Graduate Students: Project Reminder

Midpoint due is on Nov. 15 (< 3 weeks from now)

Midpoint presentations on Mon, Nov. 15.

Guidelines will be released this weekend

Make progress every day.

Keep a notebook & write as you go*, so that you are not writing both the report and making the slides at the last minute.*

CS 295B/CS 395B
Systems for Knowledge Discovery

Demographics of AMT

The University of Vermont

Topics for today

Why should we care about the demographics of AMT in the first place?

What are the demographics of AMT?

Context for Monday’s reading.

Why should we care?

What do we mean by demographics?

- Features of crowd workers
    - Age, Ethnicity, Gender
    - Mother tongue
    - Employment status

Social Science/Ethnographic research ⟷ AI/ML research

Draw ER diagram on the board

What do we mean by demographics?

- Features of crowd workers
  - Age, Ethnicity, Gender
  - Mother tongue
  - Employment status

Obvious why we should care

Social Science/Ethnographic research ↔ AI/ML research

What do we mean by demographics?

- Features of crowd workers
  - Age, Ethnicity, Gender
  - Mother tongue
  - Employment status

Less obvious why we should care

Social Science/Ethnographic research

AI/ML research


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

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139
Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Editors: Sorelle A. Friedler and Christo Wilson

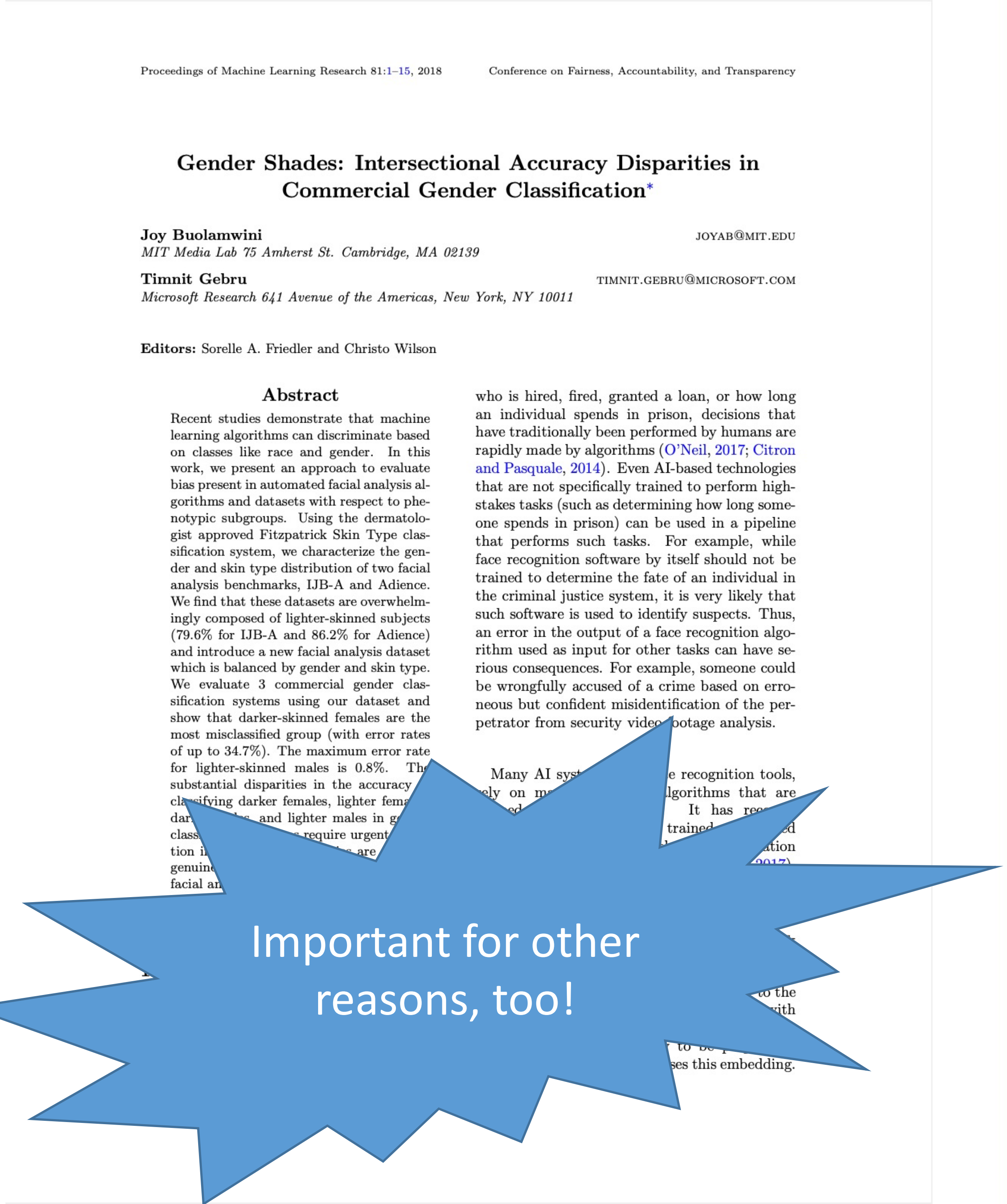
Abstract

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

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

Paper idea: empirical analysis of gender classification for computer vision

Findings: Poor performance for women, abysmal performance for dark-skinned women


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

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

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

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

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

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

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

Paper idea: empirical analysis of gender classification for computer vision

Findings: Poor performance for women, abysmal performance for dark-skinned women

Mainly attributed to class imbalance

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

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

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

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

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

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

Keywords: Computer Vision, Algorithmic Audit, Gender Classification

1. Introduction

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

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

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

Paper idea: empirical analysis of gender classification for computer vision

Findings: Poor performance for women, abysmal performance for dark-skinned women

- Prior work in NLP on bias


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

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

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

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

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

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

Keywords: Computer Vision, Algorithmic Audit, Gender Classification

1. Introduction

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

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

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

Paper idea: empirical analysis of gender classification for computer vision

Findings: Poor performance for women, abysmal performance for dark-skinned women

- Prior work in NLP on bias
- This work started discourse on bias in *variable construction*

