 2017 - Full Rankings available at http://www.ipu.org/wmn-e/classif.htm -Table Shows Single or Lower House Seats*

Image Collection

Six web scrapers were developed to retrieve images of parliamentarians along with available metadata like name and birth date where available. The images were collected from the official websites of the parliaments. Translation services were used to aid in the scraping of websites in Finnish and French. The images were collected for research purposes, which do not appear to be in violation with the use of terms of the websites. Different countries have varying policies on the use of government photos.

5.4 Label Selection

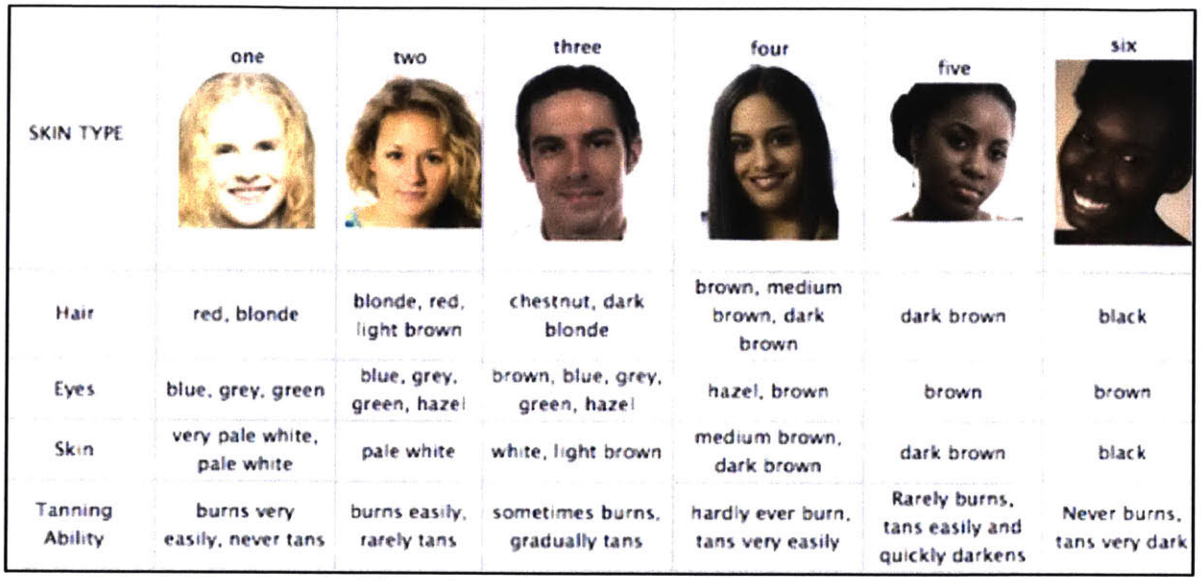
To assess accuracy for classification by not just gender but also the intersection of gender and skin type, PPB needed both demographic labels for gender and phenotypic labels for skin type.

Binary Gender Labels

In adherence with the binary construction of gender used by the algorithms audited, I chose the reductive binary gender labels of female and male.

Fitzpatrick Skin Type Labels

Though the demographic of race or ethnicity is used to label some face datasets as a way of accounting for diversity, race and ethnic labels can include faces with large intraclass variation. In the United States, individuals who self identify as Black exhibit a wide variety of facial geometries and skin types. Thus skin type was used to account for phenotypic diversity. Furthermore, skin type was chosen as a phenotypic factor of interest because default camera settings tend not to properly expose darker skin. Poorly exposed images that result from sensor optimizations for lighter skin 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.



SKIN TYPE	one	two	three	four	five	six
Hair	red, blonde	blonde, red, light brown	chestnut, dark blonde	brown, medium brown, dark brown	dark brown	black
Eyes	blue, grey, green	blue, grey, green, hazel	brown, blue, grey, green, hazel	hazel, brown	brown	brown
Skin	very pale white, pale white	pale white	white, light brown	medium brown, dark brown	dark brown	black
Tanning Ability	burns very easily, never tans	burns easily, rarely tans	sometimes burns, gradually tans	hardly ever burn, tans very easily	Rarely burns, tans easily and quickly darkens	Never burns, tans very dark

Figure 15. Fitzpatrick Skin Type Chart
Source: http://www.skincancer.org/

The Fitzpatrick classification system is used to categorize skin type in PPB. Dermatologists use the Fitzpatrick Scale to assess skin cancer risk. It is a classification of skin response to UV radiation (Fitzpatrick, 1988).

The six-point Fitzpatrick classification system is skewed to lighter skin and has three categories that can be applied to people perceived as White. 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 exploring algorithmic performance by skin type.

5.5 Benchmark Limitations

Limited to Face Detection and Face Classification

The Pilot Parliaments Benchmark (PPB) only contains one image of each unique subject in the dataset. Single images of parliamentarians do not make this a suitable benchmark for facial identification or face verification. However, the methodology used to collect these images and corresponding metadata could be extended to expand the dataset so it can be used for facial recognition benchmarking. For example the names of the parliamentarians can be used to seed searches for additional photos.

Constrained dataset

PPB is highly constrained since it is composed of official profile photos of parliamentarians. The majority of the photos are taken in a setting where pose is fixed, illumination is constant, and expressions are neutral or smiling. By using a constrained dataset to test the algorithms, variations in a performance that are impacted by pose, illumination, or expression are limited. Future work should explore an unconstrained benchmark.

Diversity

PPB consists of a limited set of parliamentarians from African and European countries. An expanded benchmark could incorporate parliamentarians from other continents. More work can be done to include other phenotypic measures like facial landmarks. The dataset is also composed of parliamentarians who by law have to be over a certain minimum age.

Label Accuracy

Gender labels were determined based on the name of the parliamentarian, gendered title and prefixes such as Mr or Ms, the pronouns used to describe the individual where available, and the appearance of the photo. The data was manually annotated by one research assistant, which introduces the potential for bias. Future work will incorporate multiple annotations per image and metrics on intercoder reliability.

Skin type labels were determined based on a subjective assessment of the subject in a photo using guidance from www.skincancer.org to guide labeling. The descriptions for the Fitzpatrick type are ill suited for individuals of Asian origin with fair skin and dark eyes. The annotator for PPB was not officially trained in assessing Fitzpatrick skin type; however, given that the majority of the parliamentarians were on opposite ends of the scale it was deemed suitable to have a non-dermatologist label the dataset. Moreover, the range of human skin tones as demonstrated in the Humanae project well exceeds a six-point classification scale that is used to assess skin type (see Figure 16). The Fitzpatrick classification system is only a starting point for labeling skin type differences.

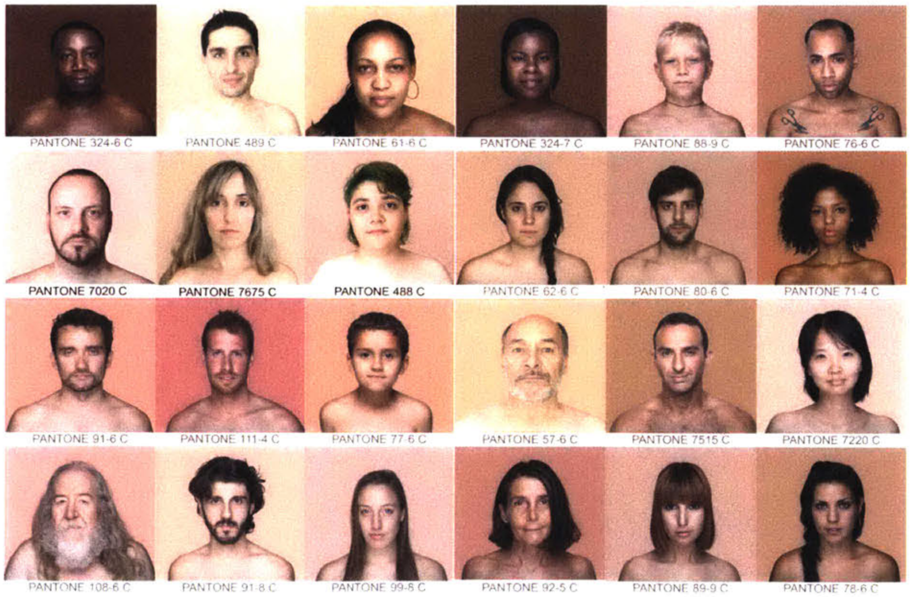


Figure 16. Humanae Project Highlighting Unique Skin Tones
Photographer: Angélica Dass

Dataset Size

PPB is a relatively small dataset comprised of 1270 images of unique individuals and would not be suitable for training deep neural networks. The benchmark is a starting point for constructing a more inclusive benchmark for auditing gender classification algorithms.

5.6 Benchmark Distribution

The PPB dataset will be released under a Creative Commons license. The release does not endorse using images of elected official in ways that suggest individuals depicted support a particular product, service, or entity. Longer-term, the aim is to expand the benchmark and add more parliamentarians from different regions of the world, including the Caribbean, Central America, South America, East Asia, South Asia, the South Pacific, Australia, and the Middle East.

5.7 Additional Datasets and Labeling

Since the goal of PPB was to create a dataset with a wider representation of skin types, the success of this aim is assessed by comparing the Fitzpatrick distribution of PPB with two baseline datasets.

The first dataset chosen for comparison is the IJB-A dataset. IJB-A is described as the most geographically diverse dataset collected by the National Institute for Standards and Technology (NIST). At the time of assessment, 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.

Finally, the Adience dataset, which was released to serve as a benchmark for gender and age classification, was also labeled with the Fitzpatrick type. The Adience benchmark contains 2,284 unique individual subjects. 2,194 of those subjects had first reference images that were discernable 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.

5.8 Gender and Skin Type Distributions

TABLE 4. PPB FITZPATRICK DISTRIBUTION

Classification	Female			Male			Total	
Skin Type	**Count**	**%Type**	**%Total**	**Count**	**%Type**	**%Total**	**Count**	**Type %**
VI	118	38.06%	9.29%	192	61.94%	15.12%	310	24.41%
V	137	59.05%	10.79%	95	40.95%	7.48%	232	18.27%
IV	26	48.15%	2.05%	28	51.85%	2.20%	54	4.25%
III	71	43.56%	5.59%	92	56.44%	7.24%	163	12.83%
II	170	39.35%	13.39%	262	60.65%	20.63%	432	34.02%
I	44	55.70%	3.46%	35	44.30%	2.76%	79	6.22%
Totals	566		44.57%	704		55.43%	1270	100.00%

TABLE 5. IJB-A FITZPATRICK DISTRIBUTION

Classification	Female			Male			Total	
Skin Type	**Count**	**%Type**	**%Total**	**Count**	**%Type**	**%Total**	**Count**	**Type %**
VI	4	13.79%	0.80%	25	86.21%	5.00%	29	5.80%
V	5	50.00%	1.00%	5	50.00%	1.00%	10	2.00%
IV	13	20.63%	2.60%	50	79.37%	10.00%	63	12.60%
III	57	25.91%	11.40%	163	74.09%	32.60%	220	44.00%
II	37	22.42%	7.40%	128	77.58%	25.60%	165	33.00%
I	7	53.85%	1.40%	6	46.15%	1.20%	13	2.60%
Totals	123		24.60%	377		75.40%	500	100.00%

TABLE 6. ADIENCE FITZPATRICK DISTRIBUTION

Classification	Female			Male			Total	
Skin Type	**Count**	**%Type**	**%Total**	**Count**	**%Type**	**%Total**	**Count**	**Type %**
VI	16	55.17%	0.73%	13	44.83%	0.59%	29	1.32%
V	60	62.50%	2.73%	36	37.50%	1.64%	96	4.38%
IV	86	48.59%	3.92%	91	51.41%	4.15%	177	8.07%
III	269	55.35%	12.26%	217	44.65%	9.89%	486	22.15%
II	657	49.14%	29.95%	680	50.86%	30.99%	1337	60.94%
I	53	76.81%	2.42%	16	23.19%	0.73%	69	3.14%
Totals	1141		52.01%	1053		47.99%	2194	100.00%

TABLE 7. PPB, IJB-A, AND ADIENCE FITZPATRICK SIX CLASS COMPARISON

DATASET	I		II		III		IV		V		VI		TOTAL
PPB	**6.22%**	79	**34.02%**	432	**12.83%**	163	**4.25%**	54	**18.27%**	232	**24.41%**	310	1270
IJB-A	**2.60%**	13	**33.00%**	165	**44.00%**	220	**12.60%**	63	**2.00%**	10	**5.80%**	29	500
ADIENCE	**3.14%**	69	**60.94%**	1337	**22.15%**	486	**8.07%**	177	**4.38%**	96	**1.32%**	29	2194

TABLE 8. PPB, IJB-A, AND ADIENCE FITZPATRICK LIGHTER AND DARKER SKIN

DATASET	LIGHTER SKIN (I , II, III)		DARKER SKIN (IV, V, VI)		TOTAL
PPB	**53.07%**	674	**46.93%**	596	1270
IJB-A	**79.60%**	398	**20.40%**	102	500
ADIENCE	**86.24%**	1892	**13.76%**	302	2194

Tables 4 - 8 show skin type distribution for unique subjects in PPB, IJB-A, and Adience. For the purposes of analysis, lighter skin will refer to faces with a Fitzpatrick skin type of I,II, or III. Darker skin will refer to faces labeled with a Fitzpatrick skin type of IV,V, or VI.

While all the datasets have more lighter skinned unique individuals, PPB is around half light at 53.07% whereas the proportion of unique subjects with lighter skin in IJB-A and Adience is 79.6% and 86.24% respectively. PPB provides substantially more darker-skinned unique subjects than the IJB-A and Adience. Even though Adience has 2194 unique subjects labeled, which is nearly twice that of the 1270 subjects in PPB, it has 302 darker skinned subjects, nearly half the 596 darker skinned subjects in PPB. Overall, PPB has a more balanced representation of lighter and darker skin as compared to the IJB-A and Adience datasets.

5.9 Intersectional Subgroup Representation

Tables 9 - 11 show the intersectional breakdown of the unique subjects of the data in respect to gender and skin type.

TABLE 9. PPB-A INTERSECTIONAL FITZPATRICK DISTRIBUTION

UV Response			
Types IV - VI	Female	Male	Types IV - VI Total
Count	281	315	**596**
%	22.13%	24.80%	**46.93%**
Types I - III	**Female**	**Male**	**Types I - III Total**
Count	285	389	**674**
%	22.44%	30.63%	**53.07%**
	44.57%	55.43%	100.00%

TABLE 10. IJB-A INTERSECTIONAL FITZPATRICK DISTRIBUTION

UV Response			
Types IV - VI	Female	Male	Types IV - VI Total
Count	22	80	**102**
%	4.40%	16.00%	**20.40%**
Types I - III	**Female**	**Male**	**Types I - III Total**
Count	101	297	**398**
%	20.20%	59.40%	**79.60%**
	24.60%	75.40%	100.00%

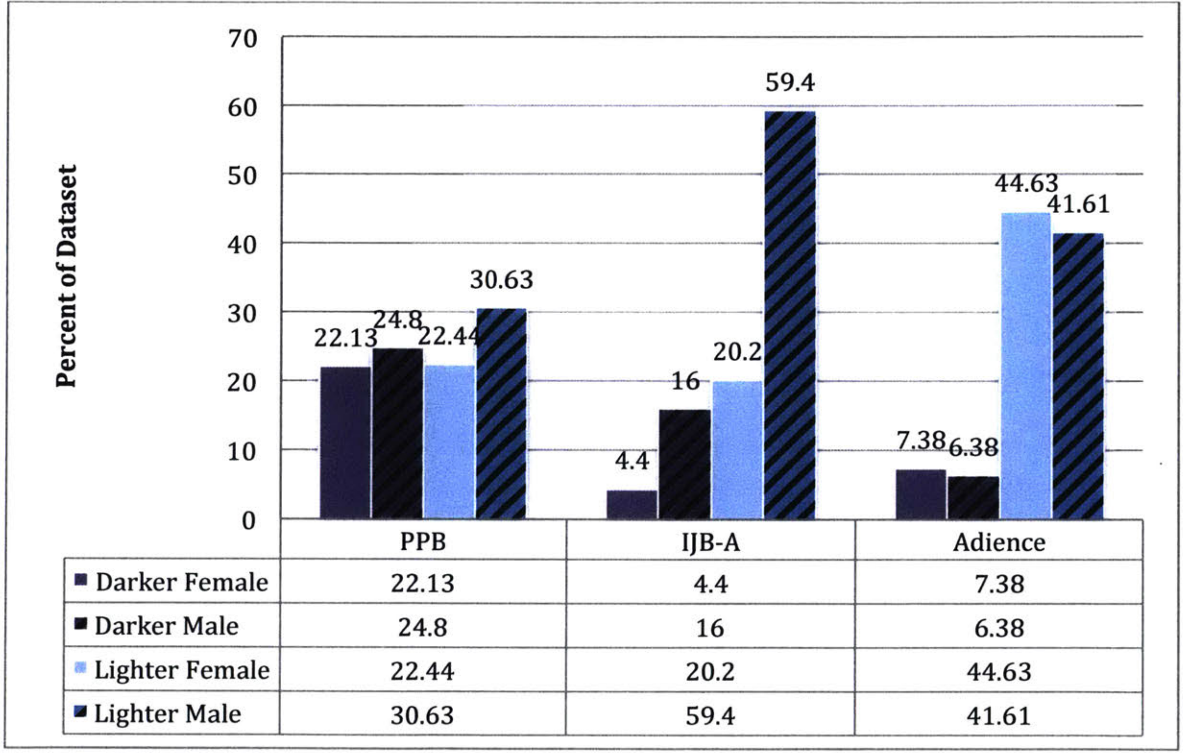
TABLE 11. ADIENCE INTERSECTIONAL FITZPATRICK DISTRIBUTION

UV Response			
Types IV - VI	Female	Male	Types IV - VI Total
Count	162	140	**302**
%	7.38%	6.38%	**13.76%**
Types I - III	**Female**	**Male**	**Types I - III Total**
Count	979	913	**1892**
%	44.62%	41.61%	**86.24%**
	52.01%	47.99%	100.00%

PPB provides the most balanced representation of darker skinned females, darker skinned males, lighter skinned females, and lighter skinned males. IJB-A has the least balanced distribution (see Table 12).

