2 WHERE DOES UNFAIRNESS COME FROM?

After implementing the hiring AI, Techie notices discriminatory decisions by the algorithm. So, Techie decides to investigate. Where does the unfairness come from and how did it get into Techie's hiring algorithm?

BIAS

The term **bias** describes a **systematic deviation**. Different research areas each have their own understanding of what exactly "bias" entails. **These different concepts of bias can lead to misunderstandings and misconceptions about how to deal with them.**

In the context of AI, bias was initially used as a term that focussed on technologically induced deviations. The discussion of societal biases in the context of AI is more recent.

BIAS IN AI CONTEXTS

Biases can influence AI models in different ways at different stages of development. Often, a bias is either due to biased data or biased training of the model. AI models then adopt biases, which can have discriminatory effects.

Two examples of biases are explained on the following pages, more can be found on the poster page of this zine!

EXAMPLES OF BIASES

MEASUREMENT BIAS

The features and labels used in the model are **inaccurate representations** of what we're really interested in, and distort the model's view of reality.

How well can a one-page PDF summarise a person's life? Certain important qualities such as kindness, ethics or network and industry connections may not end up on someone's CV.

ALGORITHMIC BIAS

Models introduce bias into the system, causing the **model to systematically favour certain predictors** over others.

During training, the AI model learns that the most interesting information is in the "career" section. It learns to focus on that section of the CV. If a candidate has hobbies that would be great for the job, this would not be recognised by the AI.

HOW TO DEAL WITH BIASES?

So, biases seep into data, models, and how we deal with them from many entry points. One way of dealing with biased datasets is called "**debiasing**". It is a term that is often used in policy documents and understood as a means to solve problems with discrimination. Debiasing describes the idea of treating a biased dataset using **fairness metrics** and technical mitigation methods to gain an "unbiased" dataset.

Even though this promise of a technical solution might sound tempting, it is **not the right way to solve societal problems**. Also, the terminology used, such as "debiasing", can be misleading because it implies that biases, including discrimination, will be mitigated technically. Despite best efforts, bias will persist, and even more so if we do not question societal roots and assumptions behind biases.

See Balayn & Gürses (2021): Beyond Debiasing. Regulating AI and its inequalities for more information

Biases seep into AI systems in various places – and there is no single approach to tackle all of them.

Still, Techie wants to know how discriminatory their AI model actually is. How can Techie measure the bias of their model?

Find out more in the next zines!

FEELS FAIR?

WHICH BIASES INFLUENCE AI?

MEASUREMENT BIAS

The features and labels used in the model are inaccurate representations of what we're really interested in.

How well can a one-page PDF summarise a person's life?

LABEL BIAS

Labels don't accurately represent the categories they are supposed to.

Binary labels misrepresent our complex reality, e.g. giving the impression that the people hired are perfect and the ones rejected are worthless.

EVALUATION BIAS

The model is evaluated using criteria that don't accurately reflect its use in the real world.

The evaluation is done by C.O.R.P.. They care about accuracy and efficiency - not fairness!

FEEDBACK BIAS

The outputs of the model are used as new training data, creating a loop in which initial biases are reinforced.

Given our AI's bias towards *white* men, more *white* men will be hired and more women of colour will be rejected.

HISTORICAL BIAS

Pre-existing societal biases are reflected in the data used to train the model.

Because of our society's sexist and racist history, biases such as "*white* men are more successful in leadership positions" are picked up by AI.

REPRESENTATION BIAS

The data used don't accurately reflect the world's diversity and complexity.

If C.O.R.P.'s dataset mainly labels *white* male applicants as hired, the AI won't learn much about women of colour. If a woman of colour applied to C.O.R.P. in the future, she wouldn't have much of a chance - even if her skills were similar to those hired.

ALGORITHMIC BIAS

Models introduce bias into the system.

During training, the AI learns that the most interesting information lies in the "career" section.

DEPLOYMENT BIAS

Occurs when the context in which the model is used differs from the training environment.

Our training dataset spanned the last 20 years. Characteristics that we value in new applicants may not be present in the old ones.

A BIASED AI IS UNFAIR...

NERD NOTES:

This selection of biases is based on van Giffen et al. (2022): Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods.

Some of the biases are also known under other names; for details, check the paper. "Debiasing" as a solution to biased datasets is mostly mentioned in policy documents, as an investigation from Balayn & Gürses (2021) shows. They found that policy documents did not usually specify what they understood as debiasing.

Debiasing is based on two steps: The first step involves fairness metrics: rules to measure fairness. The second step entails mitigation methods: tools to adjust steps in the AI lifecycle based on the identified level of fairness. However, those technical approaches do not deal with the wide variety of biases and their causes.

IDEA · REALISATION · I. Mastenbroek, I. Kühnlein, B. Lübbert DESIGN · I. Kühnlein, I. Mastenbroek, L. Wunderlich