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

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

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

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

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

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

Keywords: Computer Vision, Algorithmic Audit, Gender Classification

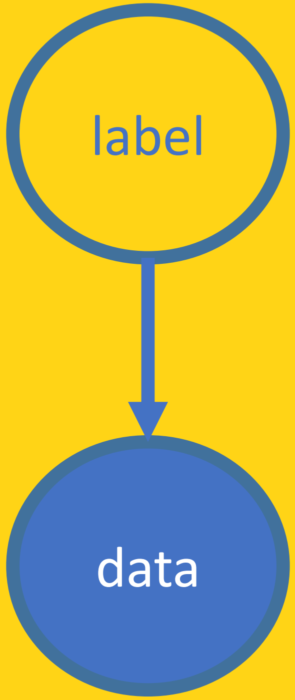
1. Introduction

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

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

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

Classic Causal Assumption



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

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

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

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

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

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

Keywords: Computer Vision, Algorithmic Audit, Gender Classification

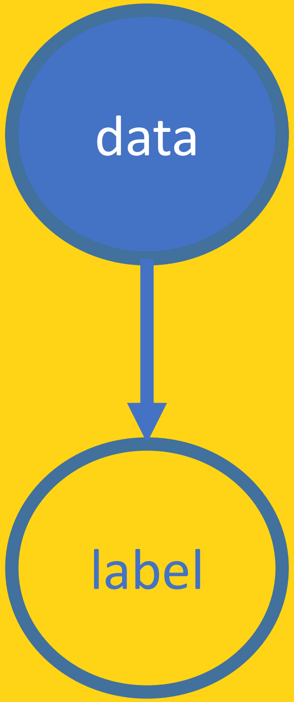
1. Introduction

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

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

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

New Causal Assumption



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

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

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

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

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

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

Keywords: Computer Vision, Algorithmic Audit, Gender Classification

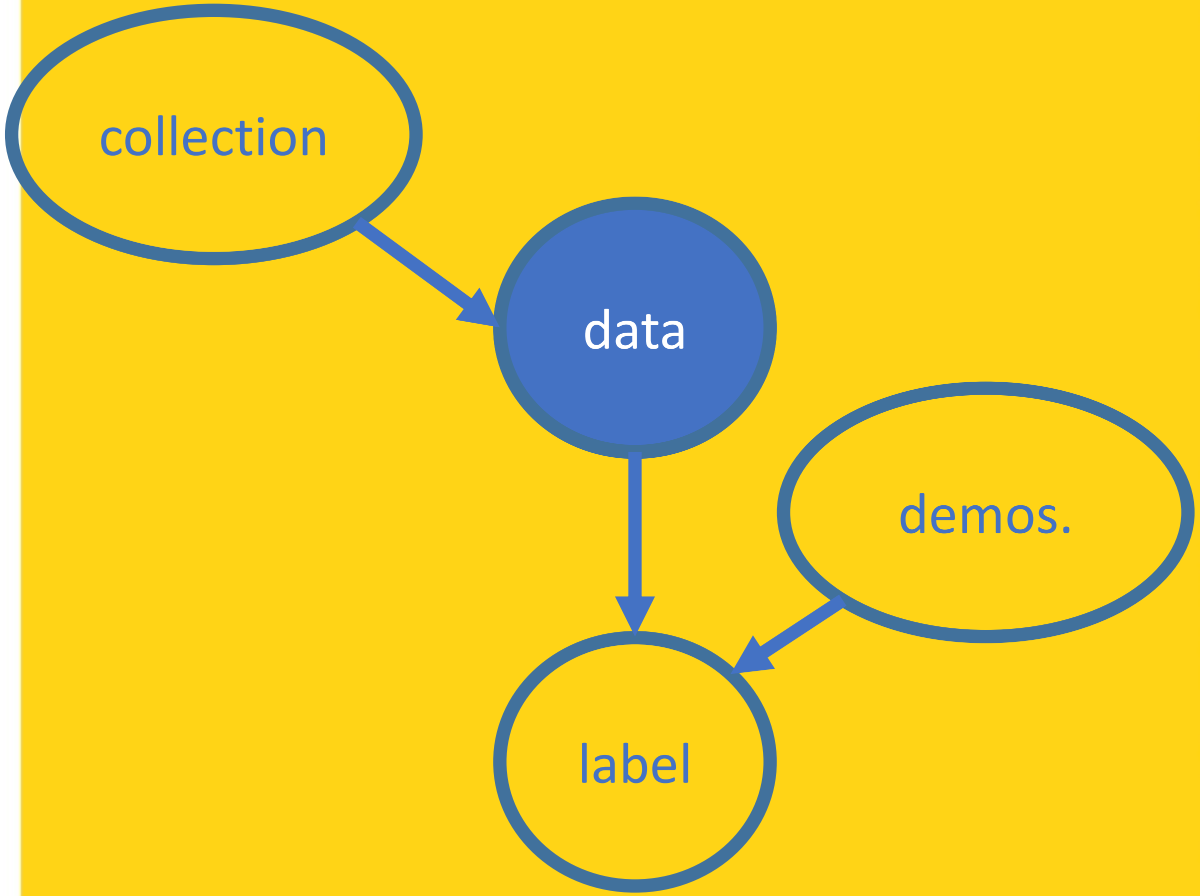
1. Introduction

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

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

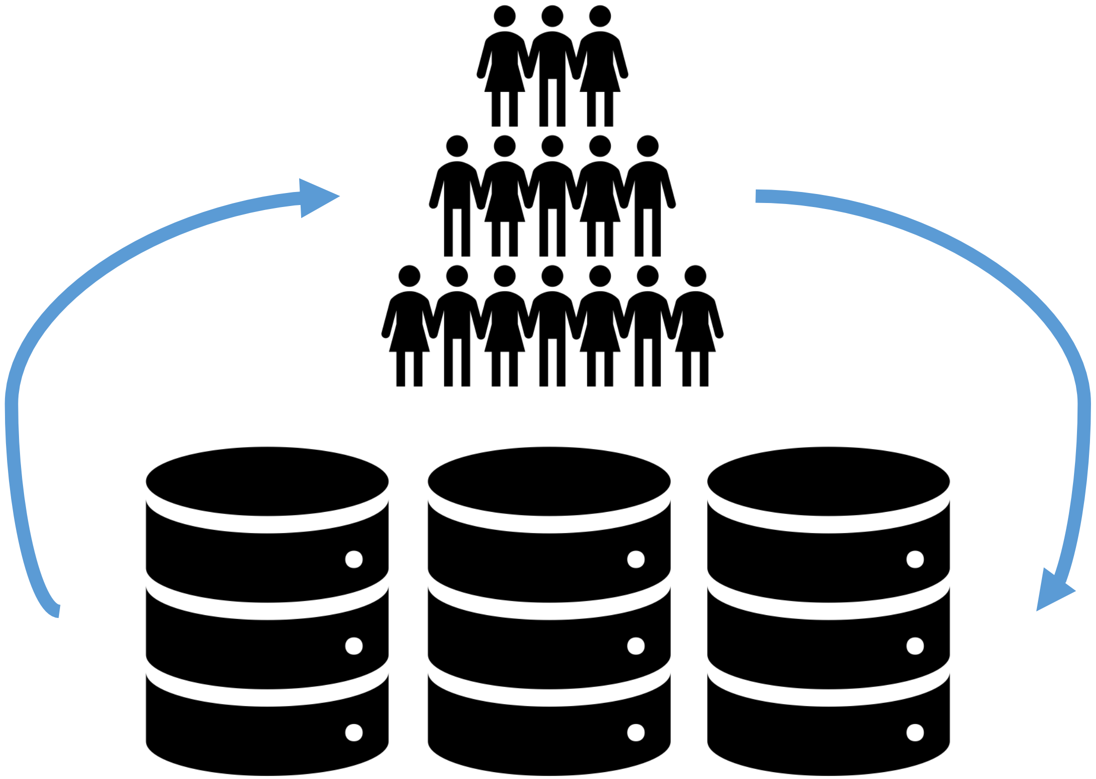
* Download our gender and skin type balanced PPB dataset at gendershades.org