Faces labeled with darker skin and female are the least represented in the IJB-A (4.4%). Faces labeled with darker skin and male are the least represented in the Adience (6.38%) dataset. Faces labeled with lighter skin and male annotations are the most represented unique subjects in all datasets. IJB-A is composed of 59.4% unique lighter skinned males. Adience is composed of 41.6%, and PPB is 30.63%.

TABLE 12. PPB, IJB-A, AND ADIENCE INTERSECTIONAL COMPARISON



	PPB	IJB-A	Adience
■ Darker Female	22.13	4.4	7.38
■ Darker Male	24.8	16	6.38
■ Lighter Female	22.44	20.2	44.63
▨ Lighter Male	30.63	59.4	41.61

5.10 Analysis

Inclusion Priorities
As a gender benchmark, Adience was constructed to have gender parity. It is unclear to what extent the creators of the benchmark considered skin type diversity. IJB-A, which serves as a recognition and detection benchmark, was constructed to be geographically diverse, but its unique subjects do not reflect the geographic distribution of gender worldwide.

Diversity of online images
Adience uses images uploaded to Flickr and by construction achieves gender parity. However, the limited representation of individuals with darker skin could be a reflection

of the Flickr user base at the time of image collection. The personal photos uploaded are presumably of family members and friends. A largely homogeneous user base can limit the phenotypic diversity of available images. Furthermore, images for the Adience dataset were automatically collected using facial detection software (Kemelmacher-Shlizerman et al., 2016). Systematic facial detection failure on individuals with darker skin could also have limited inclusion in the dataset.

Power Distribution and Media Attention Influence Data Availability

The IJB-A benchmark is composed of public figures. The gender skew of the unique subjects in IJB-A reflect structural factors that have led to public figures being predominantly male. The UN Women in Politics report shows that worldwide, women make up only 5.7% (11/193) of heads of governments as of January 2017. Available media is another factor shaping the composition of the images in the IJB-A dataset. Countries that receive less international coverage are less likely to have as many images and videos of their public figures available online.

Diversity priorities shape the composition of datasets. More inclusive datasets need the explicit prioritization of diverse demographic and phenotypic representation. One without the other is not sufficient. Collecting a gender-balanced dataset will not assure phenotypic variation nor will collecting a geographically diverse dataset assure gender parity.

Due to existing structures of power and skews in media creation, constructing diverse datasets will take intentional effort. The reliance on automated face detection can limit the diversity of images collected from online sources. The sources themselves may also lack phenotypic diversity. PPB was constructed to prioritize both gender parity and skin type representation. The benchmark only focuses on one demographic attribute, gender, and one phenotypic attribute, skin type. Additional demographics like age and phenotypes such as facial geometry can shape the creation of more representative datasets.

6. Gender Classification Performance Evaluation

6.1 Overview

Advances in deep learning have shown promising results for automated facial analysis tasks like gender classification (Levi & Hassner, 2015). Yet benchmarks like Adience and IJB-A used to evaluate and compare algorithmic performance on automated facial analysis underrepresent individuals with darker skin (see Section 5.9). As the use of automated facial analysis increases, we need to assess the accuracy of these algorithms across a diverse range of people. Even though more people of color are increasingly using and being impacted by these technologies, little work has been done to explicitly assess algorithmic performance on individuals with darker skin. This chapter presents a performance evaluation of four gender classification algorithms on the Pilot Parliaments Benchmark (PPB). PPB was developed for this thesis to support intersectional algorithmic performance evaluation on the following groups: darker-skinned females, lighter-skinned females, darker-skinned males and lighter-skinned males.[19]

Key Findings on Classifiers Evaluated

- Gender Classifiers perform better on male faces than female faces (10 - 21% accuracy difference)
- Gender Classifiers perform better on lighter skin than darker skin (9 - 20% accuracy difference)
- Classifiers perform worst on darker females (20 - 37% error rate)
- Classifiers perform best on lighter males (~1% error rate)
- The Error Rate Gap between best classified group (lighter males) and worst classified group (darker females) is as much as 36%

6.2 Target Gender Classification Algorithms

Gender classification algorithms have reached a state where technology companies sell them as a service. The algorithms are marketed in a manner that suggests universal performance across all faces. I chose to focus on binary gender classification algorithms due to increased commercial confidence in their performance and the low cost of conducting a performance evaluation on available gender classifiers. Three commercial gender classifiers and one freely available for research applications were chosen for evaluation.

Research Gender Classifier: Adience

On the research end, convolutional neural networks (CNNs) have produced the most promising results for automated facial analysis tasks like face verification. The Adience

[19] Darker and Lighter Skin Designations are based on the Fitzpatrick Skin Type Scale (see Chapter 5.).

gender classification algorithm was selected because it uses a basic implementation of a CNN architecture for classification in keeping with the state-of-the-art. The Adience gender classification model is publicly available, and it can be fine-tuned with additional training data.

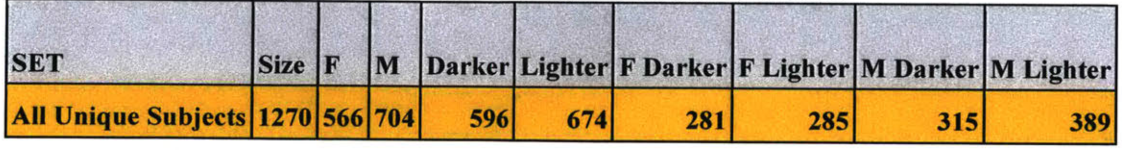
Commercial Gender Classifiers: Microsoft, IBM, and Face++
On the commercial side, I focus on gender classifiers sold in API bundles made available by Microsoft, IBM, and Face++. Microsoft's Cognitive Services Face API and IBM's Watson Visual Recognition API were chosen since both companies have made significant investments in artificial intelligence, capture significant market share 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 facial 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 company headquartered in China, was thus chosen to see if this observation holds for gender classification. Like Microsoft and IBM, Face++ also provides a publicly available demonstration of their gender classification capabilities.

6.3 Evaluation Overview

6.3.1 Gender Classification Benchmark

The Pilot Parliaments Benchmark was created to better assess algorithmic performance across intersection subgroups with regard to gender and skin type. 1270 images of unique individuals representing Rwanda, Senegal, South Africa, Sweden, Finland, and Iceland make up PPB. See section 5.1 for gender, skin type, and intersectional breakdown of benchmark by region and country.

TABLE 13. PPB SKIN, GENDER, AND INTERSECTIONAL COUNTS



SET	Size	F	M	Darker	Lighter	F Darker	F Lighter	M Darker	M Lighter
All Unique Subjects	1270	566	704	596	674	281	285	315	389

6.3.2 Evaluation Metrics

Aggregate Classification Accuracy
In following the gender classification evaluation precedent established by the National Institute for Standards and Technology (NIST), I assess the overall classification accuracy, male accuracy, and female accuracy for all fours classifiers on PPB. These accuracies are combined with the percentage of misclassified males and females to create confusion matrices for each gender classifier.

Disaggregated Classification Accuracy
The male and female accuracies are then evaluted by country and region.

Male and Female Error Rates
To conduct a demographic performance analysis, the differences in male and demale error rates for each gender classifer is compared first in aggregate and them by region and country.

Darker and Lighter Skin Type Error Rates
To conduct a phenotypic performance analysis, the differences in darker and lighter skin type error rates for each gender classifer is compared first in aggregate and then by region and country.

Intersectional Error Rates
To conduct an intersectional demographic and phenotypic analsyis, the error rates for four intersectional groups (darker-skinned females, lighter-skinned females, darker-skinned males and lighter-skinned males) are compared in aggregate and then by region and country.

Subgroup Contribution to Aggregate Error Rates
Finally, the error rates for each gender classifier are disagegated by gender, skin type, and intersectional subgroup.

6.4 Classification Accuracy

In following the precedents established by the National Institute of Standards and Technology (NIST), I evaluate the gender classifiers out of the box to approximate real world use cases. The intrepid developer looking to quickly integrate gender classification into software might not change any default configurations. Table 14 presents the accuracy of the Adience, IBM, Face++, and Microsoft gender classifiers on the Pilot Parliaments Benchmark, and Table 15 presents accuracies broken down by country.

TABLE 14. AGGREGATE GENDER CLASSIFICATION ACCURACY

	Adience		**IBM**		**Face++**		**Microsoft**	
Accuracy	*77.81%*		*88.12%*		*89.92%*		*94.78%*	
	Female	Male	Female	Male	Female	Male	Female	Male
%Classified Female	65.88%	12.66%	80.36%	5.54%	78.62%	0.88%	89.46%	0.87%
%Classified Male	34.12%	87.34%	19.64%	94.46%	21.38%	99.12%	10.54%	99.13%

N = 1270, Females = 566, Males=704

The Microsoft gender classifier leads the others with a 94.78% accuracy score, outpacing the next classifier by almost 5 percentage points. The noncommercial Adience gender classifier has the worst performance with a 77.81% accuracy score. Unlike the commercial gender classifiers, the Adience classifier is open source and trainable. With fine-tuning, better results may be achieved. Table 14 shows all algorithms have higher male classification accuracy than female classification accuracy. The differences in accuracy between male and female range from 9.66% to 21.46%. Section 6.5 offers an in-depth analysis of each classifier's male and female error rates.

TABLE 15. GENDER CLASSIFICATION ACCURACY BY COUNTRY

SOUTH AFRICA (N =437)		SWEDEN (N =349)	
Adience_SA	*69.57%*	Adience_SW	*88.10%*
IBM_SA	*85.00%*	IBM _SW	*96.28%*
Face_SA	*86.74%*	Face_SW	*94.54%*
MSFT_SA	*91.71%*	MSFT_SW	*99.14%*
SENEGAL (N =149)		FINLAND (N =197)	
Adience_SE	*71.14%*	Adience_FL	*83.52%*
IBM _SE	*69.44%*	IBM _FL	*97.97%*
Face_SE	*81.40%*	Face_FL	*95.43%*
MSFT_SE	*91.73%*	MSFT_FL	*98.48%*
RWANDA (N =75)		ICELAND (N =63)	
Adience_RW	*74.67%*	Adience_IL	*83.93%*
IBM _RW	*67.12%*	IBM _IL	*100.00%*
Face_RW	*80.95%*	Face_IL	*95.24%*
MSFT_RW	*82.86%*	MSFT_IL	*100.00%*

Since technology companies can train their algorithms on substantially larger datasets than those available to researchers, I expected all of the commercial classifiers to perform better than the noncommercial classifier. Surprisingly, the IBM gender classifier performs worse on the Senegal and Rwanda parliaments than the Adience classifier. Section 6.9 explores how systematic failure on darker skin impacts IBM's overall accuracy.

TABLE 16. IBM VS. ADIENCE ON SENEGAL & RWANDA

SENEGAL (N =149)		RWANDA (N =75)	
Adience_SE	*71.14%*	Adience_RW	*74.67%*
IBM _SE	*69.44%*	IBM _RW	*67.12%*

Adience was benchmarked to have an accuracy of 85.9%, which is in alignment with the results achieved on the Sweden (88.1%), Finland (83.52%), and Iceland (83.93%) parliaments. However, compared to the benchmark performance, accuracy drops by over 10% when Adience classifies the South Africa (69.57%), Senegal (71.14%), and Rwanda subsets (74.67%). Chapter 4 presents an overview of the estimated composition of the Adience benchmark. It is 88.68% of the faces featured that have lighter skin. This skew in the benchmark presents the potential to bias accuracy towards lighter skin.

Overall, the Microsoft gender classifier performs the best on all countries with a perfect result on Iceland. IBM's poor performance on African parliamentarians recovers with a perfect result on the Iceland benchmark. Still, as shown in Table 17, IBM has the largest disparity in performance between the African and European parliaments. Accuracy drops by 17.78% when moving from European classification to African classification. Microsoft maintains its dominant performance with an 8.27% gap between countries representing the Global North and the Global South

TABLE 17. GENDER CLASSIFICATION ACCURACY BY REGION

AFRICA (N =661)		EUROPE (N =609)		DIFFERENCE	
Adience_AF	*70.50%*	Adience_EU	*86.24%*	Adience	15.74%
IBM _AF	*79.43%*	IBM _EU	*97.21%*	IBM	17.78%
Face_AF	*85.05%*	Face_EU	*94.90%*	Face	9.85%
MSFT_AF	*90.74%*	MSFT_EU	*99.01%*	MSFT	8.27%

When announcing state-of-the-art gender classification performance using CNNs, Levi and Hassner, only reported overall accuracy for gender classification (2015). While one accuracy metric may make it easier to compare algorithms, it is not sufficient for revealing subgroup performance that may uncover systematic failures. To see if overall accuracy holds across gender, skin type, and the intersection of gender and skin type, the next sections explore subgroup error rates.

6.5 Error Rates By Gender

TABLE 18. AGGREGATE GENDER ERROR RATES

CLASSIFIER (N=1270)	Female Error (N=566)	Male Error (N=704)	Error Difference
Adience_ALL	34.12%	12.66%	21.46%
IBM _ALL	19.64%	5.54%	14.10%
Face_ALL	21.38%	0.88%	20.49%
MSFT_ALL	10.54%	0.87%	9.66%

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 algorithms evaluated (Ngan & Grother, 2015). The gender error rates on the Pilot Parliaments Benchmark replicate this trend across all algorithms as seen in Table 18. The difference between female error rates and male error rates ranged from 9.66% to 21.46%

Chapter 7 provides a deeper discussion about using different thresholds to make a case for discrimination (better treatment of one group than another) based on the error rates reported in this section.

TABLE 19. GENDER ERROR RATES BY REGION

AFRICA (N=661)	Female Error (N=290)	Male Error (N=371)	Error Difference
Adience_AF	37.24%	23.45%	13.79%
IBM _AF	32.75%	10.76%	21.98%
Face_AF	32.85%	0.58%	32.27%
MSFT_AF	19.01%	1.42%	17.60%
EUROPE (N=609)	(N=276)	(N=333)	
Adience_EU	30.62%	0.00%	30.62%
IBM _EU	6.16%	0.00%	6.16%
Face_EU	9.82%	1.20%	8.62%
MSFT_EU	1.81%	0.30%	1.51%

The tendency for these gender classifiers to perform better on male faces holds on a regional basis as well (Table 19). For both regions nearly all gender errors produced by commercial classification are a result of misclassification of images labeled female. IBM is a notable exception with a 10.76% error rate for males in the African region. All algorithms perform better on the European subset than on the African subset of faces across both genders. The commercial classifiers have an average error rate of 28.20% for females in African parliaments and an average 5.95% error rate for females in European parliaments. They have an average error rate of 4.25% for males in African parliaments and average 0.5% error rate for males in European parliaments

The noncommercial Adience classifier has the highest error rates across both regions and both genders with one exception. The Adience classifier has a 0.00% error rate for males in European parliaments, which is significantly less than the 23.45% error rate for males in African parliaments.

