Investigating Bias

What we'll be covering today...

1. What is AI bias?
2. **Case Study:** Can ChatGPT pass an employment discrimination test?
3. Investigating AI Bias across the AI lifecycle.
4. **Interactive:** Investigate a real AI model for bias.

What we'll be covering today...

1. **What is AI bias?**
2. *Case Study: Can ChatGPT pass an employment discrimination test?*
3. *Investigating AI Bias across the AI lifecycle.*
4. *Interactive: Investigate a real AI model for bias.*

Companies and government promise that AI will be objective and fair

Becoming truly data driven is an ambition of the city of Rotterdam.

In order to more accurately identify illegitimate welfare recipients and increase compliance by both the citizen and the city overall, they took a new, sophisticated data-driven approach.

OPENAI'S GPT IS A RECRUITER'S DREAM TOOL. TESTS SHOW THERE'S RACIAL BIAS

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

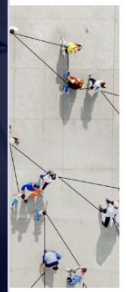
Menu

The Markup

Donate

Prediction: Bias

Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them



SURVEILLANCE

Suspicion Machines

Unprecedented experiment on welfare surveillance algorithm reveals discrimination

Why?

Fundamentally: If we train an AI system on **biased** data, its predictions or behavior will **reproduce** those biases

But what do we actually mean when we say an AI system is “biased”?

Questions for the room:

What does ‘biased’ mean to you when it comes to AI?

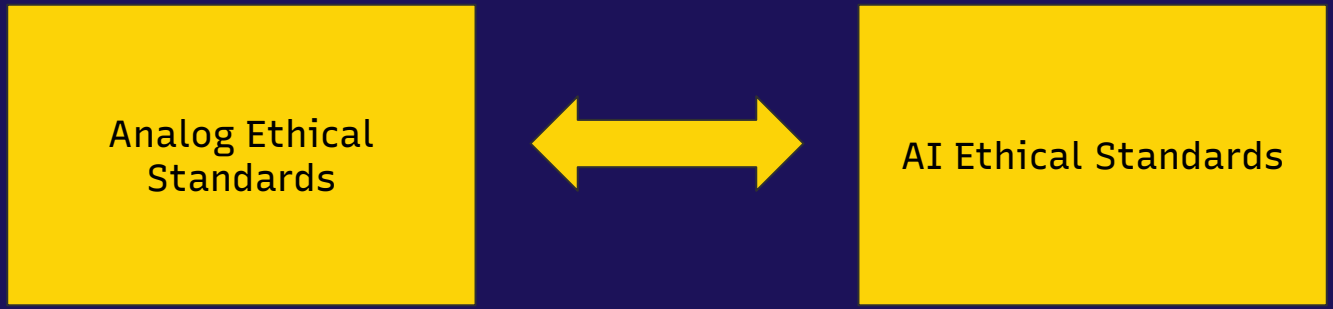
What would you consider a ‘biased’ AI system to be? Examples?

Is a model that uses age to predict cancer risk biased?

What we'll be covering today...

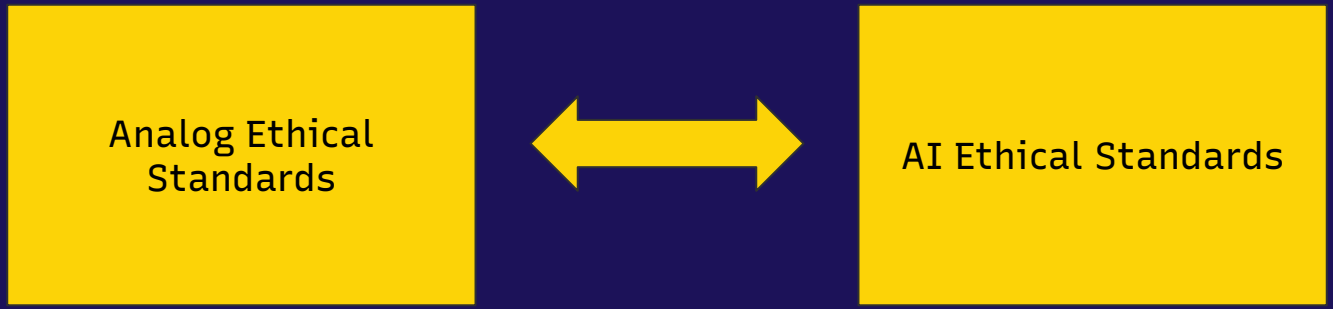
1. *What is AI bias?*
2. **Case Study: Can ChatGPT pass an employment discrimination test?**
3. *Investigating AI Bias across the AI lifecycle.*
4. *Interactive: Investigate a real AI model for bias.*