New Causal Assumption



Why does this matter?

Why does this matter?



What are the demographics?

Turkers over time

AMT opened: 2005

A lot has changed in 15 years

Many early demographic studies: 2010-2015

Behav Res (2012) 44:1–23
DOI 10.3758/s13428-011-0124-6

Conducting behavioral research on Amazon's Mechanical Turk

Winter Mason · Siddharth Suri

Published online: 30 June 2011

Abstract Amazon's Mechanical Turk is an online labor market where requesters post jobs and workers choose which jobs to do for pay. The central purpose of this article is to demonstrate how to use this Web site for conducting behavioral research and to lower the barrier to entry for researchers who could benefit from this platform. We describe general techniques that apply to a variety of types of research and experiments across disciplines. We begin by discussing some of the advantages of doing experiments on Mechanical Turk, such as easy access to a large, stable, and diverse subject pool, the low cost of doing experiments, and faster iteration between developing theory and executing experiments. While other methods of conducting behavioral research may be comparable to or even better than Mechanical Turk on one or more of the axes outlined above, we will show that when taken as a whole Mechanical Turk can be a useful tool for many researchers. We will discuss how the behavior of workers compares with that of experts and laboratory subjects. Then we will illustrate the mechanics of putting a task on Mechanical Turk, including recruiting subjects, executing the task, and reviewing the work that was submitted. We also provide solutions to common problems that a researcher might face when executing their research on this platform, including techniques for conducting synchronous experiments, methods for ensuring high-quality work, how to keep data private, and how to maintain code security.

Keywords Crowdsourcing · Online research · Mechanical turk

Introduction

The creation of the Internet and its subsequent widespread adoption has provided behavioral researchers with an additional medium for conducting studies. In fact, researchers from a variety of fields, such as economics (Hossain & Morgan, 2006; Reiley, 1999), sociology (Centola, 2010; Salganik, Dodds, & Watts, 2006), and psychology (Birnbaum, 2000; Nosek, 2007), have used the Internet to conduct behavioral experiments.[1] The advantages and disadvantages of online behavioral research, relative to laboratory-based research, have been explored in depth (see, e.g., Kraut et al., 2004; Reips, 2000). Moreover, many methods for conducting online behavioral research have been developed (e.g., Birnbaum, 2004; Gosling & Johnson, 2010; Reips, 2002; Reips & Birnbaum, 2011). In this article, we describe a tool that has emerged in the last 5 years for conducting online behavioral research: crowdsourcing platforms. The term *crowdsourcing* has its origin in an article by Howe (2006), who defined it as a job outsourced to an undefined group of people in the form of an open call. The key benefit of these platforms to behavioral researchers is that they provide access to a persistently available, large set of people who are willing to do tasks—including participating in research studies—for relatively low pay. The crowdsourcing site with one of the largest subject pools is Amazon's Mechanical Turk[2] (AMT), so it is the focus of this article.

[1] This is clearly not an exhaustive review of every study done on the Internet in these fields. We aim only to provide some salient examples.
[2] The name "Mechanical Turk" comes from a mechanical chess-playing automaton from the turn of the 18th century, designed to look like a Turkish "sorcerer," which was able to move pieces and beat many opponents. While it was a technological marvel at the time, the real genius lay in a diminutive chess master hidden in the workings of the machine (see http://en.wikipedia.org/wiki/The_Turk). Amazon's Mechanical Turk was designed to hide human workers in an automatic process of the platform.