TABLE 20. GENDER ERROR RATES BY COUNTRY

SOUTH AFRICA (N=437)	Female Error (N=181)	Male Error (N=256)	Error Difference
Adience_SA	44.20%	20.70%	23.50%
IBM_SA	28.81%	4.94%	23.88%
Face_SA	31.64%	0.40%	31.24%
MSFT_SA	20.22%	0.00%	20.22%
SENEGAL (N=149)	(N=64)	(N=85)	
Adience_SE	25.00%	31.76%	-6.76%
IBM _SE	38.71%	24.39%	14.32%
Face_SE	39.34%	0.00%	39.34%
MSFT_SE	17.74%	0.00%	17.74%
RWANDA (N=75)	(N=45)	(N=30)	
Adience_RW	26.67%	23.33%	3.33%
IBM _RW	40.00%	21.43%	18.57%
Face_RW	28.21%	4.17%	24.04%
MSFT_RW	15.91%	19.23%	-3.32%
SWEDEN (N=349)	(N=162)	(N=187)	
Adience_SW	26.14%	0.00%	26.14%
IBM _SW	8.02%	0.00%	8.02%
Face_SW	10.56%	1.07%	9.49%
MSFT_SW	1.23%	0.53%	0.70%
FINLAND (N=197)	(N=84)	(N=113)	
Adience_FL	38.46%	0.00%	38.46%
IBM _FL	4.76%	0.00%	4.76%
Face_FL	8.33%	1.77%	6.56%
MSFT_FL	3.57%	0.00%	3.57%
ICELAND (N=63)	(N=30)	(N=33)	
Adience_IL	33.33%	0.00%	33.33%
IBM _IL	0.00%	0.00%	0.00%
Face_IL	10.00%	0.00%	10.00%
MSFT_IL	0.00%	0.00%	0.00%

Table 20 shows the error rate for females and males by country. In general female error rates are higher or equal to male error rates across all countries with two notable exceptions. The Microsoft gender classifier has a higher error rate for males in Rwanda (19.23%) than females in Rwanda (15.91%). Similarly the Adience gender classifier has a higher error rate for males in Senegal (31.76%) than females in Senegal (25%). The performance of commercial classifiers on the Senegal subset varies substantially. While the Face++ and Microsoft classifiers have 0.00% error rates for males in this set, IBM has a 24.39% error rate. IBM's performance on Senegal reflects overall poorer performance on darker skin. Section 6.9 shows the IBM classifier performs worse on darker skin than lighter skin on both males and female.

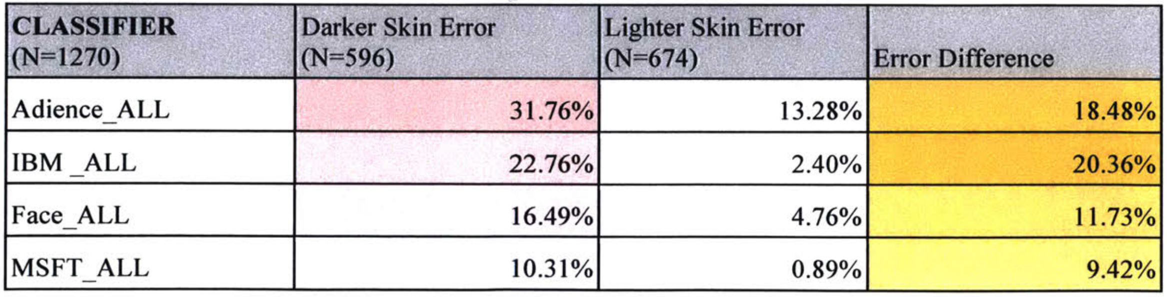
For commercial classifiers the error rates are highest for females in South Africa, Senegal, and Rwanda. For this group, error rates range from 15.91% to 40%. The IBM gender classifier has the worst classification error rate of 40% on detected females in the Rwandan parliament. Face++ is close behind with a 39.34% failure rate on detected females in the Senegalese parliament. Microsoft achieves the best performance with 15.91% error rate on detected females in the Rwandan parliament.

The Microsoft gender classification error rate for males is .53% or less for all countries except for Rwanda where the error rate is 19.23%. The Face++ classifier performs the best on males in the Rwanda subset.

The noncommercial Adience classifier performs male classification well on Sweden, Finland, and Iceland with an error rate of 0.0% on all three. The Adience classifier struggles the most with female classification for the South African parliament at an error rate of 44.2%

6.6 Error Rates by Skin Type

TABLE 21. PPB AGGREGATE SKIN TYPE ERROR RATES



CLASSIFIER (N=1270)	Darker Skin Error (N=596)	Lighter Skin Error (N=674)	Error Difference
Adience_ALL	31.76%	13.28%	18.48%
IBM _ALL	22.76%	2.40%	20.36%
Face_ALL	16.49%	4.76%	11.73%
MSFT_ALL	10.31%	0.89%	9.42%

To attend to skin type phenotypic variation, the images in PPB were labeled with one of six Fitzpatrick skin type categories. Table X shows the error rates for darker skin classified with Fitzpatrick type IV, V, and VI and lighter skin classified with Fitzpatrick type I, II, and III. All algorithms perform better on lighter skin than on darker skin in PPB. Of the commercial classifiers, Microsoft achieves the best result with an error rate of 10.31% on darker skin and an error rate of .89% on lighter skin. On darker skin, IBM achieves the worst classification of the commercial classifiers with an error rate of 22.76%. This error rate is nearly 10 times higher than the IBM error rate on lighter skin.

Overall, the open source Adience gender classifier has the overall worst performance on darker and lighter skin with error rates of 31.76% and 13.28% respectively.

TABLE 22. PPB SKIN TYPE ERROR RATES BY REGION

AFRICA (N=661)	Darker Skin Error (N=573)	Lighter Skin Error (N=88)	Error Difference
Adience_AF	32.46%	10.23%	22.23%
IBM _AF	23.34%	1.25%	22.09%
Face_AF	17.20%	1.15%	16.05%
MSFT_AF	10.75%	0.00%	10.75%
EUROPE (N=609)	(N=23)	(N=586)	
Adience_EU	13.64%	13.77%	-0.13%
IBM _EU	8.70%	2.56%	6.14%
Face_EU	0.00%	5.30%	-5.30%
MSFT_EU	0.00%	1.02%	-1.02%

When focusing on the darker skin error rates for the African region and the lighter skin error rates in for European region, observations made about the overall darker skin and lighter skin error rates in Table 22 follow the trends previously described since most of the darker skinned parliamentarian are in the African region and most of the lighter skin parliamentarians are in the European region. The lighter skin error in the African region is comparable to the lighter skin error in the European region. The darker skin error in the European region is notably less than the darker skin error in the African region. However, since there are only 23 individuals in the European subset with darker skin, looking at the aggregate accuracy on the total number of images (596) labeled with darker skin in PPB is a better indicator of darker skin classification performance. The negative error difference for European classification indicates that performance is better on darker skin, but the subset is only composed of 3.77% images labeled with dark skin. The European regional error difference between light and dark skin classification shows how low representation in a dataset can impact results of subgroup accuracy. In instances where there is low representation, oversampling can be used to increase proportional representation. However if the subgroup sample is not representative of subgroup population, oversampling can provide a misleading view of accuracy.

TABLE 23. PPB SKIN TYPE ERROR RATES BY COUNTRY

SOUTH AFRICA (N=437)	Darker Skin Error (N=349)	Lighter Skin Error (N=88)	Error Difference
Adience_SA	35.53%	10.23%	25.30%
IBM_SA	18.24%	1.25%	16.99%
Face_SA	16.33%	1.15%	15.18%
MSFT_SA	10.40%	0.00%	10.40%
SENEGAL (N=149)	(N=149)	(N=0)	
Adience_SE	28.86%	Not represented	Only Darker Skin
IBM _SE	30.56%	Not represented	Only Darker Skin
Face_SE	18.60%	Not represented	Only Darker Skin
MSFT_SE	8.27%	Not represented	Only Darker Skin
RWANDA (N=75)	(N=75)	(N=0)	
Adience_RW	25.33%	Not represented	Only Darker Skin
IBM _RW	32.88%	Not represented	Only Darker Skin
Face_RW	19.05%	Not represented	Only Darker Skin
MSFT_RW	17.14%	Not represented	Only Darker Skin
SWEDEN (N=349)	(N=16)	(N=333)	
Adience_SW	0.00%	12.46%	-12.46%
IBM _SW	0.00%	3.90%	-3.90%
Face_SW	0.00%	5.72%	-5.72%
MSFT_SW	0.00%	0.90%	-0.90%
FINLAND (N=197)	(N=7)	(N=190)	
Adience_FL	42.86%	15.43%	27.43%
IBM _FL	28.57%	1.05%	27.52%
Face_FL	0.00%	4.74%	-4.74%
MSFT_FL	0.00%	1.58%	-1.58%
ICELAND (N=63)	(N=0)	(N=63)	
Adience_IL	Not represented	16.07%	Only Lighter Skin
IBM _IL	Not represented	0.00%	Only Lighter Skin
Face_IL	Not represented	4.76%	Only Lighter Skin
MSFT_IL	Not represented	0.00%	Only Lighter Skin

Given low representation or no representation of individuals with darker skin in Sweden (16), Finland (7), and Iceland (0), the classifiers' darker error rates ranging from 0% to 42.85% lack statistical strength. All of the commercial gender classifiers perform relatively well on lighter skin across all countries where lighter skin is represented. With the exception of South Africa, Face++ has the highest lighter skin error rates of the commercial classifiers. This difference could be a reflection of a presumed optimization for Asian faces as opposed to European faces. Given significant differences in the distribution of skin types across counties, careful consideration must be made when country of origin is used as a proxy for assessing phenotypic accuracy.

6.7 Intersectional Error Rates by Skin Type and Gender

The error rate by gender and the error rate by skin type show that all algorithms perform better on the male subgroup and the lighter skin subgroup (See Table 18 and Table 21). Here, I present error rates on the intersectional subgroups of darker female, lighter female, darker male, and lighter male.

TABLE 24. PPB INTERSECTIONAL SKIN TYPE AND GENDER TYPE ERROR RATES

CLASSIFIER (N=1270)	Darker Female (N=281)	Lighter Female (N=285)	Darker Male (N=315)	Lighter Male (N=389)	D.Fem.-L. Male
Adience_ALL	36.79%	31.34%	27.30%	0.27%	36.52%
IBM _ALL	34.42%	5.28%	12.17%	0.26%	34.16%
Face_ALL	33.58%	9.86%	0.69%	1.03%	32.55%
MSFT_ALL	19.64%	1.75%	1.68%	0.26%	19.38%

Best Classified vs Worst Classified

For the entire Pilot Parliaments Benchmark, all four gender classifiers perform the worst on the darker female subgroup. With the exception of Face++, these algorithms perform the best on the lighter male subgroup with error rates ranging from .26% to 1.03% The difference in error rates for the darker females and lighter males is substantial. For the darker female subgroup, error rates range from 19.64%to 36.79%. On the commercial side, the IBM and Face++ gender classifier misclassifies 1 out of 3 darker females. The Microsoft classifier misclassifies 1 out of 5 darker females. Conversely, Microsoft and IBM only misclassify .26% (1/385) of the lighter male subgroup. Face++ does slightly worse with a 1.03% error rate on lighter males.

Notable Differences In Error Rates

In general the algorithms perform better on the lighter skin subgroup and the male subgroup than on their counterpart darker skin subgroup and female subgroup. Still, skin type and gender accuracies on their own do not provide information on the difference between intersectional subgroups. Table 25 shows that the difference in error rates between the darker female subgroup and the lighter male subgroups ranges from 19.38% and 36.52%. Though performance is better on the lighter subgroup and the male subgroup, without the intersectional break down provided in Table 25, the performance differences between lighter females and darker males would be obfuscated. The performance difference between lighter females and darker males are mixed with IBM++ performing 6.89% better on lighter females than on darker males. Microsoft has roughly equivalent performance with a difference of .07%. Both Adience and Face++ perform better on darker males than lighter females with error difference of 4.04% and 9.17% respectively.

TABLE 25. SUBGROUP DIFFERENCES IN INTERSECTIONAL ERROR RATES

CLASSIFIER (N=1270)	Female - Male Error Difference	Darker - Lighter Error Difference	D. Female -L. Male Error Difference	L.Female - Darker Male. Error Difference
Adience_ALL	21.46%	18.48%	36.52%	4.04%
IBM _ALL	14.10%	20.36%	34.16%	-6.89%
Face_ALL	20.49%	11.73%	32.55%	9.17%
MSFT_ALL	9.66%	9.42%	19.38%	0.07%

TABLE 26. INTERSECTIONAL SKIN TYPE AND GENDER ERROR RATES BY REGION

AFRICA (N=661)	Darker Female (N=266)	Lighter Female (N=24)	Darker Male (N=307)	Lighter Male (N=64)	DF-LM
Adience_AF	37.59%	33.33%	28.01%	1.56%	36.03%
IBM _AF	35.63%	0.00%	12.50%	1.75%	33.88%
Face_AF	35.57%	4.17%	0.71%	0.00%	35.57%
MSFT_AF	20.77%	0.00%	1.73%	0.00%	20.77%
EUROPE (N=609)	(N=15)	(N=261)	(N=8)	(N=325)	
Adience_EU	21.43%	31.15%	0.00%	0.00%	21.43%
IBM _EU	13.33%	5.75%	0.00%	0.00%	13.33%
Face_EU	0.00%	10.38%	0.00%	1.23%	-1.23%
MSFT_EU	0.00%	1.92%	0.00%	0.31%	-0.31%

For completeness the error rates by region and country are presented in Tables 26 and 27. Given the imbalance of darker female, lighter female, darker male, and lighter male representation across regions and countries, assessing intersectional subgroup performance by region and country is limited by low number counts. The Europe Subgroup for example, only has 2.4% (n=15) darker females and 1.3% (n=8) darker males.

TABLE 27. INTERSECTIONAL SKIN TYPE AND GENDER ERROR RATES BY COUNTRY

SOUTH AFRICA (N=437)	Darker Female (N=157)	Lighter Female (N=24)	Darker Male (N=192)	Lighter Male (N=64)	DF-LM
Adience_SA	45.86%	33.33%	27.08%	1.56%	44.30%
IBM_SA	33.12%	0.00%	5.91%	1.75%	31.36%
Face_SA	35.95%	4.17%	0.53%	0.00%	35.95%
MSFT_SA	23.38%	0.00%	0.00%	0.00%	23.38%
SENEGAL (N=149)	(N=64)	(N=0)	(N=85)	(N=0)	
Adience_SE	25.00%	Not represented	31.76%	Not represented	Only Darker Skin
IBM _SE	38.71%	Not represented	24.39%	Not represented	Only Darker Skin
Face_SE	39.34%	Not represented	0.00%	Not represented	Only Darker Skin
MSFT_SE	17.74%	Not represented	0.00%	Not represented	Only Darker Skin
RWANDA (N=75)	(N=45)	(N=0)	(N=30)	(N=0)	
Adience_RW	26.67%	Not represented	23.33%	Not represented	Only Darker Skin
IBM _RW	40.00%	Not represented	21.43%	Not represented	Only Darker Skin
Face_RW	28.21%	Not represented	4.17%	Not represented	Only Darker Skin
MSFT_RW	15.91%	Not represented	19.23%	Not represented	Only Darker Skin
SWEDEN (N=349)	(N=11)	(N=151)	(N=5)	(N=182)	
Adience_SW	0.00%	27.97%	0.00%	0.00%	0.00%
IBM _SW	0.00%	8.61%	0.00%	0.00%	0.00%
Face_SW	0.00%	11.33%	0.00%	1.10%	-1.10%
MSFT_SW	0.00%	1.32%	0.00%	0.55%	-0.55%
FINLAND (N=197)	(N=4)	(N=80)	(N=3)	(N=110)	
Adience_FL	75.00%	36.49%	0.00%	0.00%	75.00%
IBM _FL	50.00%	2.50%	0.00%	0.00%	50.00%
Face_FL	0.00%	8.75%	0.00%	1.82%	-1.82%
MSFT_FL	0.00%	3.75%	0.00%	0.00%	0.00%
ICELAND (N=63)	(N=0)	(N=30)	(N=0)	(N=33)	
Adience_IL	Not represented	33.33%	Not represented	0.00%	Only Lighter Skin
IBM _IL	Not represented	0.00%	Not represented	0.00%	Only Lighter Skin
Face_IL	Not represented	10.00%	Not represented	0.00%	Only Lighter Skin
MSFT_IL	Not represented	0.00%	Not represented	0.00%	Only Lighter Skin

Iceland has no representation of darker females or males. Rwanda and Senegal have no representation of lighter females or males. The country breakdown of results underscores the importance of representation for evaluation. For example in Sweden where only 11 of 349 parliamentarians are darker women, all algorithms perform flawlessly on this intersectional subgroup. If one were to address under representation by oversampling the darker females from the Sweden subset, a false notion of accuracy would be established.

