1 HOW TO CREATE AN AI MODEL?

Meet **Techie** – Techie is an aspiring ethical data scientist committed to using AI for social good. But that is not so easy – follow their journey towards fairer AI!

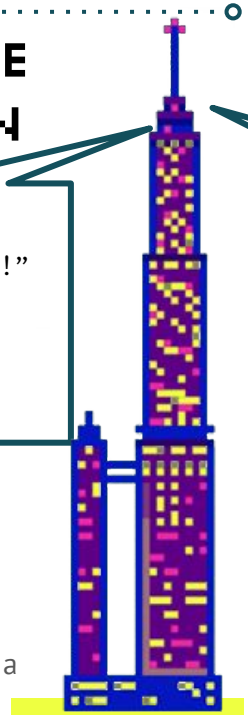
CHALLENGE DEFINITION

"Build a **fair AI** hiring model for us!"

```
Input: CV
Output: hire/reject
```

Techie is approached by the **C.O.R.P.** corporation to automate hiring processes...

...and receives a lot of data.



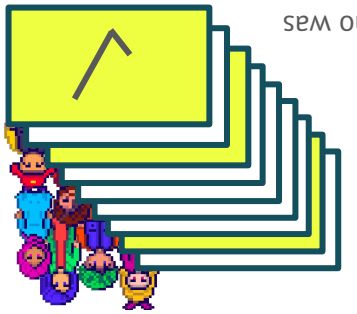
DATA COLLECTION

You'll get 1000 CV's from previous applicants, including **labels** indicating who got hired and who didn't. We want you to build an AI model that looks at a person's CV and predicts whether to hire them or not. A good model would predict "hire" if the person was in fact hired.

TRAINING

To train and test the model with **different data, Techie splits the 1000 CV's in two parts**:

They use 900 CVs to train the hiring model:



Based on this data, the model "learns" who was hired before.

The remaining 100 CVs are used to test the model. This is done to ensure that the model works on new data, so it is tested on data that was not used for training.

The results of this testing show where the model predicted correctly and incorrectly.

TESTING

To evaluate how accurate the model is, its predictions are scrutinised with the help of several **formulas**, for example, this one:

$$\text{Accuracy} = \frac{\text{True Predictions}}{\text{All Predictions}}$$

It asks: **How many predictions were true?** The result is the rate of correct predictions, meaning predictions that matched the label given beforehand (see data collection).

IS IT ACCURATE?

As soon as these rates are high enough, the model is deployed.

This rate reflects how comparable the model's predictions are with past decisions. What if I disagree with past decisions? What can I do to prevent my model from being unfair and learning societal prejudices?

DEPLOYMENT & EVALUATION

C.O.R.P. deploys the algorithm and evaluates how valuable and efficient it is for them.

But civic society has a say in this! And the public backlash says: This model is unfair.

How did that happen? And what can Techie do? Find answers in the next zines!

1

FEELS FAIR?

HOW TO CREATE AN AI MODEL?

THE QUEST BY C.O.R.P.:

Build us a fair AI hiring model:

Create a binary classification model [NN] which looks at a person's CV and predicts whether to hire them (1), or not (0). A good model would predict "hire" if the person was in fact hired.

Find out, based on a CV, if the new applicant fits the C.O.R.P. culture.

Background information:
The majority of C.O.R.P.'s current staff is white, male and earns a high-income.

CHALLENGE DEFINITION

DATA COLLECTION

Techie receives 1000 labeled CVs from C.O.R.P.

PRE-PROCESS DATASET

ADJUST MODEL

PRE-PROCESS DATASET

TRAINING

900 CVs used to train hiring algorithm

TESTING

100 CVs used to test algorithm

IS IT ACCURATE?

YES

NO

DEPLOYMENT

C.O.R.P. uses the AI model to hire new employees. For these new decisions, we cannot use a label to check whether the AI model is making them properly. This makes it a black box.

REAL WORLD EVALUATION

IS YOUR AI FAIR?

TESTING THE ACCURACY

Techie tests the AI's performance in the hiring task by comparing its predictions to past human made hiring decisions, called Ground Truth.

	AI predicted hire	AI predicted reject
Label by C.O.R.P.: hire	TRUE POSITIVE	FALSE NEGATIVE
Label by C.O.R.P.: reject	FALSE POSITIVE	TRUE NEGATIVE

Techie calculates accuracy metrics, which give insight into the rate of "correct" predictions. If these proportions are not satisfactory, Techie refines the model.