Myth: Turkers are anonymous

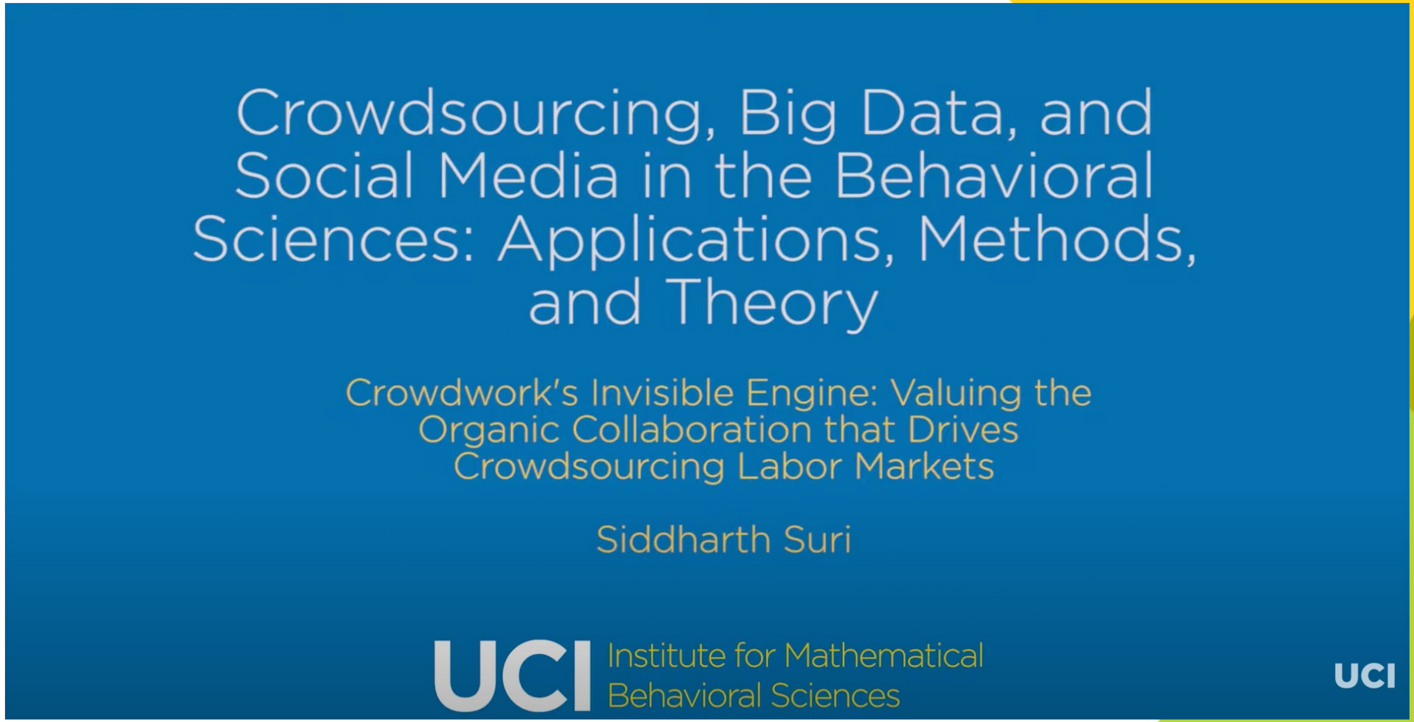
We studied how well the privacy attitudes of MTurk workers mirror the privacy attitudes of the larger user population. We report results from an **MTurk survey** of attitudes about **managing one's personal information online and policy preferences about anonymity**. We compare these attitudes with those of a **representative U.S. adult sample** drawn from a separate survey a few months earlier. **MTurk respondents were younger and better educated**, and more likely to use social media than the representative US adult sample. Although they reported a similar amount of personal information online, **U.S. MTurk workers put a higher value on anonymity and hiding information**, were more likely to do so, had more privacy concerns than the larger U.S. public. **Indian MTurk workers were much less concerned than American workers about their privacy and more tolerant of government monitoring**. Our analyses show that these findings hold even when controlling for age, education, gender, and social media use. Our findings suggest that privacy studies using MTurk need to account for differences between MTurk samples and the general population.