In an era where algorithms can be trained on 100s of millions of images, it may seem attending to issues of representation are of less concern. (Surely 100s of millions of images have enough representation so that oversampling is feasible at least on the training side). Benchmarks, however, are substantially smaller. For example the Labeled Faces in the Wild (LFW) benchmark that has served as the gold standard for facial recognition contains 13,233 images. As documented in Chapter 5, the benchmark has significant gender and ethnic skews. In addition, the latest NIST benchmark that is the United States national benchmark to assess algorithmic performance on the task of facial verification is only composed of 4.4% darker females, which amounts to 22 unique subjects. We need to have better standards. By creating more inclusive benchmarks with attention to gender and skin type we can (1) prove the suitability of using an algorithm for high stakes decision making and (2) catch systematic errors early.

Not attending to subgroup accuracy can no longer be an option. The evaluation of commercial gender classifiers provides evidence to show accuracy can differ substantially between intersectional subgroups. In the best case darker females were at least 32 times more like to be misclassified than lighter males. In the worst-case darker females were at least 136 times more likely to be misclassified than lighter males. When we put these numbers next to the 4/5th threshold recommendation for assessing disparate impact, we see that compared to lighter males darker females are at greater risk to experience harm from misclassification.

Though this intersectional subgroup evaluation focuses on gender classification, other automated facial analysis tasks like facial detection, facial identification, and facial verification should be assessed using intersectional subgroups.

6.8 Error Rates and Image Quality:South African Case

TABLE 28. SOUTH AFRICA - GENDER AND SKIN TYPE CLASSIFICATION ERROR RATES

	Total	Female	Male	Darker Skin	Lighter Skin
Adience_SA	30.43%	44.20%	20.70%	35.53%	10.23%
IBM_SA	15.00%	28.81%	4.94%	18.24%	1.25%
Face_SA	13.26%	31.64%	0.40%	16.33%	1.15%
MSFT_SA	8.29%	20.22%	0.00%	10.40%	0.00%

I 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 parliament images tend to be of higher resolution with less pose variation when compared with the images from African parliaments. The South African parliament, however, has comparable image resolution and has the largest skin type representation of all the parliaments. Members with lighter skin make up 20.14% (n=88) of the images, and members with darker skin make up the remaining 79.86% (n=349) of images. Table 28 shows that all algorithms perform worse on female and darker skin subgroups when compared to their counterpart male and lighter skin subgroups. The Microsoft gender classifier performs the best, with zero errors on the male and lighter skin subgroups.

TABLE 29. SOUTH AFRICA - INTERSECTIONAL SKIN TYPE AND GENDER ERROR RATES

SOUTH AFRICA (N=437)	Darker Female (N=157)	Lighter Female (N=24)	Darker Male (N=192)	Lighter Male (N=64)
Adience_SA	45.86%	33.33%	27.08%	1.56%
IBM_SA	33.12%	0.00%	5.91%	1.75%
Face_SA	35.95%	4.17%	0.53%	0.00%
MSFT_SA	23.38%	0.00%	0.00%	0.00%

As seen in Table 29 all the error for Microsoft comes from misclassifying images of females with darker skin. The table also shows that all algorithms perform worse on the women of color in the dataset. The Adience algorithm fails at a rate of 45.86% for this subgroup, which is 15.43% higher than its error for the entire dataset. All the commercial algorithms (IBM, FACE++, MSFT) have at least double the error rates for darker skinned females when compared to the overall error rates. On average, classification of lighter skinned males contributes the least to overall error rates.

TABLE 30. PPB - INTERSECTIONAL SKIN TYPE AND GENDER ERROR RATES

ALL (N=1270)	Darker Female (N=281)	Lighter Female (N=285)	Darker Male (N=315)	Lighter Male (N=389)
Adience_ALL	36.79%	31.34%	27.30%	0.27%
IBM _ALL	34.42%	5.28%	12.17%	0.26%
Face_ALL	33.58%	9.86%	0.69%	1.03%
MSFT_ALL	19.64%	1.75%	1.68%	0.26%

Examining algorithmic performance on the South African subset of PPB (Table 29) reveals trends that closely match the algorithmic performance on the entire dataset (Table 30). Thus, I conclude variation in performance due to the image characteristics of each country subset does not fully account for the differences in error rates between intersectional subgroups. In other words, the presence of more darker-skinned individuals is a better explanation for error rates than a deviation in how images of parliamentarians are composed and produced.

6.9 Subgroup Contribution to Aggregate Error Rates

In addition to using the balanced accuracy measure, reporting subgroup contribution to overall error will help attend to systematic errors.

Table 31 and 32 show how much each subgroup contributes to the overall error rate or each gender classifier.

TABLE 31. ERROR CONTRIBUTION BY GENDER AND SKIN TYPE

ALL (N=1270)	Female (N=566)	Male (N=704)	Darker (N=596)	Lighter (N=674)
Adience_ALL	68.25%	31.75%	68.98%	31.02%
IBM _ALL	74.32%	25.68%	89.19%	10.81%
Face_ALL	95.16%	4.84%	74.19%	25.81%
MSFT_ALL	90.77%	9.23%	90.77%	9.23%

For the commercial gender classifiers, around three quarters or more of the overall error comes from the misclassification of female faces. Likewise with skin type, three quarters or more of the total errors comes from the misclassification of darker skin. With the noncommercial Adience classifier, over two thirds of the overall error come from misclassifying females when looking at gender and darker skin when looking at skin type. Face++ has the largest female contribution to error at 95.16% and Microsoft has the highest dark skin contribution to error at 90.77%.

TABLE 32. ERROR CONTRIBUTION BY INTERSECTIONAL SUBGROUP

ALL (N=1270)	Darker Female (N=281)	Lighter Female (N=285)	Darker Male (N=315)	Lighter Male N=(389)
Adience_ALL	37.59%	30.66%	31.39%	0.36%
IBM _ALL	64.19%	10.14%	25.00%	0.68%
Face_ALL	72.58%	22.58%	1.61%	3.23%
MSFT_ALL	83.08%	7.69%	7.69%	1.54%

In the intersectional subgroup breakdown, we see across the board that darker females account for the largest proportion of misclassifications for all algorithms. Even though darker females make up 22.13% of the benchmark, they constitute between 64.19% to 83.08% of error for the commercial classifiers. White males who make up 30.63% of the benchmark contribute only 0.68% to 3.23% of the total errors from these classifiers. With the Adience classifier, error is more evenly distributed between darker females (37.59%), lighter females (30.66%), and darker males (31.39%), Lighter males contribute to only 0.36% of the Adience error. Considering that the Adience dataset is estimated to be 86.24% White, the low error rate on lighter skin seems to be reflective of dataset representation. Notably, even though it is estimated that lighter females (44.46%) and lighter males (41.61%) make up a roughly equal proportion of the Adience dataset, the lighter female group contributes 30.66% of the error whereas lighter males contribute less than 1%. These results suggest representation in training may not be the only factor to consider when composing new training datasets. The algorithms as well as datasets need to be reexamining and reconstructed to improve classification performance. Nonetheless, balanced benchmark datasets enable better assessment of subgroup performance regardless of how algorithms were trained.

7. Discussion

7.1 Overview

"To study algorithmic fairness is to study the interactions between different spaces that make up the decision pipeline for a task" - Sorelle Friedler (2016).

I look at gender classification through the construct space, observation space and decision space to uncover key assumptions and limitations that need to be overcome to achieve a richer understanding of algorithmic fairness as it relates to the accuracy of verifiable traits like gender. The discussion of the intersectional benchmark results presented in Chapter 6 will be guided by an analysis of each space in regard to the performance of four gender classifiers. The Pilot Parliament Benchmark (PPB) was created to enable intersectional benchmarking with attention to gender as well as skin type.

Key Findings

- Aggregate Selection Rate analysis using the 4/5ths threshold does not reveal statistical evidence for disparate impact based on gender or race
- Subgroup Selection Rate analysis using the 4/5ths threshold reveals statistical evidence for racial and gender discrimination for darker females
- Absolute Accuracy analysis reveals instances of gender, racial, and intersectional disparate impact when using the 4/5ths disparate impact threshold or 95% accuracy minimum
- Existing legal frameworks are not well equipped to deal with phenotypic or intersectional discrimination
- The PPB intersectional evaluation reveals systematic subgroup error

The notable accuracy gaps in relation to gendered skin type and intersectional subgroups show (1) the importance of attending to phenotypic attributes when assessing the performance of automated facial analysis and (2) the utility of intersectional benchmarking in revealing systematic error.

7.2 Construct Space: Gender Classification as Gender Presentation

For classification algorithms, defining the construct of interest is fundamental for establishing the related classes of that construct. The gender classifiers evaluated in this work all assume a binary construction of gender that is classified as either male or female. There is an underlying assumption that gender is fixed and can be known based on the characteristics of a face. For computer vision algorithms, the terms gender and sex estimation are often used interchangeably with the binary labels of male and female as opposed to the labels man and women which more explicitly deal with gender identity or the labels masculine or feminine which deal with degrees of gender expression. We need to examine precisely what these algorithms attempt to estimate. By construct, a classifier that attempts to estimate assigned sex based on biological features is not the same as a gender classifier that attempts to estimate gender identity based on gender norms. Geometric based gender estimation approaches, which have been less successful, are based on the assumption that facial geometries vary in discriminant ways between the sexes. Appearance based methods take in implicit gender cues, which are not limited to facial geometries. Here classification is not solely based on biological sex as determined by chromosomes nor is it based on the individual's self-identity. Instead, gender is shaped by socially, culturally, and historically influenced manifestations of gender display (Goffman, 1979).

Deep learning algorithms inextricably link the assessment of gender to the visual gender display norms inferred from a training set of images. Since gender display norms are culturally influenced, curating diverse training data must be given careful attention. This means attributes like hairstyle, accessories like nose rings, or other characteristics that have strong cultural gender norms, which can change based on a population of interest, need to be considered in data gathering and benchmarking.

The binary construction of gender also erases individuals who do not fit into socially constructed male or female gender norms. Cultures around the world recognize genders that transcend a binary view of gender. Indigenous American and Siberian tribes recognize "two-spirit" individuals (Jacobs, 2005). Increasingly, transgender individuals are gaining increased legal recognition. In 2014, the Supreme Court of India officially recognized hijras who do not fit neatly into a male and female gender binary as belonging to a third sex. A person's assigned biological sex may not match the individual gender identity (self-concept) or gender expression (behavior) that can change over time. Regardless of assigned sex, gender nonconforming individuals may intentionally take on androgynous modes of gender display. The naive binary gender classification that is used today lags behind expanding and centuries old understandings of gender. The gender displays of groups that are absent or underrepresented in training data and binary labels will not be well modeled. Even when gender display is somewhat binary within a culture, across cultures display norms may differ. For example, if it is customary for women in one population to wear short hairstyles, but a model has been trained to associate short hair with men, this variable gender display norm could lead to inaccurate classification of new female images that break the learned norm.

By construct, binary gender classification models based on deep learning predict gender based on gender display. Beyond appearance based models gender classification, there are other models that use facial geometry to determine gender or sex based on difference between male and female face geometries influenced by sex hormones (Burton et al., 1993) Geometric methods have not been as successful as appearance-based models suggesting that factors beyond facial geometry signal gender. Even when facial geometry is being explicitly measured, phenotypic gender difference between populations should be accounted for by including sufficient male and female training samples of each target subgroup.

In keeping with the appearance-based approach, the convolutional neural network (CNN) based Adience gender classifier reflects state-of-the-art methods for using deep learning for computer vision. The exact implementations of the commercial gender classifiers are proprietary, but given the companies' investments in deep learning, it is likely they are created with CNNs. The explicit view of gender as a binary construct and the implicit use of gender display to estimate gender make gender classification models especially susceptible to stereotypes of visual masculinity and femininity present in training data. By conceptualizing appearance based gender classification as an imprecise exercise of inferring gender display norms and not estimating biological sex, machine learning practitioners can learn from considerable scholarship that has explored how gender display manifests in different cultures. Insights on cross-cultural gender display norms can inform the creation of more inclusive training sets and benchmarks.

7.3 Observation Space

Understanding that gender is socially constructed, we seek an observation space that is used as a proxy to estimate the gender construct. The observation space is composed of features —attributes that provide useful information to inform classification. Discussions of algorithmic fairness have mainly centered on predictive models that use explicitly defined features to determine class membership. Observable features for predictive models are chosen as proxies for the unobservable construct of interest such as creditworthiness. Creditworthiness can be assessed using a set of features, including income, occupation, and age. By United States law, a set of protected attributes like gender cannot be legally used to determine access to credit. Feature selection introduces the possibility for bias and the possibility of redundant encodings. A redundant encoding is a feature that is highly correlated to a protected variable like ethnicity. Zip codes are highly correlated to socioeconomic status, and can also encode attributes like race. Redlining practices of the past that limited access to credit, insurance, and loans based on the zip codes of individuals have not been eliminated by algorithmic automation. Training data that reflects redlining in the past in the real world can embed racial bias into virtual predictive models even if no race labels are used. Because of the possibility of redundant encodings, Hardt, Price, and Srebro argues it is best to explicitly attend to protected features like gender and ethnicity when attempting to create fair algorithms (2016). By explicitly attending to difference in predictions between different subgroups, various techniques can actively be used to curb discrimination.

7.3.1 Visual Redundant Encodings

To date, redundant encodings have been defined by intentional feature selection. As a result, algorithm designers can choose to exclude features known to redundantly include protected data. For gender classification using convolutional neural networks, features are not explicitly understood variables. Instead, discriminative features are inferred based on the images provided in training data as exemplars of classes of interest, like male or female presented faces. Though the observation space for gender classification models is visual, unlike the predictive models discussed, there is still the possibility for redundant encodings. An image of an individual can contain cues that are highly correlated to perceived gender, perceived ethnicity, or perceived religion. Even if no categorical information about gender, ethnicity, or religion is available, these attributes can still be redundantly encoded. Human bias that leads to redundant encodings through feature selection in the case of predictive models can be introduced through bias in training data selection for vision-based models. Studies that attempt to use face images to infer intrinsic qualities about individuals such as trustworthiness or criminality (see the controversial 2016 Wu and Zhang study on inferring criminality from faces), risk perpetuating stereotypes that can be a reflection of external factors like ethnic profiling more than intrinsic factors. Visual redundant encodings like categorical redundant encodings can reflect "the prejudice of prior decision makers" as Barocas and Selbst would say (2014). In human-focused computer vision classification, diverse phenotypic representation is needed for each class of interest to mitigate visual redundant encodings. Without care, visual redundant encodings can give rise to digital phrenology that propagates prejudice under the guise of objective analysis.

7.3.2 Phenotype Awareness

In the spirit of fairness through awareness, explicitly attending to the face phenotype can help mitigate demographic bias that is based on phenotype correlations. The finding that the gender classification algorithms perform the worst on females with darker skin provides evidence that the algorithms can systematically fail on demographic groups whose class association is in part influenced by perceptions of darker skin. The finding that classification is substantially worse for darker females than darker males also shows that intersectionality – the intersection of multiple identities (gender and perceived ethnicity based on skin type) – cannot be ignored. Beyond gender classification, if an automated facial analysis system is tested to work well on individuals with darker skin, but the majority of those individuals are male, it cannot be assumed the system will also work well on female faces and vice versa. The current gender imbalance of darker skin in the National Benchmark for facial detection and facial recognition is cause for concern when it comes to assessing the accuracy of algorithms.

7.3.3 Relating Phenotype and Demographics

The overlap of phenotypic and demographic groups is not one to one, but the impact of skin type on ethnic and racial perception should not be underestimated. In the US, individuals with type VI skin would likely be perceived as Black. Individuals with type IV skin may belong to a wide range of ethnic groups but would not be considered White. Failure rates on darker skin as defined by the Fitzpatrick skin type can serve as a proxy for failure rates on non-White individuals using US racial classification norms. Canada uses the term visible minority to underscore that certain phenotypic characteristics make membership to a demographic group more explicit. Here darker skin can be viewed as a visual redundant encoding for belonging to the visible minority in the Canadian context.