NERD NOTES:

A binary model is a mathematical representation of a system or process of which the outcome is either 0 or 1. We chose a simple **binary classification** model (hire vs not hire), as it provides the easiest introduction into how fairness is measured.

During training we use labeled data and know the **Ground Truth** (in Techie's case, was this person previously hired/not hired). This historical knowledge allows for testing the model accuracy. After model deployment, we lack access to such historical decisions.

This transition turns the AI into a **black box** where its decision-making process is not directly observable or comparable to the "Ground Truth".

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

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

3 HOW TO MEASURE FAIRNESS?

Fairness means treating different groups equally. But how can we implement this fluid concept in AI contexts? On the following pages, you'll learn about fairness metrics – approaches to make fairness measurable to act upon.

SENSITIVE ATTRIBUTES

Who are the "different groups" that should be treated equally? The definition of those groups relies on **sensitive attributes, binary divisions based on privilege**:

Privileged

e.g. people read as male, *white*, non-queer

e.g. people read as non-male, non-*white*, queer

Non-privileged

The test data can be used to **see if the AI model behaves differently for different groups**. To do so, **fairness metrics** are used. Many different fairness metrics exist and each one is a specific definition of fairness. On the following pages, there are some examples.

USING FAIRNESS METRICS IS NOT ENOUGH

Feels fair? Each of the metrics may sound fair at first, but achieving fairness is not so easy. The metrics are meant to provide information for action. Measuring fairness is not enough. Moreover, they are all based on binary divisions, contradict each other and are not uncontested! Read about the metrics' critique on the poster page of the next zine.
What else can Techie do to build fair(er) AI? Explore fairer processes on the next zine pages!

Deep dive on the back side!

INDIVIDUAL FAIRNESS METRICS

FAIRNESS THROUGH AWARENESS

Fairness means that **similar people receive similar classifications regardless of their group.**

That means only similar individuals can be compared, not all individuals. Who even gets to decide what similarity means?

FAIRNESS THROUGH UNAWARENESS

Fairness means that **sensitive attributes are not explicitly used** in the decision-making process.

Often, there are ways to predict sensitive attributes, so complete unawareness is impossible. Is unawareness even fair?

GROUP FAIRNESS METRICS

DEMOGRAPHIC PARITY

Fairness means that **both groups have the same likelihood of being accepted.**

What about qualified but less confident groups? There would be less applications from them, but the ratio of qualified ones could be higher, and they would be dismissed...

TREATMENT EQUALITY

Fairness means that **errors by the AI system impact both groups similarly.**

This still relies on an unchecked ground truth. Even if the errors are the same, the hiring rates could be different – and unfair!

MEASURE FAIRNESS?

Fairness metrics work by comparing predictions for individuals or groups. But how are they compared? There are two approaches:

Individual fairness metrics

...rely on comparing the predictions of individuals, for example the metrics fairness through awareness or unawareness (on the zine's pages). Here, the prediction for an individual is compared to the outcomes for another similar or different individual.

Group fairness metrics

...rely on comparing the predictions of the AI model with the actual labeled data for different groups, for example demographic parity or treatment equality (on the zine's pages). Comparing predictions per group is based on a confusion matrix - a table that gives an overview where and for which group the AI predicted correctly:

	AI predicted hire	AI predicted reject
Label by C.O.R.P.: hire	**TRUE POSITIVE** for Group A AND Group B	**FALSE NEGATIVE** for Group A AND Group B
Label by C.O.R.P.: reject	**FALSE POSITIVE** for Group A AND Group B	**TRUE NEGATIVE** for Group A AND Group B

Every fairness metric is a certain definition of fairness. Expressed mathematically, they could look like this:

The model is fair, if

$$\text{Group A: } \frac{\text{All Positives}}{\text{All}} \approx \frac{\text{All Positives}}{\text{All}} \text{ :Group B}$$

This translates into:

The ratio of predicted hires to the total number of applications is similar across groups. **40% female applicants – 40% of the hires should be female!**
It defines fairness as the same likelihood for both groups to be hired. It is called demographic parity - and you have already seen this definition on the zine's pages!

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

NERD NOTES:

Many different fairness metrics exist. Their high number shows how hard finding one specific and overarching, applicable definition of fairness is – and this fuzziness is a strength of fairness as a concept.

We have shown one metric in detail here to explain how the metrics work. When comparing other metrics in more detail, it becomes clear that they also sometimes contradict each other or that one metric's results would describe an AI system as fair while another metric would not.

The fairness metrics are a selection of the fairness metrics presented in Mehrabi et al. (2022): A survey on bias and fairness in Machine Learning; Verma & Rubin (2018): Fairness definitions explained.

