ECS171: Machine Learning

# L17: Ethics in AI

Instructor: Prof. Maike Sonnewald
TAs: Pu Sun &  Devashree Kataria

MOCO Amsterdam garden

# Intended Learning Outcomes: Not on final

- Reflection on consequences of AI development on society
- Reflection on *unintended* consequences of biased algorithms and data

Note: Non-exhaustive list here in class

# Ethics in society, and of technology use and development

Ethics definition:

- Moral principles that govern a *person's* behavior or the conducting of an activity

Ethics of technology:

- Sub-field of ethics addressing ethical questions of the Technology Age,
- The shift in society where personal computers and devices provide quick and easy transfer of information

Ethics of AI:

- Branch of the ethics of technology specific to artificial intelligence (AI) systems

# Ethics of AI

Specific to topics that are considered to have particular ethical stakes:

- Algorithmic biases, fairness, automated decision-making, accountability, privacy, and regulation.

Covers future or emerging challenges:

- Machine ethics (how to make machines that behave ethically)
- Lethal autonomous weapon systems and arms race dynamics
- AI safety and alignment
- Technological unemployment
- AI-enabled misinformation

Some applications may have important ethical implications e.g.: healthcare, education, military

Questions also include if AI systems have a moral status (AI welfare and rights), artificial superintelligence and existential risks

# Rapid changes also raise profound ethical concerns

AI systems can have:

- Embed biases (race, gender, sexual orientation, religious etc)
- Contribute to climate degradation
- Threaten human rights and more

Such risks associated with AI have already begun to compound on top of existing inequalities, resulting in further harm to already marginalised groups.

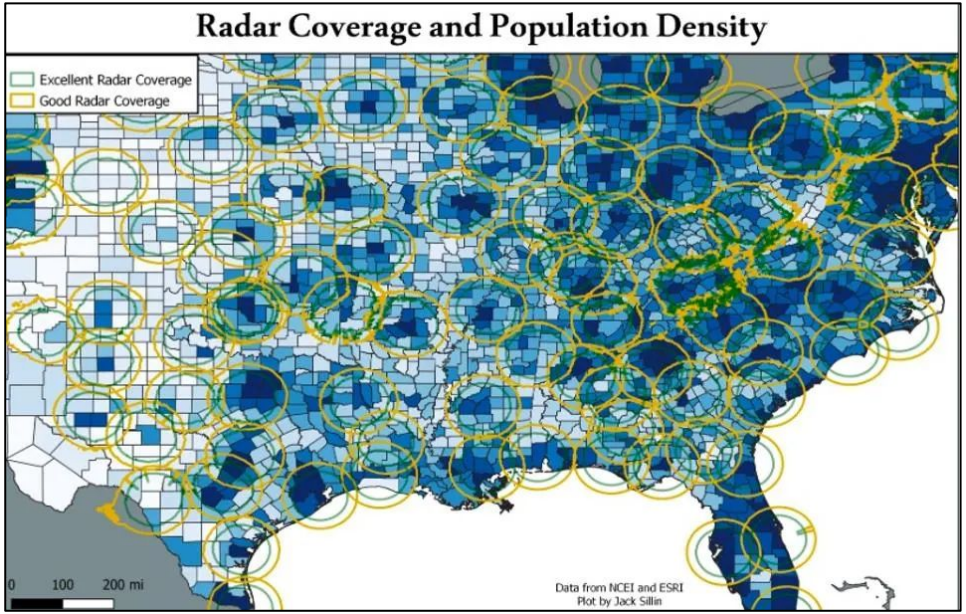
# Weather prediction: Data collection bias

What populations have good data collected have better forecasts

Extreme events e.g. hurricanes have big impact:

- Should you evacuate?
- Where will there be a flood?