SSRN 2014

SOUPS 2014

Talk by Sid Suri (computer scientist @ Microsoft Research)

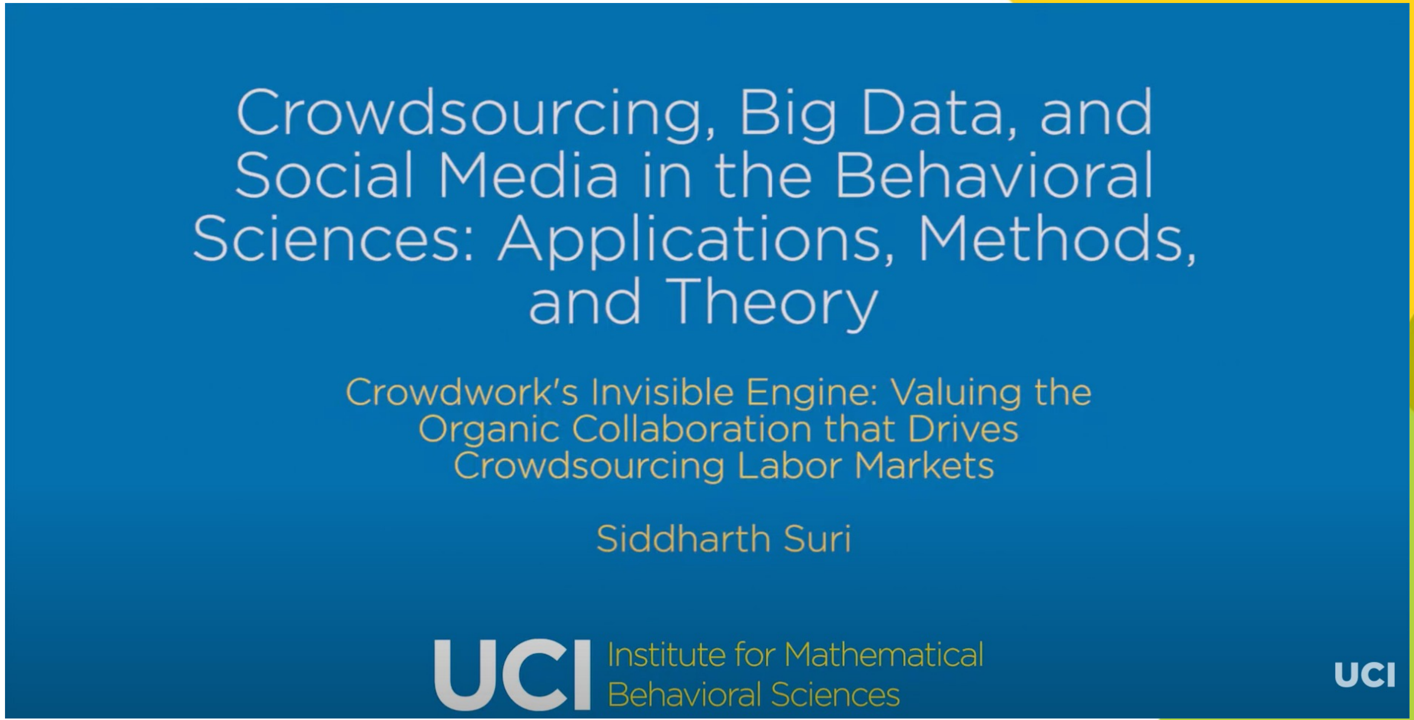
Collaboration with work Mary Gray (ethnographer @ Microsoft Research)



https://www.youtube.com/watch?v=rWSGFA-jme0

Talk by Sid Suri (computer scientist @ Microsoft Research)

Collaboration with work Mary Gray (ethnographer @ Microsoft Research)



https://www.youtube.com/watch?v=rWSGFA-jme0

- 80% US-based
- Indian Turkers highly collaborative
- Most Turkers have other work
- High degree of heterogeneity in how system is used


Inside the world of a Mechanical Turker

Context for Monday's readings

The story of my paper

Research doesn't happen the way it's written in papers

- Original idea: compiling Automan programs*
- List of big problems in crowdsourcing from Sid Suri
- Accepted on first submission

* Aside: How we think about labor has changed

Aside: Academic IRBs and AMT

Student question on Automan: was this granted IRB ap val?

Proposals to use AMT must be sub

However, de-identified crowdwork u

(SurveyMan ran with a consent form + my con

IRBs are NOT ethics review boards

How do we learn about Turkers

Tough nut to crack…

Idea: Use machine learning and multiple data sets to deduce the identities and demographic information from their A zon ids?

JK/LOL

Just f*cking ask them.

Option A: survey
Option B: interview

Variability in Methodological Training

Important to reflect on research cultures

Systems building

- security
- threat model: adversarial behavior
- assumption: start from a place of no trust

Social science

- ethnography
- thread model: measurement error
- assumption: trust is easy to lose