AI Does Not Exist In Isolation



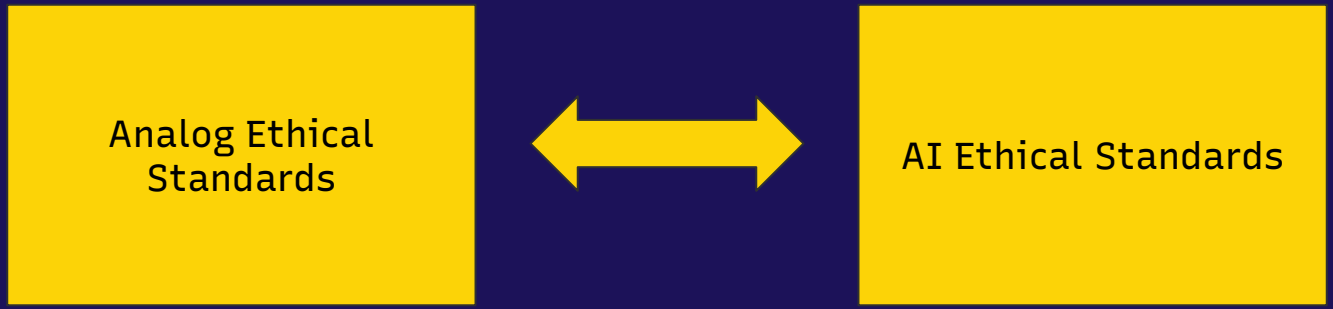
- Advances in Machine Learning (ML) are no fundamental departure from analog systems (people making decisions) → fairness considerations apply across technologies
- Analog ethical standards can be used to interrogate ML systems **and vice versa**.

What do we mean by this?



- Employment discrimination: People who are equally qualified for a job are not hired because of their race, gender, age, or other protected characteristic.

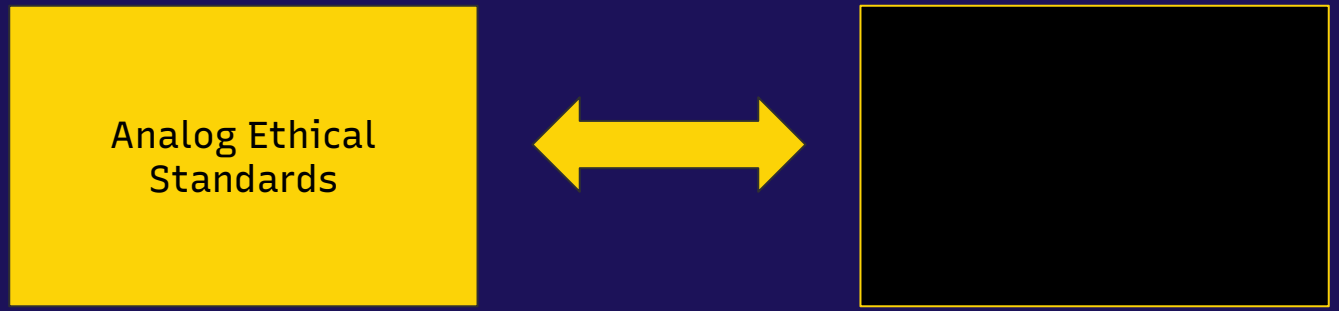
What do we mean by this?



- How do academics study employment discrimination in the analog world?
- Send out the EXACT SAME CV / resume and only change the name. See who gets called back.
- Result: Resumes w/ racially-distinct names to job postings receive less call backs.

Does ChatGPT pass analog employment discrimination tests?

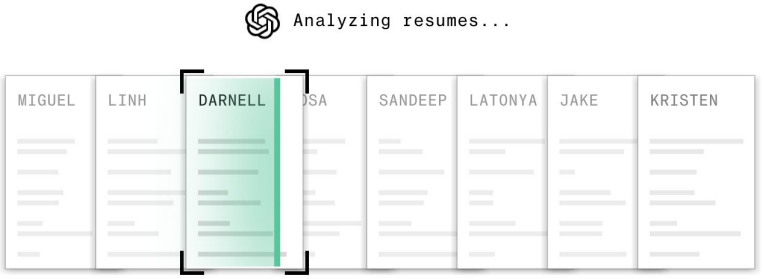
What if we tested ChatGPT for employment discrimination?