Biases in data can disproportionately impact already vulnerable communities



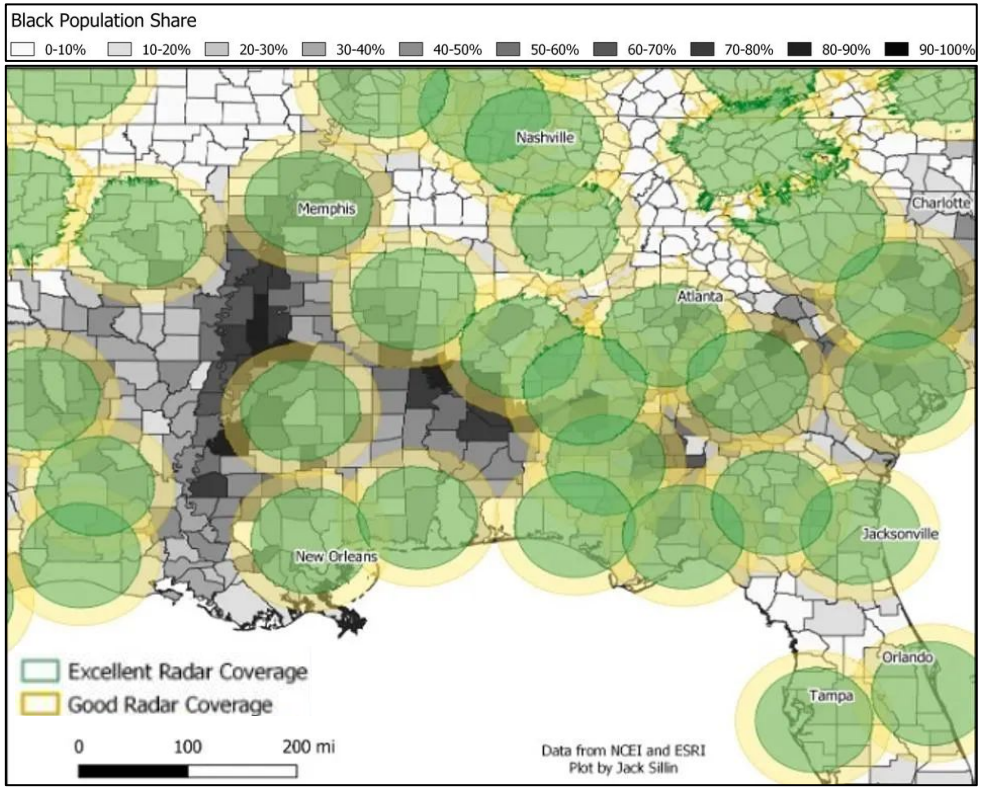
# Weather prediction: Data collection bias

What populations have good data collected have better forecasts

Extreme events e.g. hurricanes have big impact:

- Should you evacuate?
- Where will there be a flood?

Biases in data can disproportionately impact already vulnerable communities




https://www.forbes.com/sites/marshallshepherd/2021/03/20/are-black-and-rural-residents-in-the-south-more-vulnerable-to-tornadoes-due-to-radar-gaps/?sh=35068b534988

# Facial recognition: Civil liberties

Advances in Facial Recognition Technology Have Outpaced Laws, Regulations

- Calls for Federal Government Take Action on Privacy, Equity, and Civil Liberties Concerns
-

Some innocuous uses:

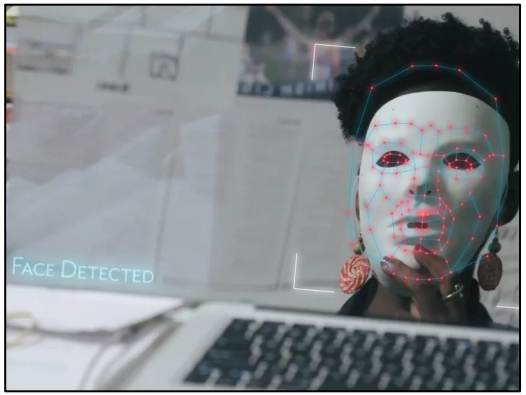
- Unlocking phones
- Finding friends

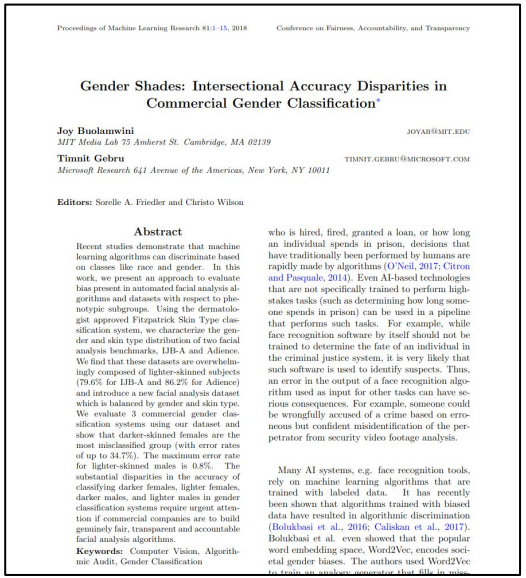
# Facial Recognition: Racial and cultural bias