Skin type of course is not the only phenotypic attribute that can be assessed. Prior research has explored facial regions that impact gender classification when using appearance-based approaches like principal component analysis. By masking facial regions such as the forehead, eyebrows, eyes, nose, lip, and chin, Özbudak et al. produced experimental results that show the nose was the most influential region for gender classification followed by the forehead. The chin was the least influential. They claim their study to be the first to test for the influence of racial features on gender classification.

However, their results do not report how many of the 480 images they used were categorized into Asian, European, or African faces. The images were reportedly taken from the FERET gray scale database, but the demographic breakdown of the images is not reported thus making it difficult to assess the statistical strength of the results that Asian faces had 0% error and African faces had 13% error. Furthermore, the gender breakdown of the error is not reported so it is unclear if misclassification is equally distributed across gender or predominantly due to either male or female error. These uncertainties call for further scholarship that attends to the impact of phenotypic characteristic on gender classification that extends beyond skin type. For some populations, skin type may be correlated with other phenotypic characteristics. Future analysis should examine the influence of different facial regions on gender classification.

7.3.4 Skin Type is an Imprecise Ethnic Proxy

Still, it should be noted that skin type alone is not an adequate proxy for ethnic or racial classification. For a given skin type, multiple ethnicities can be associated. The following individuals hailing from four different regions could be classified as Type IV on the Fitzpatrick scale though ethnicities differ (see Figure 17).

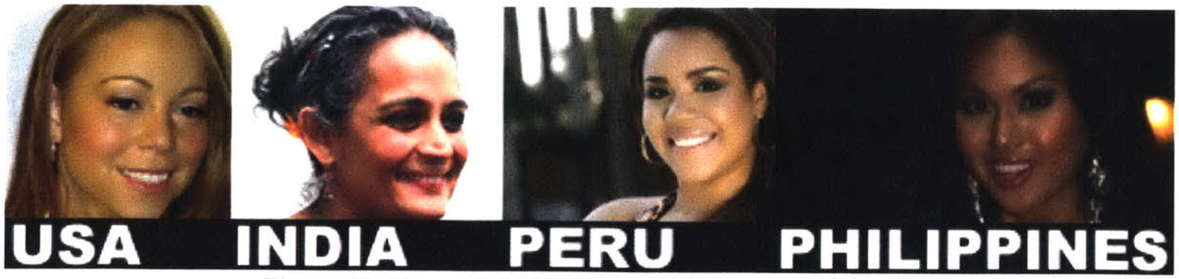


Figure 17. Fitzpatrick Skin Type IV Across Four Regions

Even though people perceived as belonging to specific ethnicities have skin types associated with these perceptions, skin type alone does not determine ethnicity nor does ethnicity necessarily define skin type. Sandra Laing - a south African woman born to Afrikaner parents with a traceable legacy of three generation of Afrikaner heritage was labeled "coloured" by the Apartheid South African government because of her hair texture and skin color. Self-perception is also variable. Public figures like Tiger Woods who is described as African-America by the media, self-identify in ways that embrace a multi-ethnic heritage.

In the US, hypodescent also known as the "one-drop rule" was used to erase multifaceted ethnic identities. Depending on the state, if an individual had 1/32 to 1/4 or more African ancestry, laws legalizing Black/White segregation deemed the individual to be Black discounting the other portion of ancestry.

Hypodescent laws in part account for the large intraclass variation of people perceived and self-identifying as Black or African-American in the United States. The following images of public figures who identify or are perceived as Black in the United Stated show this range of intraclass variation (see Figure 18).

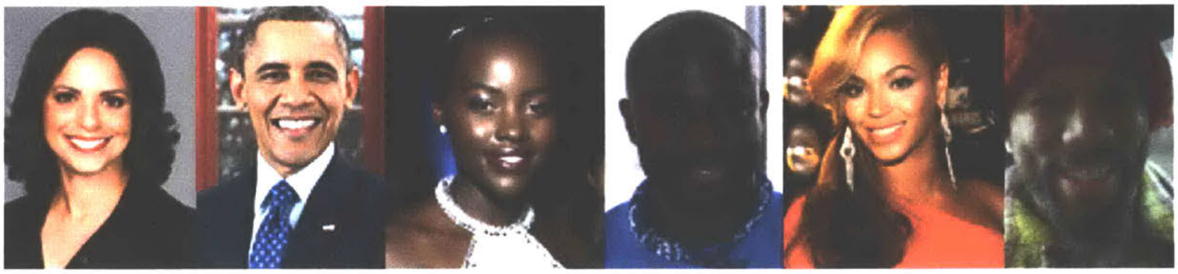


Figure 18. Individuals Perceived of Self-Identifying as Black/African-American in United States

Skin type is a limited proxy for ethnicity, and ethnicity is an unstable predictor of skin type. This is not to say there can be no correlation between the two. Given interclass variation in regard to phenotypes associated with an ethnicity, assessing phenotype directly is more useful than using demographic proxies when we want to evaluate how specific facial characteristics influence classification accuracy. Most critically, when attempting to create inclusive benchmarks, we need to account for intraclass variation within demographic groups.

Since the highest failure rates were on people of African descent in this study, I conclude this section with skin type implication for the Black demographic in the United States. Further work is needed to assess intraclass variation in other demographic groups that have been created around the world based on political, social, and cultural factors. The Fitzpatrick scale has 3 categories for classifying White skin and 3 categories for people with non-White skin. The National Survey of Black Americans (Jackson & Neighbors, 1997) has a 5-point system for assessing skin color of African-Americans to provide more nuance than the Fitzpatrick scale that was developed to measure skin response to UV radiation. With intraclass variation in mind, a benchmark that only included light skinned persons who may be more likely to be celebrities with readily accessible images would not adequately represent the phenotypic variation of the Black population in the United States.

The phenotypic characteristics of public figures who self-identify or are classified in a demographic group may not always be representative of intragroup variation. The NIST strategy of using public figures to collect seemingly representative samples must also consider the way in which the privileging of lighter skin can skew the phenotypic representation of those positioned to be public figures. In this study, the Pilot Parliaments Benchmark exemplifies how visible representatives of populations do not always reflect the phenotypic characteristics or distribution of a population. Even though White South

Africans account for 8.9% percent of the population, phenotypically lighter skin accounted for 20% of South African[20] elected or appointed officials in PPB.

Some countries with colonial histories also have an overrepresentation of lighter skinned politicians that do not phenotypically represent the majority of the population. Lighter skin overrepresentation, beyond politics, can also impact the reliability of using countries as ethnic proxies. As touched on in section 2.5.2, the NIST Gender report used country of origin as a proxy for ethnicity to determine the gender classification accuracy across ethnicities. The visa images of visitors from different countries around the world to the United States comprised the datasets. This approach is limited given that (1) ethnicity is an unstable proxy for phenotype characteristics, (2) individuals with the ability to obtain visas may not be phenotypically representative of the population, and (3) countries with multiethnic populations like Brazil are ill suited for mono-ethnic approximation.

To increase algorithmic transparency in the future, explicit attention must be given to phenotype characteristics when assessing the performance of automated facial analysis algorithms. Assessing performance across skin types using the Fitzpatrick scale is a minimum starting point. Disparities in performance between skin types can be used by analysts to determine the demographic groups defined in a given country that are most at risk for inaccurate classification. In this this, intersectional benchmarking revealed that females with darker skin who would fit into the African-American/ Black race category of the US could have a high risk for misclassification by the gender classifiers tested. Still, it cannot be assumed that all females of darker skin coming from different populations around the world will be classified in the same way. These results show that further assessments are needed before confidently deploying automated gender classifiers on multiethnic populations. Classification accuracy must be contextualized to fit the target population on which technology will be used. Fairness awareness in the context of computer vision requires phenotype awareness along with an understanding of the historical, political, and social factors that shape demographic distinctions.

7.3.5 Data Quality and Sensors

In the observation space, bias can arise due to under or over representation in data observed. The quality of available data on specific groups can also be limited. In the case of computer vision, in addition to assuring adequate diversity in training data, the quality of data is directly associated with quality of sensor readings that produce the original digital image in a training set.

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. Illumination is of particular importance

[20] 2011 South Africa Census is available at https://www.statssa.gov.za/publications/P03014/P030142011.pdf

when doing an evaluation based on skin type. Default camera settings are often optimized to expose lighter skin better than darker skin. Underexposed or overexposed images that present significant information loss or a lack contrast can make accurate classification challenging.

With full awareness of the challenges with pose and illumination, I intentionally chose an optimistic sample of constrained images that were taken from the parliamentarian websites. Each country subset had its peculiarities. Images from Rwanda and Senegal had more pose and illumination variation than the images from the other counties. The Swedish parliamentarians all had photos that were taken with a shadow on the face. Of all the subsets, the South African subset had the most consistent pose and illumination for the images. The South African subset also was composed of a substantial number of lighter skinned and darker skinned subjects. Given the diversity of the training set, the high image resolution, and the consistency of illumination and pose, the 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.

7.4 Decision Space

In machine learning, the decision space is the central focus. Perfecting prediction or classification even if the mechanism for either is not fully understand remains the priority. While debates about transparency and accountability in AI tend to question constructs, feature selection, or data composition, debates on fairness AI focus on the outcome of algorithmic prediction or classification. To expand the conversation about fairness we cannot concern ourselves only with the decision space without also questioning how decisions are reached and who determines how they are used. Regardless of algorithmic accuracy, larger ethical questions remain about whether or not gender classification should be used in the first place and to what extent those impacted by classifiers are informed of their use and have the agency to opt out. Chapter 2 explores potential misuses of gender classification, which enhances covert surveillance and introduces the potential for gender discrimination in face-based target advertising. This section examines criteria to evaluate the readiness of gender classification from a technical perspective. We should remain skeptical about the appropriateness of using gender classification and explore pathways to engage the public in governing the use of automated facial analysis.

Since bias is a social and technical concern, being able to explain how different attributes --be they categorical or visual redundant encodings--impact algorithmic decisions provides a map into how to mitigate unwanted bias. Enthusiasm over improved accuracy on benchmarks once deemed challenging has overshadowed analysis of error rates that can inform questions surrounding algorithmic fairness.

For gender classification, I define fairness as having comparable accuracy and error rates between subgroups. In addition to looking at overall accuracy on the Pilot Parliaments benchmark, I explore error rates with an attention to demographic, phenotypic, and intersectional subgroups. The purpose of this exploration is to examine the decision space in such a way that it can offer answers about factors that contribute to gender misclassification. Prediction and explanation are both vitally important if we want to identify and mitigate subgroup bias in classification accuracy.

7.4.1 From Aggregate Results to Intersectional Analysis

The results of the overall gender classification accuracy show the obfuscating nature of single performance metrics. Taken at face value, the accuracy of classifiers ranging from 77.81% to 94.78% on the PPB, suggests that some classifiers are suitable to use on the entire population 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 break down of the results show that performance differs substantially for distinct subgroups. Classification is 10 – 21% worse on female than males and 9 – 20% worse on darker skinned than lighter skinned subjects.

Though helpful in seeing systematic error, gender analysis and skin 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 females, lighter females, darker males, and lighter male subgroups provides more answers. Across the board darker females constitute the majority of misclassification for all gender classifiers ranging from 64.19 – 83.08% for commercial classifiers. Lighter females contribute to 7.69% to 22.5% of the misclassification for commercial classifiers. Darker males contribute 1.61% to 25% of these classifications. Lighter males contribute 0.68%to 3.23% of misclassifications. We can see that the most improvement is needed on darker females specifically, and more broadly speaking the 10 – 21% gap between male and female classification should be closed. When examining the gap in lighter skin and darker skin classification, we see that even though darker females are most impacted, darker males are still more misclassified than lighter males.
These results raise more questions.

In this thesis, I consider gender classification, but what differences might subgroup error analysis reveal in other automated facial analysis tasks? A benchmark dataset that underrepresents darker females which is true of existing benchmarks would not be suitable for finding this kind of disparity. What steps can be taken to create more inclusive benchmarks to uncover subgroup disparities, and what steps can be taken to mitigate the disparities that are uncovered? Most critical to questions around fairness, how do we establish permissible accuracy thresholds for fairness. Appropriate thresholds will be context specific. Next, I look at ways we can construe fairness for binary gender classification.

Are differences in subgroup classification error rates grounds for disparate impact claims?

The gender, skin type, and intersectional error analyses (see Chapter 6) show subgroup differences with gender classification accuracy. By using different measures for discrimination that have been defined in legal literature, it is possible to examine these results through selection rate analysis and the 4/5ths threshold.

In the United States selection rate is used to measure disparate impact. In the context of gender classification, I define discrimination in terms of a gender classifier having accuracy differences between two groups that exceed the disparate impact threshold. Here the comparison group will be the intersectional subgroup that has the highest accuracy deemed A_best and another subgroup deemed A_worst.

Selection Rate: $\frac{1-p1}{1-p2} = \frac{1-Error\_worst}{1-Error\_best} = \frac{A\_worst}{A\_best}$

Table 33 shows the selection rates between the best-classified group and the worst classified group.

TABLE 33. SUPBGROUP SELECTION RATES

	Selection Rate			
ALL	Female/Male	Darker/Lighter	D.Fem/L.Male	D.Fem/D.Male
Adience_ALL	84.66%	78.69%	63%	86.95%
IBM _ALL	91.19%	79.14%	66%	74.67%
Face_ALL	87.44%	87.69%	67%	66.88%
MSFT_ALL	94.39%	90.49%	81%	81.74%

If we use the 4/5ths precedent to assess discrimination, we reach the following conclusions. The selection rate between males and females is above 80% for all classifiers and does not meet the 4/5ths threshold. The selection rate between lighter and darker skin is on the border of the discrimination threshold for the Adience (78.69%) and IBM (79.14%) classifier. When we look at the selection rate between the worst classified group which is darker females across the board and the best classified group which is lighter males for the Adience, IBM, and Microsoft classifiers and darker males for the Face++, we have 3 instances that surpass the discrimination threshold. The selection rate between darker females and lighter males for the Adience (63%), and IBM(66%) classifier and the selection rate between darker females and darker males for the Face++ (66.88%) are all well below the 80% mark. The Microsoft classifier is on the border with an 81% selection rate between darker females and lighter males.

Can we conclude there is no gender discrimination even though darker women are significantly misclassified? Can we conclude there is no potential for race discrimination since darker skin misclassification is on the border or above the 80% threshold? Using the current threshold and data, there is not enough statistical evidence to make a claim for gender discrimination in the aggregate or enough evidence to make a strong racial discrimination claim using darker skin as a proxy for non-White. Even though there is statistical evidence that shows darker females are misclassified at a rate that well exceeds the discrimination threshold, we lack legal frameworks that address intersectional discrimination. The 4/5ths rule applies to protect classes which includes females and Blacks, but it does not include explicitly include Black females.

Beyond the legal implications, selection rate analysis faces two key limitations. One, even if the rate between groups is within a suitable threshold, the overall classification rates may not be acceptable for the task. For example, if gender classification for the best-classified group in at 70% and the worst classified group is at 68%, the selection rate of 97.14% may be acceptable. However, the suitability for use in a high stakes scenario is questionable given the 30 and 32% error rates on the groups respectively. Two, the use of selection rate prioritizes group accuracy over individual accuracy. In a case where exceptional accuracy is achieved for one subgroup and acceptable accuracy is achieved for another, using selection rate analysis to prohibit automated classification may limit benefits for individuals in a subgroup that will be correctly classified.

Should a member of the worst performing group who would be correctly classified not receive the benefits of classification? Hypothetically, the benefit of accurate classification could mean the ability to use an automated system that saves time and reduces costs. Even if the system fails on 1 out of 5 individuals in a disadvantaged subgroup, 4 out of 5 individuals in the subgroup still receive the benefit of efficiency.

Beyond the perceived benefit of efficiency, how might we factor in that the efficiency introduced might have externalities relating to security and privacy that are not factored in? These open-ended questions provide reasons to reevaluate how we assess the relation between aggregate accuracy, subgroup error rates, and algorithmic fairness.
The selection rate can be used to assess relative differences for assessing fairness in opportunity. Comparing the selection rate for men and women when it comes to receiving a positive credit rating can be helpful for analyzing gender fairness. Using selection rate to assess fairness of the classification accuracy for verifiable traits like gender is ill posed. Relative performance is not as important as absolute performance. The suitability of a verifiable classifier should be determined by ensuring that the accuracy of all subgroups of interest is above an absolute threshold and not a ratio-based threshold. As a thought experiment, let us require the minimum threshold of accuracy to be 80% for all subgroups of interest.