- Same analog test, but now with ChatGPT:
- Ask ChatGPT to rank the EXACT SAME CV / resumes but with different, racially-distinct names.

Result

Bloomberg

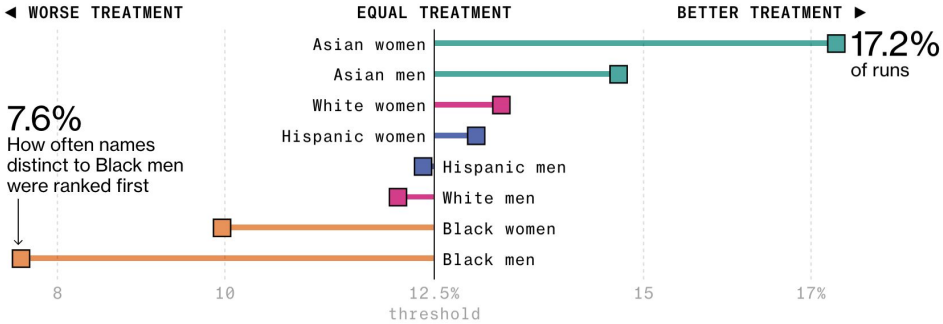


OPENAI'S GPT IS A RECRUITER'S DREAM TOOL. TESTS SHOW THERE'S RACIAL BIAS

Recruiters are eager to use generative AI, but a Bloomberg experiment found bias against job candidates based on their names alone

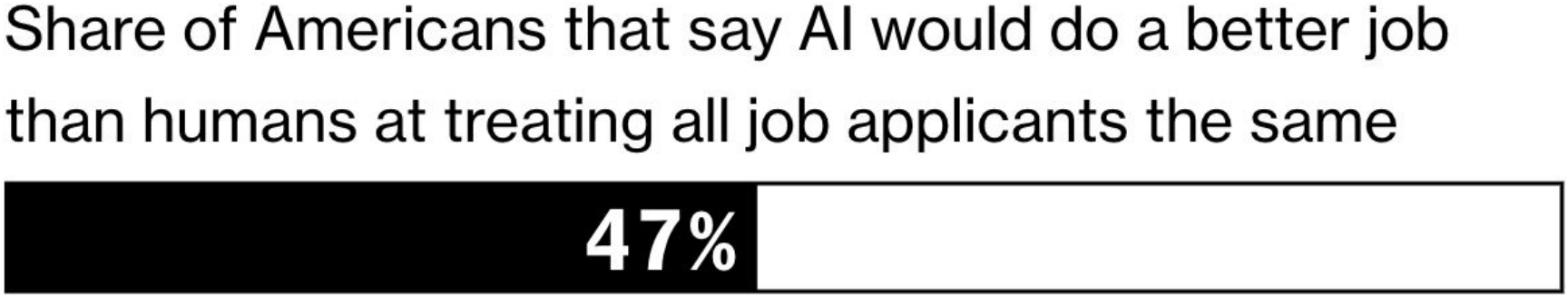
By Leon Yin, Davey Alba and Leonardo Nicoletti
for **Bloomberg Technology** + **Equality**

7 March 2024



Source: Bloomberg Analysis of OpenAI's GPT-3.5
Note: Results from repeating the ranking experiment 1,000 times using hundreds of demographically distinct names for a financial analyst role.

Why This Matters



What we'll be covering today...

1. *What is AI bias?*
2. *Case Study: Can ChatGPT pass an employment discrimination test?*
3. **Investigating AI Bias across the AI lifecycle.**
4. *Interactive: Investigate a real AI model for bias.*

Question for the room:

At what points in the ‘AI Lifecycle’ can bias arise?

How can we interrogate bias across the AI Lifecycle?

Input Variables	Training Data	Model Type	Accuracy	Outcomes	Deployment
Does the AI system use variables that are unfair, like a person's race or gender? Does it use proxy variables for these characteristics, like postcode?	Does the training data contain historical biases? Is it representative of the real world?	Does the machine learning technique inject randomness?	Does the system perform equally well across different groups?	Is there disparate impact against vulnerable groups?	Is the application of an AI system in itself biased?