Many tools rely on training data that is not representative of the population

- Examples where non-white faces have algorithms not perform well
- Different cultures are mis-represented e.g. bride from India classified as performer

More: "Coded Bias" documentary directed by Shalini Kantayya



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

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 miss-

# Large Language Models: Language and view

Current large language models are predominately trained on English-language data, they often present the Anglo-American views as truth

Systematically downplaying non-English perspectives as irrelevant, wrong, or noise

E.g. political ideologies like "What is liberalism?" emphasizing aspects of human rights and equality, while equally valid aspects like "opposes state intervention in personal and economic life" from the dominant Vietnamese perspective and "limitation of government power" from the prevalent Chinese perspective are absent.

# Hiring decisions: AI system turn out to be biased against female and minority candidates

Amazon's AI-powered recruitment tool was trained with its own recruitment data accumulated over the years, during which time the candidates that successfully got the job were mostly white males

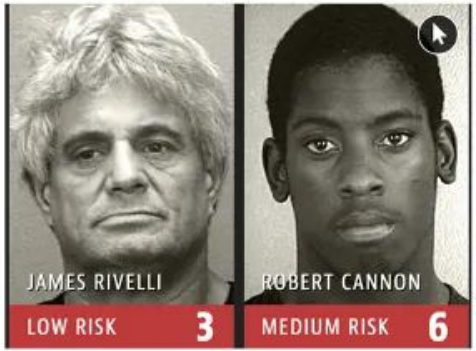
The algorithms learned the (biased) pattern from the historical data and generated predictions for the present/future that these types of candidates are most likely to succeed in getting the job. Therefore, the recruitment decisions made by the

# Bias in software used in legal system

If software is trained on biased data, the ‘risk’ assessment of individuals can be biased in a manner a human observer is more likely to determine more accurately

- Enthusiasm for use can outpace our understanding of bias
- Unintended consequences






# Large Language Models

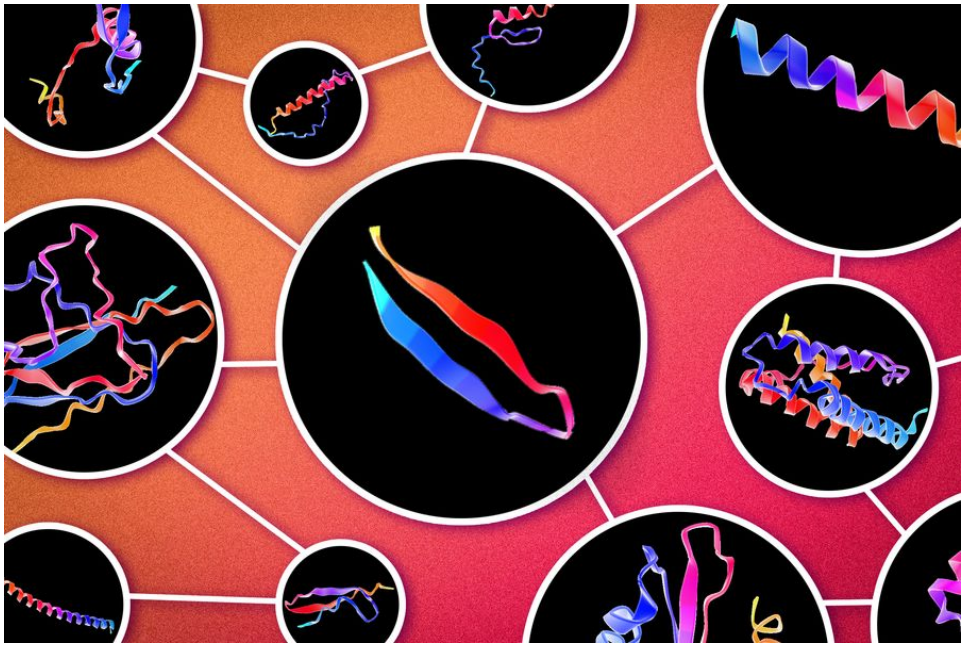
Tools like chatGPT, Bard etc offer capabilities for gathering information and accelerating insight