TABLE 34. AGGREGATE GENDER AND SKIN TYPE ACCURACY

	Accuracy			
ALL	Female	Male	Darker	Lighter
Adience_ALL	65.88%	87.34%	68.24%	86.72%
IBM _ALL	80.36%	94.46%	77.24%	97.60%
Face_ALL	78.62%	99.12%	83.51%	95.24%
MSFT_ALL	89.46%	99.13%	89.69%	99.11%

Using this criterion, we can see that if our subgroups are just demographic by gender, IBM and Microsoft pass the test since the accuracies for males and females are above 80%. If we only split the subgroup by skin type, Face++ and Microsoft pass the test. Using absolute performance and not relative performance we can use the 4/5ths threshold to state there is a statistical case for gender discrimination with the commercial Face++ classifier. There is also a case for racial discrimination with the IBM classifier. The Adience classifier has the poorest performance across gender and skin type, but it is not sold commercially. In addition, the Adience gender classification model can be retrained and improved before being used. Keep in mind 80% accuracy is a generous threshold for a binary classifier. A human parity standard which states automation technology is suitable for adoption when it matches or exceeds human performance on a task negates the use of the 4/5ths rule for gender classification. If we find human performance on estimated gender to be 95% and required absolute performance for all subgroups to be at 95% or higher, all the classifiers evaluated would be deemed to demonstrate both gender and racial discrimination.

TABLE 35. INTERSECTIONAL GENDER AND SKIN TYPE ACCURACY

	Accuracy			
ALL	Darker Female	Lighter Female	Darker Male	Lighter Male
Adience_ALL	63.21%	68.66%	72.70%	99.73%
IBM _ALL	65.58%	94.72%	87.83%	99.74%
Face_ALL	66.42%	90.14%	99.31%	98.97%
MSFT_ALL	80.36%	98.25%	98.32%	99.74%

If we evaluated intersectional subgroups using either the 80% or 95% thresholds, we see that none of the classifiers definitely pass the test mainly due to poor performance of classification on darker females. Microsoft comes the closes with a borderline 80.36% accuracy rate on darker females. The dividing lines matter. Establishing the subgroups of interest that must have accuracies that exceed the minimum threshold is a critical decision that impacts fairness evaluations. How might we establish which subgroups are most relevant to test? How do we establish appropriate accuracy thresholds?

7.4.2 Impact Population, Benchmark Data, and Training Data.

The subgroups of interest for a gender classifier should reflect the population that will most likely be impacted. The demographics and phenotypic composition of the impacted population should be reflected in benchmark datasets used to determine suitable use. Classifiers that learn on training sets that are inclusive of the impacted population will be best positioned to do well on population inclusive benchmarks.

If a gender classifier is being used for a mono-ethnic population with small intraclass phenotypic variation, the gender subgroups may be sufficient for determining acceptability of use. However, the reality is that automated facial analysis is being deployed on multiethnic populations at transportation hubs like airports and embedded in consumer products like smart phones that are sold in global markets. For companies like Microsoft and IBM that operate on a global scale, the mono-ethnic assumption is not credible or viable. Face++, which appears to focus mainly on China, reports their facial recognition software is embedded in Lenovo laptops that are sold globally by IBM. It is not uncommon for large corporations to use subcontractors to add specific functionality to products. Due to increased globalization, any company that is offering facial analysis for high stakes tasks like authentication or surveillance must ensure the technology works well across the sepia spectrum of human faces.

How then do we proceed with evaluating the accuracy of automated facial analysis on multiethnic populations? While census data offer demographic information for some populations down to precinct levels, phenotypic information is not readily available. We have also seen that phenotypes and ethnicities do not have a direct relationship. To address this question, let us return to the overarching goal of generalizability for facial classification. Developers of gender classifiers have the ultimate goal of creating classifiers that works well on all faces. One way of setting the standard to reach that goal is to work towards creating a globally representative benchmark. Admittedly, such a benchmark will be a work in progress since capturing the variation between 7 billion people will not happen all at once. Since classifiers can be used in a variety of contexts and population dynamics continue to change, developing a globally representative benchmark will give a better picture of the overall state of the art with automated facial analysis technology.

A global benchmark would include at minimum balanced gender representation across specific age brackets and phenotypic traits. The age brackets can follow the categories already set out by the National Institute of Standards and Technology. For phenotypic traits, at a minimum skin type can be factored in given the interaction of skin reflections, camera calibration settings, and illumination. Factoring other phenotypic factors like eye shape or nose/lip/chin ratios can help define additional components to increase diversity. Following from the use of 500 unique subjects in the NIST benchmark, let us pose that each intersection subgroup requires at least 500 unique subjects. So then there are 9 age subgroups split by decade, 6 skin types and 2 genders to result in 108 intersectional subgroups. With each subgroup containing 500 unique individuals, the total benchmark would need 54,000 images. The number can be reduced when we take into account the difficulty of distinguishing children under the age of 8 and that 3 of the Fitzpatrick skin types are applicable to White skin. With this reduction we now have 8 age groups, 4 skin types, and 2 genders resulting in a benchmark of 32,000 unique individuals.

Using a phenotypically inclusive benchmark as a starting point, the state of the art of automated facial analysis algorithms can be more rigorously assessed. Such a benchmark will need to be not only phenotypically inclusive but phenotypically balanced to enable meaningful subgroup error analysis. A geographically inclusive dataset like IJB-A should not be conflated with a phenotypically representative dataset. Including a few examples of subjects with phenotypes that differ from the majority is not enough to make a suitable benchmark. The Pilot Parliaments benchmark is a starting point towards making a benchmark that is phenotypically balanced in regard to skin type. Further work will be needed to include unique subjects that represent more phenotypic diversity around the world. At the very least, the Central and South American parliaments that are in the top 10 ranking for women's representation in parliaments can be included.

While performance on a global benchmark can be used to determine the technical suitability of using an algorithm on a target population, citizens themselves should have a voice in the use of these technologies in local jurisdictions. For example, following the notice-and-comments process used for environmental impact statements, citizens can weigh in on discriminatory impact assessments (Selbst, 2017). These impact assessments posed by Selbst in the context of predictive policing can be extended to automated facial analysis, which is increasingly adopted by law enforcement. For the case of automated facial analysis, discriminatory impact assessments should incorporate local benchmarks showing the accuracy on a sample of faces selected to be demographically representative of the jurisdiction of concern can serve as a complement to a global benchmark. By constructing a phenotypically diverse benchmark and requiring the reporting of accuracy, error rate, and contribution to overall error rate of each intersectional subgroup, we can increase transparency. By requiring that all subgroups perform at an agreed upon minimum accuracy we can increase fairness. Accountability will need to be established through a repeatable process where vendors are required to check the subgroup accuracy of their classifiers periodically on both the expanding global benchmark and local spot checks.

7.4.3 Revisiting the articulation of intersectionality

Though there are explicit laws against gender-based discrimination and race-based discrimination, laws around phenotypic discrimination are not defined. For people seeking redress on issues of colorism, defining intraclass discrimination proves challenging in the current legal landscape. To coincide with existing legal frameworks using skin type as a cue for race can enable phenotypic accuracy assessments to be used as evidence for potential disparate impacts. Even if color is associated with race, the intersection loophole remains. The selection rate analysis showed that by using the 4/5ths rule there was not a strong statistical case for gender or racial discrimination even though darker females who can be translated as non-White women were significantly more likely to be misclassified than lighter males, a group we can translate as White men. In 1990 Kimberle Crenshaw introduced the term "intersectionality" to address how intersecting identities can lead to outcomes that cannot be assessed in analytic silos. Her earlier 1989 analysis of discrimination cases brought forward concerning Black women showed the limitation of anti-discrimination law that treated sex and race discrimination separately. In a sex-based discrimination case brought against General Electric, since discrimination against Black women did not indicate discrimination against all women, the judge rejected the claim of sex-based discrimination. For a race-based discrimination class action lawsuit brought against Travenol, two Black women provided statistical evidence of *overall* race-based discrimination, but the defendant was able to limit back pay to just Black woman and exclude redress for Black men.

These cases highlight two dangers with a single-issue view of discrimination. In the former case, intraclass variation is overlooked to deny discrimination. Females are seen as a monolith, thus Black women should not be treated as a hybrid case. In the later case, intraclass similarity is overlooked to limit responsibility. Black women are not seen to represent Black people as a whole. A single-issue approach to discrimination marginalizes people who have intersectional issues and enables the abdication of intraclass responsibility. Intersectionality is a necessary analytical frame that can help identify issues of equity would otherwise be dismissed. Going back to gender classification, looking at the performance on females alone or darker-skinned individual alone is not enough to identify the potential result in disparate impact for women of color. The results of the Pilot Parliaments Benchmark show the importance of applying an intersectional analytic frame to gender classification in particular and automated systems that make determinations about individuals in general.

7.5 Synthesis

In examining the construct space, observation space, and decision space for gender classification, I show the need to revisit how we approach definitions of gender classification, how we configure training datasets, and how we analyze algorithmic performance on gender classification. I argue gender classification in computer vision is an exercise of inferring gender display norms. Since these norms vary from culture to culture, it is of critical importance to create datasets that are informed by a nuanced understanding of gender identity, gender expression, and their relationships to gender display. When gathering data to represent gender norms, I emphasize the importance of being aware of phenotypic differences like skin type and the complex relationship between skin type and perceived ethnicity. Finally, I look at how existing measures for discrimination prove inadequate for assessing the fairness of classifiers that estimate verifiable traits like gender instead of estimating an unknown future behavior like likelihood to default. I present intersectional subgroup analysis as a way to more rigorously assess gender classification across phenotypic and demographic categories. As automation becomes increasingly embedded in decision-making, this thesis shows intersectional curation of training data and intersectional analysis of algorithmic accuracy will inform our understanding of algorithmic fairness.

8. Conclusion

Future Work

This thesis focused on the diversity of benchmark datasets and the performance of gender classification algorithms in regard to gender and skin type. Future work is needed to advance scholarship on dataset representation and intersectional evaluation of algorithms not limited to gender classification. More face datasets should be assessed for gender and skin type representation as well as other demographic and phenotypic factors like age and eye shape respectively. The Pilot Parliaments Benchmark dataset only focused on individuals from European and African countries. Larger phenotypic representation is needed for the Americas, Asia, Australia, and Pacific Islands. For efficiency this thesis evaluated four gender classifiers. Gender classifiers made available from the research community like IMDB-WIKI and tech companies like Amazon, which embeds gender classification in its Rekognition services should be evaluated to see if the disparities uncovered in these results persist. Unsupervised learning techniques should also be employed for cluster analysis that can reveal novel subgroups inferred from the classification accuracies. Beyond gender classification, intersectional subgroup evaluation with attention to phenotype can be applied to all automated facial analysis tasks including face detection and facial recognition.

Summary

Advances in automated facial analysis have led to renewed enthusiasm about the potential for deep learning techniques to outperform humans on facial perceptual tasks like gender classification. Automated gender classification is increasingly used to customize product experiences and enable covert soft biometric surveillance. Leading technical companies operating in global markets now sell gender classification powered by deep learning breakthroughs. Law enforcement officials are increasingly employing automated facial recognition software that relies on analogous machine learning approaches used for gender classification. Increased adoption of gender classification suggests the technology has reached maturation, yet previous research and performance reports on the accuracy of these classifiers fail to rigorously address how phenotypic differences impact accuracy. There is also little research on how intersectional phenotypic and demographic factors influence classification. Since skin reflectance influences sensor readings and can prove challenging in low illumination conditions, I decided to evaluate skin type as a phenotypic attribute of concern.

To increase scholarship on classification accuracy in relation to skin type and gender, I first assessed the suitability of existing facial analysis benchmarks for this task. I evaluated the gender parity and skin type representation for the government administered IJB-A dataset and Adience benchmark developed to assess the state of the art of research gender classifiers. This work demonstrates that significant gender and phenotypic skews persist in influential datasets. Analysis revealed the government benchmark was composed of only 24.6% women. The distribution of skin type for this benchmark was heavily skewed to lighter skin (79.6%). I then looked at an intersectional breakdown of the benchmark in regard to gender and skin type distribution. The least represented group, darker-skinned women made up only 4.4% of the benchmark compared to the most represented group, lighter males, who made up 59.4% of the benchmark. For the Adience benchmark, gender parity was reached, but darker skin represented only 13.76% of the benchmark. Since the skin type representation was heavily skewed, I concluded neither benchmark would be suitable to serve as a tool to assess gender classification performance in an intersectional manner. Nonetheless, the process of assessing representation resulted in skin type annotations of unique subjects in the datasets that can be used for phenotypic evaluations in the future. To become more inclusive, the Adience and IJB-A benchmarks should be extended to increase phenotypic representation particularly of men and women with darker skin. Other influential datasets used in human-focused computer vision should be evaluated for phenotypic representation and extended to become more representative if needed.

Given the representational shortcomings of the Adience and IJB-A datasets, the Pilot Parliaments Benchmark (PPB) was created. The benchmark is composed of parliamentarians and appointed officials from three European countries and three African countries to balance for skin type. The benchmark contains 1270 unique individuals split into four intersectional subgroups: darker females, lighter females, darker males, and lighter males. This new benchmark can be extended by the research community for more in-depth intersectional analysis of gender classification.

Equipped with the newly constructed Pilot Parliaments Benchmark, I then selected four gender classification algorithms to evaluate: Adience, IBM, Microsoft, and Face++. The evaluation of the Adience, IBM, Face++, and Microsoft gender classification algorithms using the PPB, shows that these gender classifiers perform better on male faces than female faces (9 -10%) and perform better on lighter skin than darker skin by (10 - 21%). All classifiers perform the worst on the darker female faces, and the majority perform the best on lighter male faces. The difference between performance on the best-classified group and the worst classified group is as much as 36%. Based on these results, I replicate the established finding that gender classification tends to perform better on male faces than on female faces (Ngan & Grother, 2015). To my knowledge this is the first study that looks at the impact of phenotypic and demographic attributes on gender classification accuracy. The performance results indicate that only assessing accuracy by gender or skin type will not reveal important subgroup systematic error.

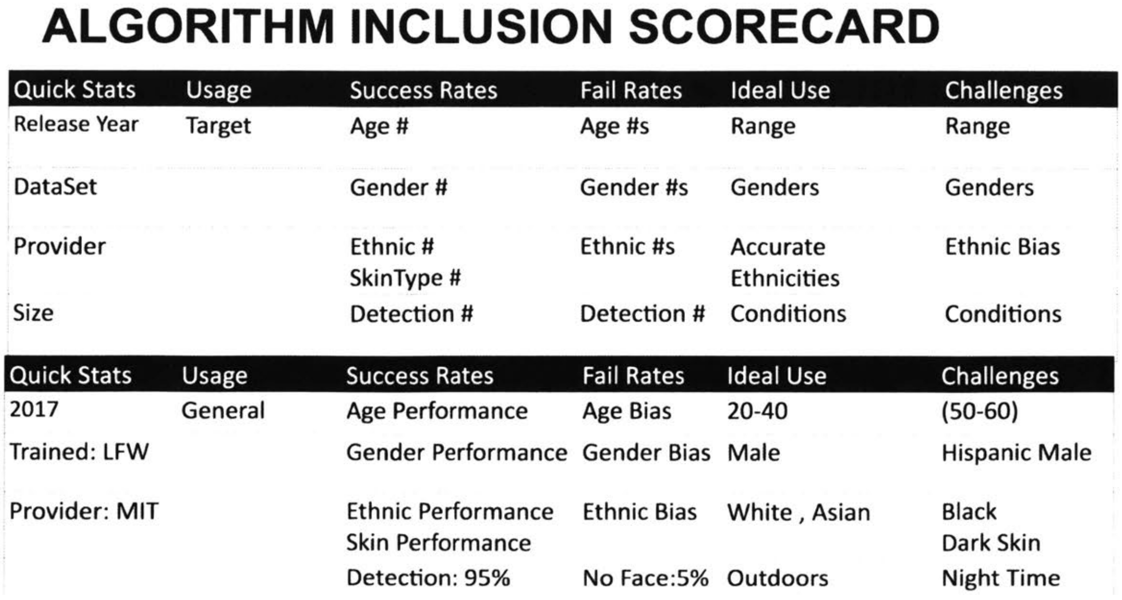
Intersectional subgroup analysis revealed that for the best performing classifier, darker females are 32 times more likely to be misclassified than lighter males on a benchmark that is has an overall accuracy of 94.78% on the fairly balanced Pilot Parliaments Benchmark. Subgroup error analysis showed that of the misclassified, darker females contributed 37 - 83% of the error rates, lighter females contributed 8 - 30% of the error rates, darker males contributed 2 - 31% of the error rates, and lighter males contributed .4 - 3% of the error rates. The disparities revealed in these error rates show the utility of explicitly checking for intersectional subgroup performance. Even on a fairly balanced benchmark, systematic subgroup failure can be obfuscated by aggregate accuracy numbers.