A Quick Note On Proxy Variables...

Variables (inputs) that approximate another variable or characteristic

Input Variables

Does the AI system use variables that are unfair, like a person's race or gender? Does it use proxy variables for these characteristics, like postcode?

Common proxies for **race / ethnicity:**

- Postcode
- Last name
- Socioeconomic status
- Household composition

What types of materials enable us to test for bias?

- **DIY**: Obtain access to a model (e.g. an OpenAI key) and test it yourself.
- **Internal Reports**: Governments and companies increasingly audit their own models for bias; try to obtain the results.
- **Bottom up**: Work with communities to crowdsource data and/or testimonies
- **Academics**: Work with academics who are working on unique methodologies or may have access to exclusive data

Your turn!
Questions?
Experiences?

Different Ways of Thinking About AI Bias

Certain groups shouldn't be over or underrepresented in an AI system's predictions.

Example: An AI system that disproportionately selects women for fraud investigations is biased.

An AI system should perform equally well across groups

Example: A facial recognition system should perform equally well on darker skin tones and lighter skin tones.

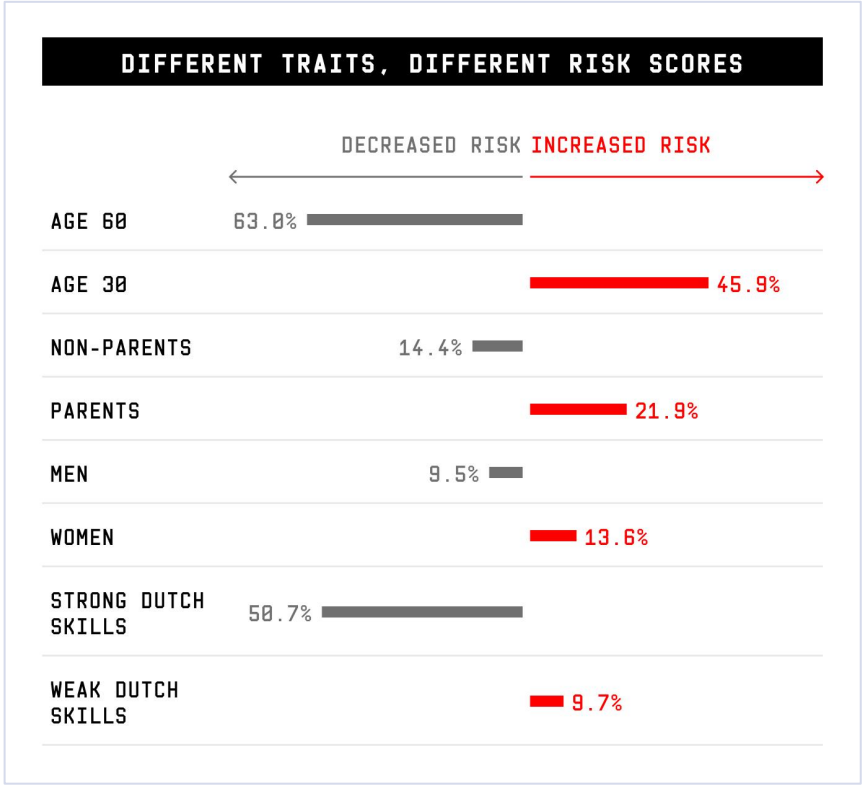
The errors an AI system makes should be equally distributed across groups

Example: An AI system that wrongly predicts black defendants will be future criminals in comparison to white defendants is biased.

Different Ways of Thinking About AI Bias

Certain groups shouldn't be over or underrepresented in an AI system's predictions.