- Lower bar for e.g. non-native english speakers in writing
- Disseminate information
- Many more

# New materials and medication

Many advances in medicine, materials, fundamental science and other fields are benefitting from AI:

- Do old things faster and better
- Potential for something entirely new

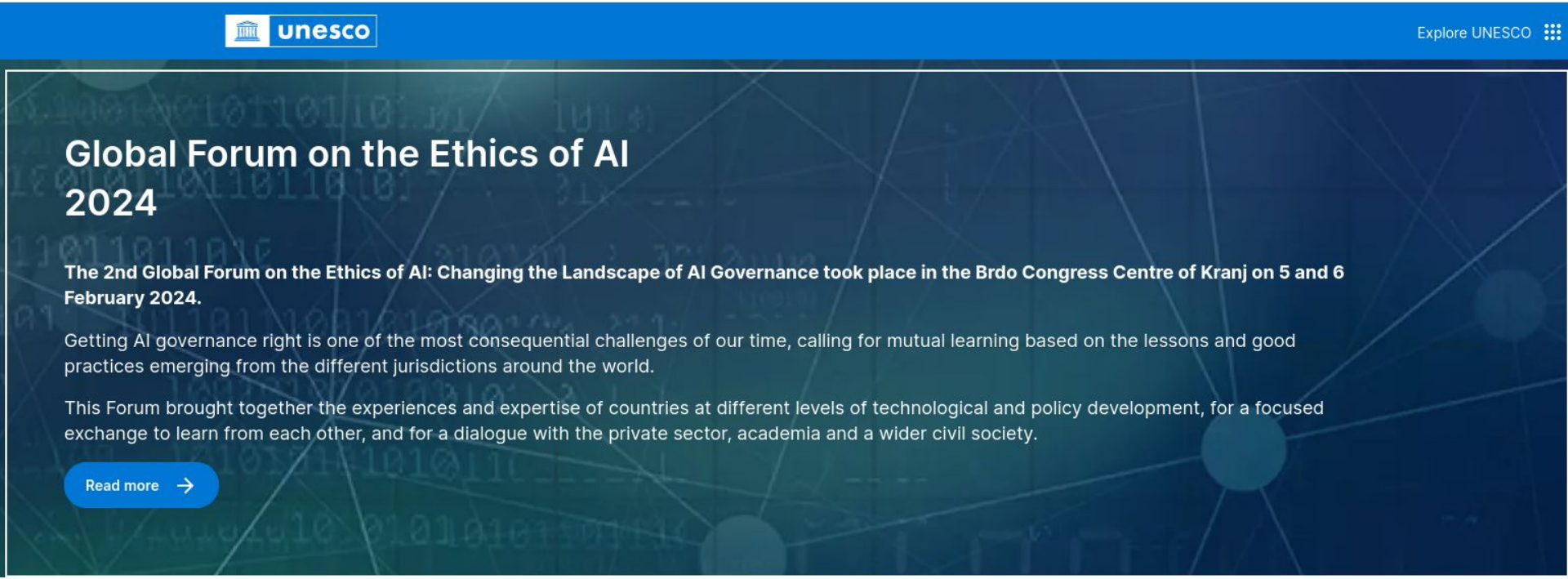


# Open questions….

AI is having a profound effect on humanity, do AI developers, as representatives of future humanity, have an ethical obligation to be transparent in their efforts?

- Should all code be open source? Available does equate to transparent and understandable
- Should all datasets be publicly available?
- What impact would the above have on the *rate* of progress?

# It is unclear how/if we should regulate AI



# Acknowledgements


- Class covered all material from Setareh Rafatirad with modifications and additions
    - Amount of math was the same

# Class time for evaluations!

https://eval.ucdavis.edu

Help students taking the class after you: We care about the success of the future students!

Tell us, instructor and TAs *what went well* and what you think would benefit students taking the class in a different quarter

Constructive comments that are actionable are the most useful…