Finally, I analyzed the difference in classification accuracy between subgroups (darker females, lighter females, darker males, lighter males). Subgroup accuracies were compared to see the applicability of using legal measures for discrimination in gender classification that have been used in legal cases to establish disparate impact in the United States. Existing legal measures like selection rate that define discrimination based on relative performance to the most privileged group prove inadequate when establishing accuracy benchmarks for gender classification. Here discrimination means the accuracy of a group is less than 4/5ths of the accuracy of the best-classified group. Since gender classification is used for high stakes decision-making, absolute minimum accuracy standards that must be achieved across all subgroups will ensure that systems are suitable to use on phenotypically diverse populations.

Final Thoughts

The intersectional dataset evaluation demonstrates existing face datasets do not reflect the increasingly diverse populations that are exposed to automated facial analysis systems. Most alarmingly the government IJB-A dataset described as being geographically diverse severely underrepresents females with darker skin while significantly over representing lighter males and lighter-skinned individuals in general. Like prior benchmarks, it also fails to achieve gender parity with a ratio of 1 female to every 3 males. Datasets that are largely male and pale provide a false sense of universal progress when used as benchmarks and can encode bias when used for training machine learning algorithms. To provide a more realistic picture of the current state-of-the-art for automated facial analysis tasks like gender classification, benchmarks need to be more phenotypically and demographically representative. Performance metrics on more inclusive benchmarks must be disaggregated to show subgroup performance. As shown in this work, intersectional analysis can highlight performance disparities between subgroups that are otherwise obfuscated by aggregate measures. Minimal accuracy standards must be achieved on each subgroup of interest before an automated facial analysis algorithm can be used reliably for high stakes decisions resulting from biometric classification.

Commercial products that use gender classification should provide data on how well their services perform across a range of demographic and phenotypic attributes. At the minimum, the skin type phenotype for which there is an existing scientific classification scale, namely the Fitzpatrick Scale, can be used. Datasheets that outline how an artifact performs under various physical conditions have long been provided for hardware components. Inclusion datasheets can be used for algorithms to outline how they perform in relation to various social, cultural, and where appropriate, phenotypic conditions. Figure 19 provides an example of what such a report could look like.



ALGORITHM INCLUSION SCORECARD

Quick Stats	Usage	Success Rates	Fail Rates	Ideal Use	Challenges
Release Year	Target	Age #	Age #s	Range	Range
DataSet		Gender #	Gender #s	Genders	Genders
Provider		Ethnic # SkinType #	Ethnic #s	Accurate Ethnicities	Ethnic Bias
Size		Detection #	Detection #	Conditions	Conditions

Quick Stats	Usage	Success Rates	Fail Rates	Ideal Use	Challenges
2017	General	Age Performance	Age Bias	20-40	(50-60)
Trained: LFW		Gender Performance	Gender Bias	Male	Hispanic Male
Provider: MIT		Ethnic Performance Skin Performance	Ethnic Bias	White , Asian	Black Dark Skin
		Detection: 95%	No Face:5%	Outdoors	Night Time

Figure 19. Algorithmic Inclusion Scorecard

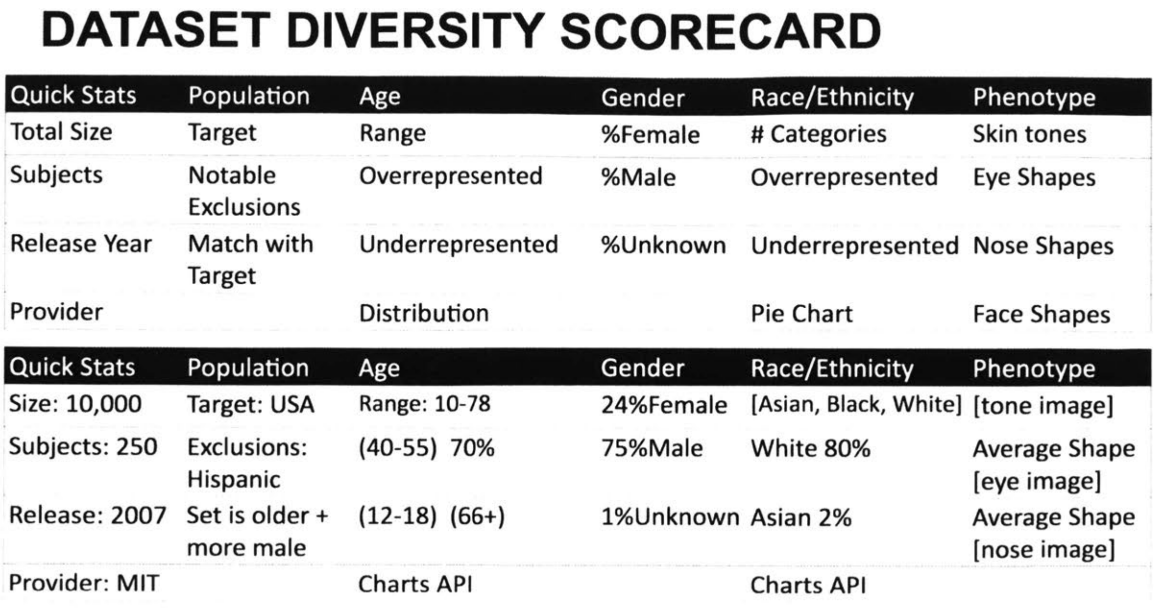
Caution should be taken in using nationality as a proxy for ethnicity or ethnicity as a proxy for phenotypic characteristics like skin type or eye shape. Though previous studies have used nationality as a proxy for ethnicity, ethnicity and race distinctions are unstable, overlapping, and defined differently in nations around the world. The rules for membership evolve overtime. The United States census has an option for selecting Hispanic ethnicity that can be applied to any race on the census. For a given nation, migrations can shift population demographics and phenotypic dimensions. In a multiethnic country like Brazil, an aggregate accuracy measure would not adequately map to a specific ethnicity or phenotype. For a given ethnic or racial classification, intraclass variation can be wide. Attention must be taken to adequately represent phenotypic variations within these classifications. In the United States, attempting to diversify a dataset by including only light-skinned African-Americans would be insufficient for full phenotypic representation as it relates to skin type for all African-Americans.

Without care a dataset that is geographically diverse may not be phenotypically inclusive. Still, nationality or ethnicity can be used to help guide curation as a starting point but not as an end to itself. The Pilot Parliaments Benchmark was constructed based on expected skin type distribution in European and African countries, but the final analysis was performed on subgroups based on phenotypic characteristics. Light-skinned women in South Africa were grouped with the light-skinned women from European parliaments.

Discussions about fairness, accountability, and transparency in artificial intelligence fueled automation have largely focused on predictive models that make inferences about an uncertain future. Less attention has been focused on the accuracy of verifiable tasks like gender classification that are tackled by machine learning. The fairness discourse can be broadened by acknowledging the accuracy disparities on verifiable tasks constitute another form of unfairness that must be actively assessed. Algorithmic accountability for human-focused computer vision necessitates measures that explicitly attend to phenotypic differences between groups. For the case of gender classification, skin type and national origin were used in concert to assess classification performance on distinct subgroups in this work. For tasks like iris verification or hand tracking, the relevant phenotypic differences will need to be established. We cannot assume largely homogenous or heavily skewed datasets can form the basis of unquestioned metrics that are presented in a universal manner. **A single accuracy measure without subgroup analysis should be explicitly recognized as a cursory and incomplete analysis. National benchmarks and competitions need to specifically include subgroup accuracy as part of overall performance scores if the goal is to create systems that work well for all of humanity and not a few data rich groups**. Algorithmic accountability as it relates to human-focused computer vision must include transparent rigorous evaluation across phenotypic and demographic factors.

Accountability must also start at the conceptualization stage when a construct or target variable is defined. **When supervised computer vision models are used to assess constructs like gender, the models have been trained to learn the visual displays of the construct.** Thus gender classification as is practiced is an exercise of learning gender display norms. Gender display is socially, historically, and culturally influenced. An individual's display of gender in one cultural or temporal space may not be the same in another. Social expectations for gender expression to fall along masculine or feminine continuums and what is perceived as masculine or feminine change over time. Moreover, the cultural recognition of hijras and two-spirit people has existed for over a millennia, yet current binary gender classifiers by construct fail to account for non-binary gender identities and also do not account for transgender identities. Advanced identity representation provides alternative ways of representing group membership that can inform a more robust construction of gender for machine learning in the future (Harrell, 2009). Instead of using a binary flag to denote belonging or exclusion to a class, Harrell's system uses measures of centrality in relation to members who most typify class identity. Establishing archetypes can still inculcate bias, but by stating the assumptions that are made about group membership, we can better represent identities that fall outside of normative assumptions and understand the ways in which these assumptions can be shaped to become more inclusive.

Next, accountability should continue during the data collection and labeling stages. **Actively checking for demographic and phenotypic representation as well as gender parity should become part of common practice rather than merely a commendable option.** Transparency in the demographic and phenotypic composition of training data and benchmarks will increase credibility and confidence in using inferences resulting from datasets to make performance claims. Figure 20 presents an image of what a data diversity report could look like at a minimum.



DATASET DIVERSITY SCORECARD

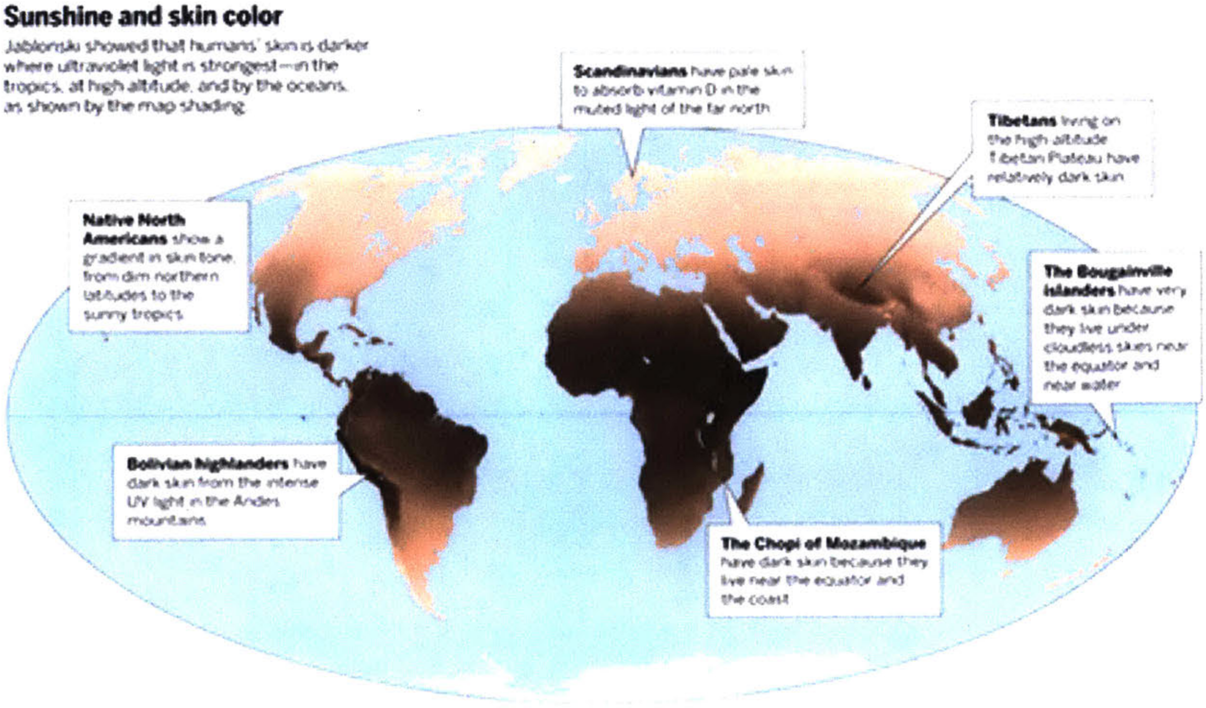
Quick Stats	Population	Age	Gender	Race/Ethnicity	Phenotype
Total Size	Target	Range	%Female	# Categories	Skin tones
Subjects	Notable Exclusions	Overrepresented	%Male	Overrepresented	Eye Shapes
Release Year	Match with Target	Underrepresented	%Unknown	Underrepresented	Nose Shapes
Provider		Distribution		Pie Chart	Face Shapes

Quick Stats	Population	Age	Gender	Race/Ethnicity	Phenotype
Size: 10,000	Target: USA	Range: 10-78	24%Female	[Asian, Black, White]	[tone image]
Subjects: 250	Exclusions: Hispanic	(40-55) 70%	75%Male	White 80%	Average Shape [eye image]
Release: 2007	Set is older + more male	(12-18) (66+)	1%Unknown	Asian 2%	Average Shape [nose image]
Provider: MIT		Charts API		Charts API	

Figure 20. Dataset Diversity Scorecard

Inaction, ambivalence, and a reliance on skewed data and aggregate accuracy metrics will not just undermine the development of artificial intelligence but will be a form of gross negligence. Inattention to an algorithm's effectiveness on a variety of subgroups can perpetuate harms in any other domain that is touched by automated decision-making. If predictive models trained on largely homogenous data are used for medical diagnostics, the people who are least represented in the data risk receiving the wrong medical advice and treatments. The promise of personalized medicine may only become a true option for people who are data rich, that is to say well represented or modeled by existing data gathering processes. Given the skews that can exist in benchmarks, we must increase the rigor with which benchmarks are constructed and be transparent about differences in performance between different groups. Results must be published with context.

Researchers, policy makers, and industry practitioners who aim to create generalizable models, serve global public interests, or reach broader markets need to place more attention on the under-sampled majority, namely women and people with non-White skin who have been highlighted in this work. I return to the Jablonski map of skin type distribution to show that even though existing face datasets tend to be largely pale and male, they are not reflective of the beautiful sepia spectrum that makes up the majority of humanity now and in the future.



Sunshine and skin color

Jablonski showed that humans' skin is darker where ultraviolet light is strongest—in the tropics, at high altitude, and by the oceans, as shown by the map shading.

Scandinavians have pale skin to absorb vitamin D in the muted light of the far north

Tibetans living on the high altitude Tibetan Plateau have relatively dark skin

Native North Americans show a gradient in skin tone, from dim northern latitudes to the sunny tropics

The Bougainville islanders have very dark skin because they live under cloudless skies near the equator and near water

Bolivian highlanders have dark skin from the intense UV light in the Andes mountains

The Chopi of Mozambique have dark skin because they live near the equator and the coast

Figure 21. Jablonski Skin Map

As we move into the automation era, we risk propagating and obscuring the bias of the past if we fail to change the ways we design, develop, deploy, and evaluate artificial intelligence. To create a future where full-spectrum inclusion is a reality in our datasets, benchmarks, and automated decision-making processes, we must proceed with intention.

Because automation increasingly impacts people's lives, we cannot place data or data-centric technologies like artificial intelligence in a vacuum. Just as when an aerospace engineer moves from textbook models of ideal planes to real aircraft, we have to attend to the real world pressures and frictions that result from bias and external conditions. Acknowledging social, cultural, and historic turbulence will be necessary if artificial intelligence is ever to ascend to the elusive stratosphere of fairness and inclusion.