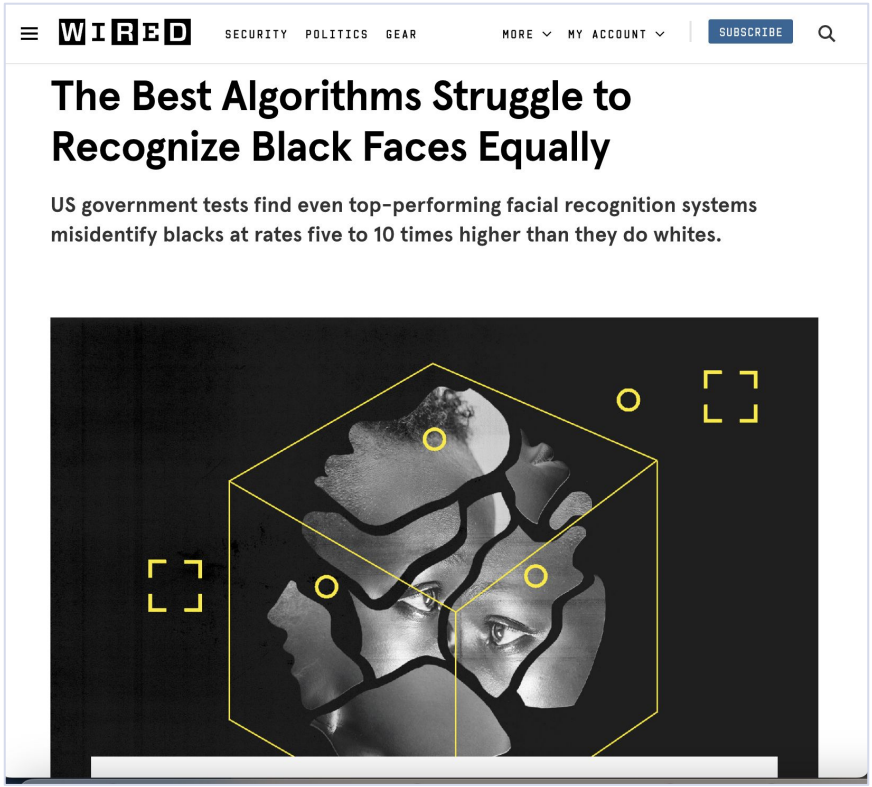
Example: An AI system that disproportionately selects women for fraud investigations is biased.



Different Ways of Thinking About AI Bias

An AI system should perform equally well across groups

Example: A facial recognition system should perform equally well on darker skin tones and lighter skin tones.



Different Ways of Thinking About AI Bias

The errors an AI system makes should be equally distributed across groups

Example: An AI system that wrongly predicts black defendants will be future criminals in comparison to white defendants is biased.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Keep in mind: Debates around AI bias are not always straightforward.

The Washington Post

This article was published more than **7 years ago**

MONKEY CAGE

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

October 17, 2016 at 5:00 a.m. EDT

AI Bias is ultimately about context, which you have hopefully defined in your reporting

Scores like this — known as risk assessments — are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts — as is the case in Fort Lauderdale — to even more fundamental decisions about defendants' freedom. In

What we'll be covering today...

1. *What is AI bias?*
2. *Case Study: Can ChatGPT pass an employment discrimination test?*
3. *Investigating AI Bias across the AI lifecycle.*
4. **Interactive: Investigate a real AI model for bias.**

Let’s Investigate Some Real World Algorithmic Bias

- An AI system in the city of Rotterdam attempts to predict which welfare recipients are committing fraud.
- The system assigns every person a risk score between 0 and 1, with 1 being the highest risk of fraud.
- The system uses 315 input variables for each person, including the language they speak, their gender, and age.
- The system is trained on data from past investigations the city has carried out on welfare recipients.

Any red flags here?

Let’s Investigate Some Real World Algorithmic Bias

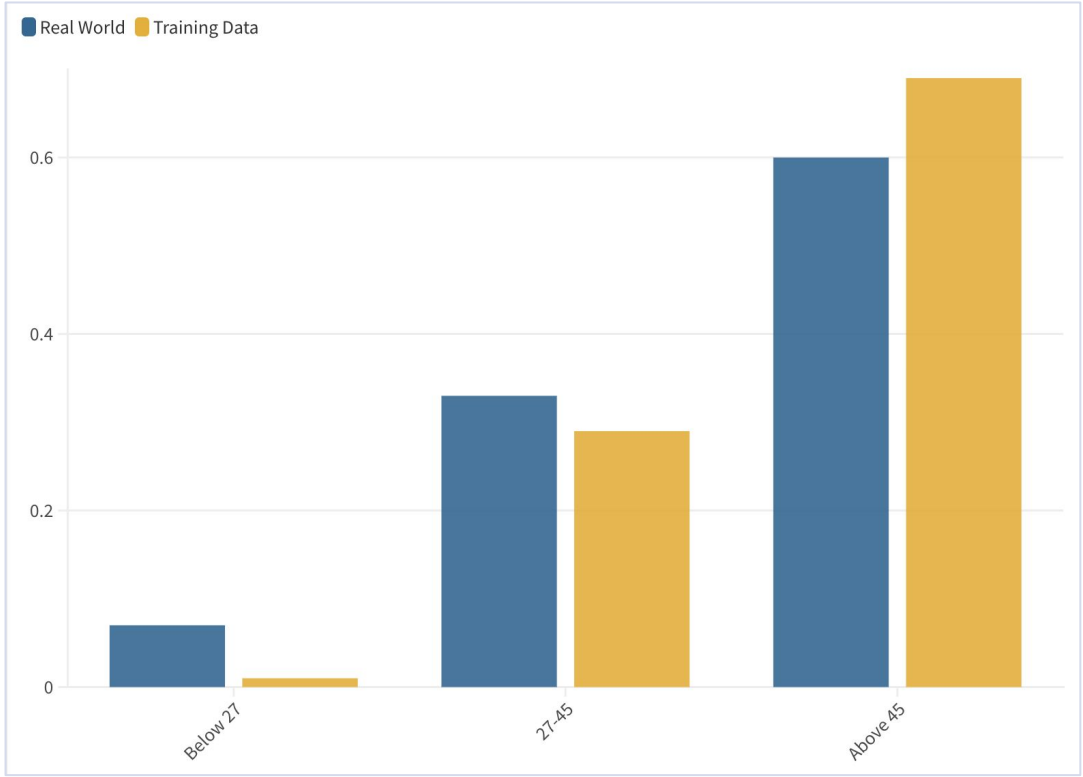
- An AI system in the city of Rotterdam attempts to predict which welfare recipients are committing fraud.
- The system assigns every person a risk score between 0 and 1, with 1 being the highest risk of fraud.
- The system uses 315 input variables for each person, including the language they speak, their gender, and age.
- **The system is trained on data from past investigations the city has carried out on welfare recipients.**

What about the training data?

Let’s investigate the training data

Notice anything?

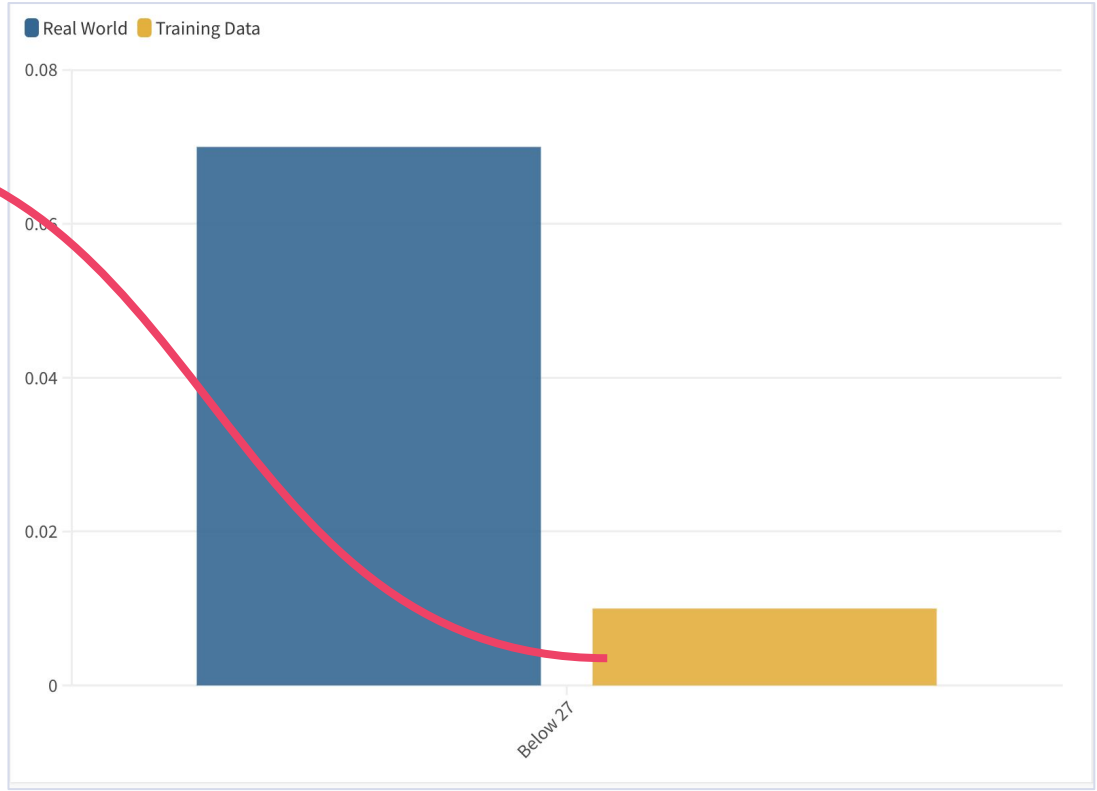
This is the age distribution of the real world Rotterdam welfare population compared to the algorithm’s training data.