References

Abdat, F., Maaoui, C., & Pruski, A. (2011). Human-Computer Interaction Using Emotion Recognition from Facial Expression. *2011 UKSim 5th European Symposium on Computer Modeling and Simulation.* doi:10.1109/ems.2011.20

AJL. (n.d.). Retrieved from https://www.ajlunited.org/

Amazon Web Services, Inc. (n.d.). Amazon Rekognition – Deep learning-based image analysis. Retrieved from https://aws.amazon.com/rekognition/

Angwin, J., Mattu, S., Larson, J., & Kirchner, L. (2016, May 31). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Baldassarri, S., Hupont, I., Cerezo, E., & Abadi-a, D. (2013). Affective-aware tutoring platform for interactive digital television. *Multimedia Tools and Applications, 74*(9), 3183-3206. doi:10.1007/s11042-013-1779-z

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3). http://dx.doi.org/10.15779/Z38BG31

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1996). Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *Lecture Notes in Computer Science Computer Vision ECCV '96,* 43-58. doi:10.1007/bfb0015522

Bessendorf, A. (2015, December). *From Cradle to Cane: The Cost of Being a Female Consumer* (Rep.). Retrieved https://www1.nyc.gov/assets/dca/downloads/pdf/partners/Study-of-Gender-Pricing-in-NYC.pdf

Bruce, V., Burton, A. M., Hanna, E., Healey, P., Mason, O., Coombes, A., . . . Linney, A. (1993). Sex Discrimination: How Do We Tell the Difference between Male and Female Faces? *Perception, 22*(2), 131-152. doi:10.1068/p220131

Buolamwini, J. (2016, December 14). The Algorithmic Justice League. Retrieved from https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148#.lh6ftfaoa

Buolamwini, J. (2016, May 16). InCoding - In the beginning. Retrieved from https://medium.com/mit-media-lab/incoding-in-the-beginning-4e2a5c51a45d

Buolamwini, J. (2017, May 29). Algorithms aren't racist. Your skin is just too dark. Retrieved from https://hackernoon.com/algorithms-arent-racist-your-skin-is-just-too-dark-4ed31a7304b8

Buolamwini, J. (n.d.). How I'm fighting bias in algorithms. Retrieved from https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

Burton, M., Bruce, V., & Dench, N. (1993). What's the Difference between Men and Women? Evidence from Facial Measurement. *Perception, 22*(2), 153-176. doi:10.1068/p220153

Census 2011 (Rep.). (2012, October 30). Retrieved https://www.statssa.gov.za/publications/P03014/P030142011.pdf

Cheney, J., Klein, B., Jain, A. K., & Klare, B. F. (2015). Unconstrained face detection: State of the art baseline and challenges. *2015 International Conference on Biometrics (ICB)*. doi:10.1109/icb.2015.7139089

Chihaoui, M., Elkefi, A., Bellil, W., & Amar, C. B. (2016). A Survey of 2D Face Recognition Techniques. *Computers, 5*(4), 21. doi:10.3390/computers5040021

Collins, P. H. (2015). Intersectionality's Definitional Dilemmas. *Annual Review of Sociology, 41*(1), 1-20. doi:10.1146/annurev-soc-073014-112142

Crawford, K., Whittaker, M., Elish, M. C., Barocas, S., Plasek, A., & Ferryman, K. (2016, September 22). *The AI Now Report : The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term* (Rep.). Retrieved https://artificialintelligencenow.com/media/documents/AINowSummaryReport_3_RpmwKHu.pdf

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *The University of Chicago Legal Forum, 140*, 139-167.

Dantcheva, A., Elia, P., & Ross, A. (2016). What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. *IEEE Transactions on Information Forensics and Security, 11*(3), 441-467. doi:10.1109/tifs.2015.2480381

Dass, A. (n.d.). Humanae. Retrieved from http://humanae.tumblr.com/about

Data and Society Report on Activities (Rep.). (2016). Retrieved https://datasociety.net/pubs/ar/DS_Report-on-Activities_2015-2016.pdf

Dey, S., & Samanta, D. (2014). *Unimodal and Multimodal Biometric Data Indexing*. De Gruyter.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS Risk Scales: Demonstrating accuracy equity and predictive parity* (Rep.). Northpointe.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*. doi:10.1145/2090236.2090255

Ekman, P. (2005). Facial Expressions. *Handbook of Cognition and Emotion,* 301-320. doi:10.1002/0470013494.ch16

Everingham, M., Eslami, S. M., Gool, L. V., Williams, C. K., Winn, J., & Zisserman, A. (2014). The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision, 111*(1), 98-136. doi:10.1007/s11263-014-0733-5

Face API. (n.d.). Retrieved from https://azure.microsoft.com/en-us/services/cognitive-services/face/

Face Attributes. (n.d.). Retrieved from https://www.faceplusplus.com/attributes/#demo

Farinella, G., & Dugelay, J. (2012). Demographic classification: Do gender and ethnicity affect each other? *2012 International Conference on Informatics, Electronics & Vision (ICIEV)*. doi:10.1109/iciev.2012.6317383

Fellous, J. (1997). Gender discrimination and prediction on the basis of facial metric information. *Vision Research, 37*(14), 1961-1973. doi:10.1016/s0042-6989(97)00010-2

Fermi, D., B, F. N., Radhakrishnan, R., & Kartha, S. S. (2017). A survey on different face detection algorithms in image processing. *International Journal of Innovative Research in Science, Engineering and Technology, 6*(1), 151-156. doi:10.15680/ijirset

ISSN(Online) : 2319-8753

Fitzpatrick, T. B. (1988). The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Archives of Dermatology, 124*(6), 869. doi:10.1001/archderm.1988.01670060015008

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *ArXiv.org*. Retrieved from https://arxiv.org/abs/1609.07236.

Furl, N., Phillips, P., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis. *Cognitive Science, 26*, 797-815.

Fuss, D. (1989). *Essentially speaking: Feminism, nature and difference*. Routledge.

Garvie, C., Frankle, J., & Bedoya, A. (2016). *The perpetual line-up: Unregulated police face recognition in America*. Washington, DC: Georgetown Law, Center on Privacy & Technology.

Goffman, E. (1979). *Gender advertisements*. Cambridge, MA: Harvard University Press.

Goldstein, A., Harmon, L., & Lesk, A. (1971). Identification of human faces. *Proceedings of the IEEE, 59*(5), 748-760. doi:10.1109/proc.1971.8254

Golomb, B., Lawrence, D. T., & Sejnowski, T. J. (1990). *SEXNET: A neural network identifies sex from human faces.* Denver, Colorado: Advances in Neural Information Processing Systems 3.

Goodman, B., & Flaxman, S. (2016). European Union regulations on algorithmic decision-making and a "right to explanation". *ArXiv.org*. Retrieved from https://arxiv.org/pdf/1606.08813v1.pdf.

Grother, P. J., & Ngan, M. L. (2017, February 19). Face Recognition Vendor Test (FRVT) Performance of Face Identification Algorithms NIST IR 8009. Retrieved from https://www.nist.gov/node/558561

Grother, P., Ngan, M., & Hanaoka, K. (2017, April 12). *Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification* (Rep.). Retrieved http://biometrics.nist.gov/cs_links/face/FRVT/frvt_report_2017_04_12.pdf

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., . . . Wang, G. (2017). Recent advances in convolutional neural networks. Retrieved from arXiv:1512.07108v5

Hammal, Z., Couvreur, L., Caplier, A., & Rombaut, M. (2007). Facial expression classification: An approach based on the fusion of facial deformations using the transferable belief model. *International Journal of Approximate Reasoning, 46*(3), 542-567. doi:10.1016/j.ijar.2007.02.003

Han, H., & Jain, A. K. (2014). *Age, gender and race estimation from unconstrained face images* (Rep.). Michigan State University.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *ArXiv.org*. Retrieved from https://arxiv.org/pdf/1610.02413.pdf.

Harrell, D. F. (2009). Computational and cognitive infrastructures of stigma. *Proceeding of the Seventh ACM Conference on Creativity and Cognition - C&C '09.* doi:10.1145/1640233.1640244

Hern, A. (2016, September 28). 'Partnership on AI' formed by Google, Facebook, Amazon, IBM and Microsoft. Retrieved from https://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms

Hersch, J. (2006). Skin-Tone Effects among African Americans: Perceptions and Reality. *American Economic Review, 96*(2), 251-255. doi:10.1257/000282806777212071

Ingold, D., & Soper, S. (2016, April 21). Amazon Doesn't Consider the Race of Its Customers. Should It? Retrieved May 3, 2017, from https://www.bloomberg.com/graphics/2016-amazon-same-day/

Intelligence Advanced Research Project Activity. (n.d.). Retrieved from https://www.challenge.gov/agency/intelligence-advanced-research-project-activity/

Inter-Parliamentary Union, & United Nations Entity for Gender Equality and the Empowerment of Women. (2017). Women in politics 2017. Retrieved from http://www.ipu.org/english/surveys.htm#MAP2017

Inter-Parliamentary Union. (2017, July 1). Women in national parliaments. Retrieved from http://www.ipu.org/wmn-e/classif.htm

Jackson, J. S., & Gurin, G. (n.d.). National Survey of Black Americans, 1979-1980. *ICPSR Data Holdings*. doi:10.3886/icpsr08512

Jackson, J. S., & Neighbors, H. W. (1997). National Survey of Black Americans, Waves 1-4, 1979-1980, 1987-1988, 1988-1989, 1992. *ICPSR Data Holdings*. doi:10.3886/icpsr06668

Jacobs, S. (2005). *Two-spirit people: Native American gender identity, sexuality, and spirituality*. Urbana, Ill.: University of Illinois Press.

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., & Brossard, E. (2016). The MegaFace Benchmark: 1 Million Faces for Recognition at Scale. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* doi:10.1109/cvpr.2016.527

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W., & Jain, A. K. (2012). Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security, 7*(6), 1789-1801. doi:10.1109/tifs.2012.2214212

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., . . . Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* doi:10.1109/cvpr.2015.7298803

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. Working Paper. "Inherent Trade-Offs in the Fair Determination of Risk Scores".* (Working paper). doi:https://arxiv.org/pdf/1609.05807.pdf

Larson, J., Angwin, J., Kirchner, L., & Mattu, S. (2016, August 01). How we analyzed the COMPAS recidivism algorithm. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Levi, G., & HassNer, T. (2015). Age and gender classification using convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR).* doi:10.1109/cvprw.2015.7301352

Lubin, G. (2016, October 12). 'Facial-profiling' could be dangerously inaccurate and biased, experts warn. Retrieved from http://www.businessinsider.com/does-faception-work-2016-10

Lynch, J. (n.d.). *United States House Committee on Oversight and Government Reform held a Hearing on Law Enforcement's Use of Facial Recognition Technology - Written Testimony* (Publication). Retrieved from https://oversight.house.gov/wp-content/uploads/2017/03/Lynch-EFF-Statement-FRT-Study-3-22.pdf

Makinen, E., & Raisamo, R. (2008). Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(3), 541-547. doi:10.1109/tpami.2007.70800

Mathias, M., Benenson, R., Pedersoli, M., & Gool, L. V. (2014). Face Detection without Bells and Whistles. *Computer Vision ECCV 2014 Lecture Notes in Computer Science,* 720-735. doi:10.1007/978-3-319-10593-2_47

MIT Media Lab to participate in $27 million initiative on AI ethics and governance. (2017, January 10). Retrieved from http://news.mit.edu/2017/mit-media-lab-to-participate-in-ai-ethics-and-governance-initiative-0110

Ngan, M., & Grother, P. (2015, April). *NISTIR 8052 Face Recognition Vendor Test (FRVT) Performance of Automated Gender Classification Algorithms* (Rep.). Retrieved http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8052.pdf

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* New York: Crown.

Ozbudak, O., Kirci, M., Cakir, Y., & Gunes, E. O. (2010). Effects of the facial and racial features on gender classification. *Melecon 2010 - 2010 15th IEEE Mediterranean Electrotechnical Conference*. doi:10.1109/melcon.2010.5476346

Ozbudak, O., Kirci, M., Cakir, Y., & Gunes, E. O. (2010). Effects of the facial and racial features on gender classification. *Melecon 2010 - 2010 15th IEEE Mediterranean Electrotechnical Conference*. doi:10.1109/melcon.2010.5476346

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. *Procedings of the British Machine Vision Conference 2015*. doi:10.5244/c.29.41

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'toole, A. J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception, 8*(2), 1-11. doi:10.1145/1870076.1870082

Phillips, P., Moon, H., Rauss, P., & Rizvi, S. (n.d.). The FERET evaluation methodology for face-recognition algorithms. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. doi:10.1109/cvpr.1997.609311

Phillips, P., Moon, H., Rizvi, S., & Rauss, P. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(10), 1090-1104. doi:10.1109/34.879790

Phillips, P., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing, 16*(5), 295-306. doi:10.1016/s0262-8856(97)00070-x

Picard, R. W. (2000). *Affective computing*. Cambridge, MA: MIT Press.

The Pink Tax : How Gender-Based Pricing Hurts Women's Buying Power (Publication). (2016, December). Retrieved https://www.jec.senate.gov/public/_cache/files/8a42df04-8b6d-4949-b20b-6f40a326db9e/the-pink-tax---how-gender-based-pricing-hurts-women-s-buying-power.pdf

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98-125. doi:10.1016/j.inffus.2017.02.003

Raisamo, R., & Makinen, E. (2008). An experimental comparison of gender classification methods. *Pattern Recognition Letters, 29*(10), 1544-1556. doi:10.1016/j.patrec.2008.03.016

Reid, D., Samangooei, S., Chen, C., Nixon, M., & Ross, A. (2013). Soft Biometrics for Surveillance: An Overview. *Handbook of Statistics - Machine Learning: Theory and Applications Handbook of Statistics,* 327-352. doi:10.1016/b978-0-444-53859-8.00013-8

Rothe, R., Timofte, R., & Gool, L. V. (2015). DEX: Deep EXpectation of Apparent Age from a Single Image. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. doi:10.1109/iccvw.2015.41

Russakovsky, O., & Deng, J. (2015). ImageNet large scale visual recognition challenge. Retrieved from https://arxiv.org/pdf/1409.0575v3.pdf

Schneiderman, H. (2004). Feature-centric evaluation for efficient cascaded object detection. *Proceedings of the 2004 IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition, 2004. CVPR 2004. doi:10.1109/cvpr.2004.1315141

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* doi:10.1109/cvpr.2015.7298682

Selbst, A. D. (2017). Disparate Impact in Big Data Policing. *Georgia Law Review.* doi:10.2139/ssrn.2819182

Shan, C. (2012). Smile detection by boosting pixel differences. *IEEE Transactions on Image Processing, 21*(1), 431-436. doi:10.1109/tip.2011.2161587

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A, 4*(3), 519. doi:10.1364/josaa.4.000519

Stotzer, R. L. (2009). Violence against transgender people: A review of United States data. *Aggression and Violent Behavior, 14*(3), 170-179. doi:10.1016/j.avb.2009.01.006

Suh, B., Ling, H., Bederson, B. B., & Jacobs, D. W. (2003). Automatic thumbnail cropping and its effectiveness. *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology - UIST '03.* doi:10.1145/964696.964707

Sun, Y., Wang, X., & Tang, X. (2015). Deeply learned face representations are sparse, selective, and robust. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* doi:10.1109/cvpr.2015.7298907

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *2014 IEEE Conference on Computer Vision and Pattern Recognition.* doi:10.1109/cvpr.2014.220

Thomee, B., Elizalde, B., Shamma, D. A., Ni, K., Friedland, G., Poland, D., . . . Li, L. (2016). YFCC100M: The new data and new challenges in multimedia research. *Communications of the ACM, 59*(2), 64-73. doi:10.1145/2812802

Turk, M., & Pentland, A. (1991). Face recognition using eigenfaces. *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* doi:10.1109/cvpr.1991.139758

Viola, P., & Jones, M. (n.d.). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.* doi:10.1109/cvpr.2001.990517

Visual Recognition - IBM Bluemix. (2016, May 01). Retrieved from https://console.bluemix.net/catalog/services/visual-recognition/

Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. Retrieved from https://arxiv.org/pdf/1611.04135v1.pdf

Yamamoto, K., Kobayashi, H., Tagami, Y., & Nakayama, H. (2016). Multimodal Content-Aware Image Thumbnailing. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion.* doi:10.1145/2872518.2889413

Yan, J., Lei, Z., Wen, L., & Li, S. Z. (2014). The Fastest Deformable Part Model for Object Detection. *2014 IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/cvpr.2014.320

Yan, J., Zhang, X., Lei, Z., & Li, S. Z. (2013). Real-time high performance deformable model for face detection in the wild. *2013 International Conference on Biometrics (ICB)*. doi:10.1109/icb.2013.6612972

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Learning face representation from scratch. *ArXiv.org*. Retrieved from https://arxiv.org/pdf/1411.7923.pdf.

Zafeiriou, S., Zhang, C., & Zhang, Z. (2015). A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding, 138*, 1-24. doi:10.1016/j.cviu.2015.03.015

Zhang, C., & Zhang, Z. (2014). Improving multiview face detection with multi-task deep convolutional neural networks. *IEEE Winter Conference on Applications of Computer Vision*. doi:10.1109/wacv.2014.6835990

Zliobaite, I. (2015, October 31). *A survey on measuring indirect discrimination in machine learning*. doi:arXiv:1511.00148v1

Zwick, R., & Green, J. G. (2007). New Perspectives on the Correlation of SAT Scores, High School Grades, and Socioeconomic Factors. *Journal of Educational Measurement, 44*(1), 23-45. doi:10.1111/j.1745-3984.2007.00025.x