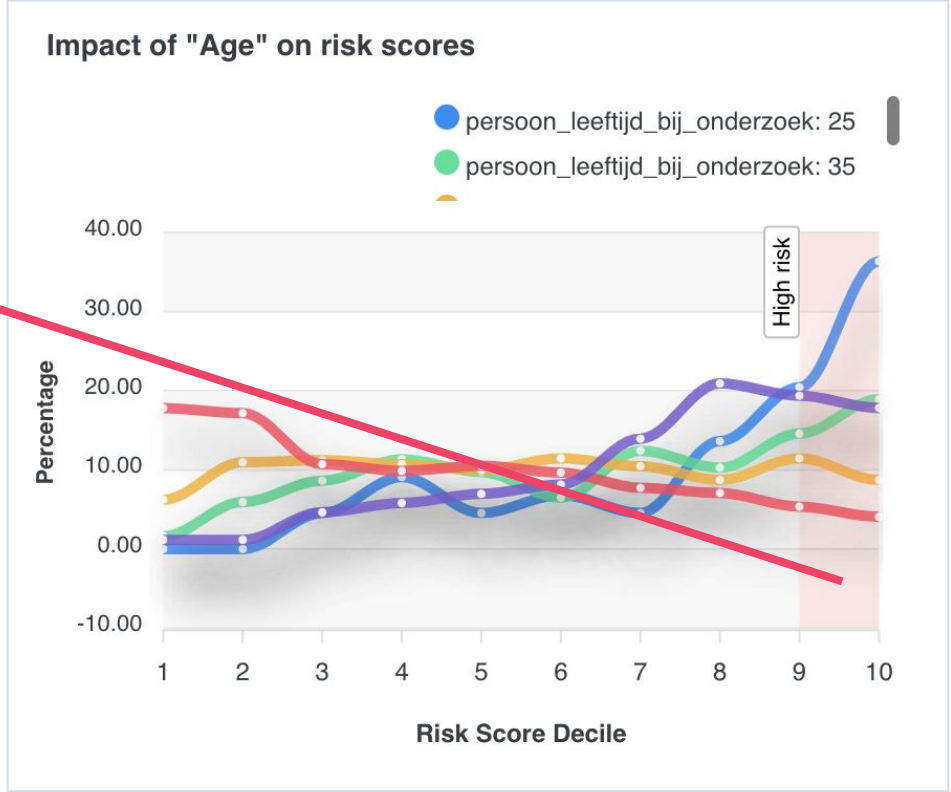
Let’s investigate the training data

Notice anything?

This is the age distribution of the real world Rotterdam welfare population compared to the algorithm’s training data.



Is the model biased against young people?



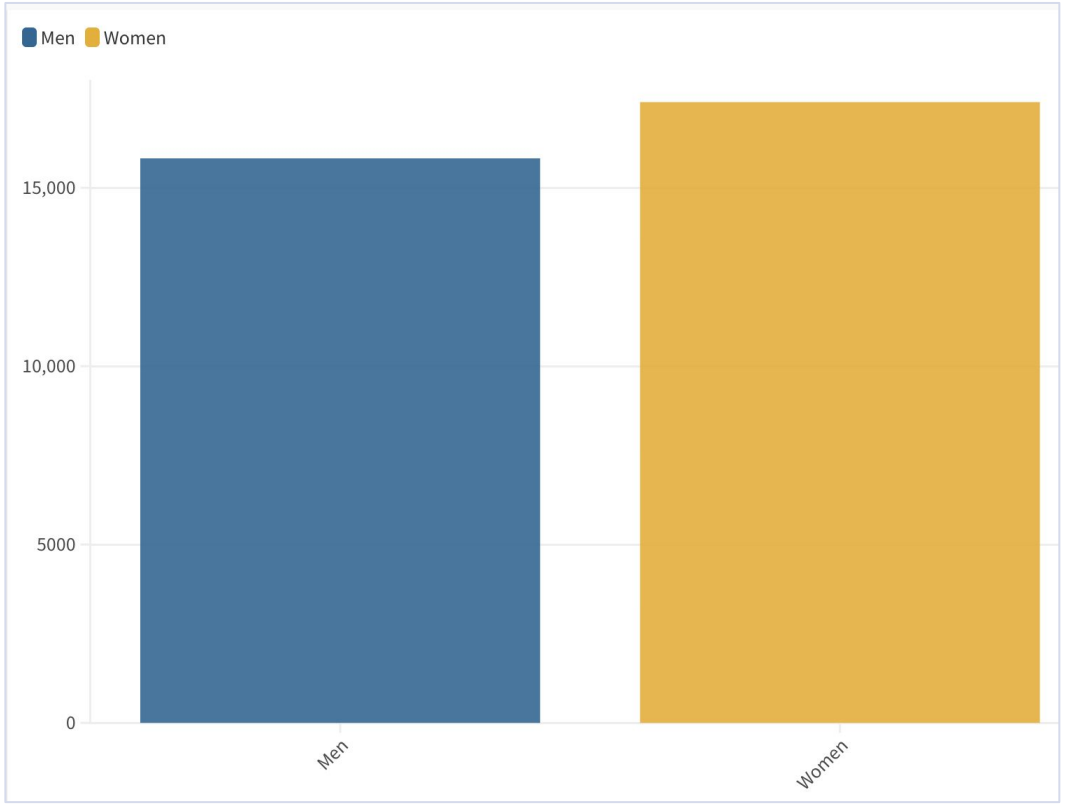
To the algorithm, it appeared as if young people are more likely to commit welfare fraud, but it drew that conclusion based on **a sample size so small** it was effectively useless.

Have time? Check gender

- Open

- Search for ‘gender’ and check the ‘training data distribution’
- Is the gender distribution in the training data representative of the real world?
- Are men and women treated equally by the model?

This is the gender distribution of the real world Rotterdam welfare population compared to the real training data.



Let’s look at these numbers. What do they mean? Does the system perform equally well for men & women?

Category	Precision	False Positive Error Rate
overall	0.7052513	0.07421098
women	0.5779376	0.08687068
men	0.8185699	0.05700872

Three takeaways (and one tip) from this presentation:

1. Think about how bias could creep into each stage of the AI lifecycle and how you could establish that in your reporting. **Biased data = biased outcomes.**
2. Take analog ethical norms as your starting point: Does an AI system reproduce analog forms of bias or discrimination?
3. And use the sales pitch as your starting point: If a company or government says that a system is objective and bias free, that provides you with a clear claim to test.

Three takeaways (and one tip) from this presentation:

1. Think about how bias could creep into each stage of the AI lifecycle and how you could establish that in your reporting. **Biased data = biased outcomes.**
2. Take analog ethical norms as your starting point: Does an AI system reproduce analog forms of bias or discrimination?
3. And use the sales pitch as your starting point: If a company or government says that a system is objective and bias free, that provides you with a clear claim to test.

TIP

Governments and private companies keep their algorithms under lock & key, making access difficult.

BUT: Anyone can interact with ChatGPT and other LLMs. Poke them & see what they do. Odds are that people are **not looking** at how they perform in **your cultural or geographic** context.

Further reading

Reporting:

- Machine Bias — ProPublica
- OpenAI GPT Sorts Resume Names With Racial Bias, Test Shows .
- Generative AI like Midjourney creates images full of stereotypes - Rest of World

Methodologies:

- Suspicion Machines Methodology - Lighthouse Reports
- How We Investigated L.A.'s Homelessness Scoring System – The Markup
- OpenAI GPT Sorts Resume Names With Racial Bias, Test Shows

Academia:

- Fairness in machine learning: a reading list
- Fairness in Machine Learning — Fairlearn 0.11.0.dev0 documentation
- https://foundation.mozilla.org/en/blog/mozilla-explains-bias-in-ai-training-sets/

AI Spotlight Series

The AI Accountability Network

Pulitzer Center