4 HOW TO CREATE A FAIR ML AI SYSTEM?

Remember the AI development process from the first zine? Fairness has to be considered in every step. Find out how on the next pages!

DATA COLLECTION

Datasets have limitations and lack diversity. Collecting more and the right data – together with the people affected – is the best way. Also, there are technical ways of enlarging datasets: For example, reweighting (enlarging underrepresented groups in a dataset) or synthetic data (calculated guesses on missing data).

CHALLENGE DEFINITION

Who is in charge? Who defines the challenges? Striving for AI justice means thinking about the lived experiences of groups whose lives are affected by AI. Choosing not to use AI should always be an option to prevent harm.

TRAINING

Models are usually build to be most accurate. What if we **told the model that fairness is also important?**

That could be done by incorporating fairness metrics into the training objective, such that the model optimises for both accuracy and equal treatment while learning.

TESTING

It is not enough to test a model's accuracy; fairness also has to be considered.

Testing the AI system across a **wide range of scenarios and demographics** and **implementing intersectional testing protocols** to be aware of multipliers of disadvantage is an important step in developing fairer AI models.

IS IT FAIR?

Solutions to societal problems won't arise from turning flawed metrics into technical tools of control. Testing the fairness of an AI model by using fairness metrics should not be a way to "fairness wash" an AI but a self-critical process.

Keeping the many people affected by the AI model in mind and taking their experiences into account is more important than using a specific fairness metric to declare an AI model as fair.

DEPLOYMENT & EVALUATION

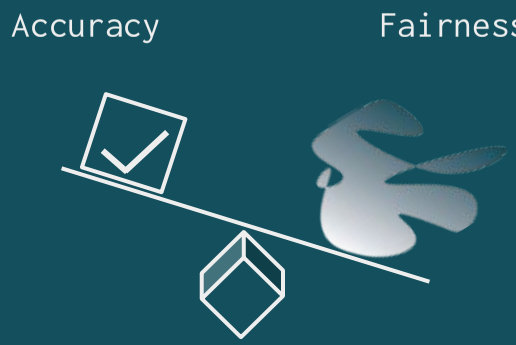
When the AI model is not explainable, the public does not get a say. But fairness is about power sharing!

Explainable AI ensures that applicants receive an explanation of the decision. **Open sourcing** the algorithm for crowdsourced testing might help to bring in a variety of perspectives. Both are approaches for improving the algorithms' fairness.

MEASURE FAIRNESS?

All fairness metrics are based on binary categories and thus make it difficult to account for the nuances of life. There are **two lines of criticism of fairness metrics:**

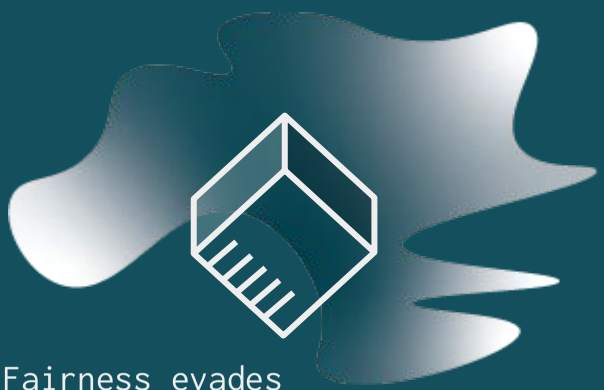
ACCURACY FOCUSED CRITICISM:



Trying to mitigate bias makes AI models less accurate.

Reducing existing bias means changing data sets. These changes could lead to more fairness but also to less accuracy. There needs to be a process of weighing up which criteria to aim for in AI models.

INTERSECTIONALLY REASONED CRITICISM:



Using metrics to measure fairness is too one dimensional.

Class, gender, ethnicity and other individual characteristics are not binary, they intersect and overlap. Fairness – treating different groups equally – is not even always the right approach. When it comes to justice, approaches such as equity, providing equal opportunities to participate, are even more important. All these concepts are fluid and cannot be captured in formulas. Recognition is more important than metrics: Do the people affected have a say in the process?

Using fairness metrics to measure fairness is not enough. As we move towards fairer AI, considering fairness in every step of AI development will be unavoidable. Find out how on the zine's pages.

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

NERD NOTES:

Let's consider Goodhart's law, a saying that is usually quoted as **"When a measure becomes a target, it ceases to be a good measure".**
What could that mean in our context of AI and fairness metrics?

It emphasises the danger of a particular metric being misused once it becomes a means of control. If an AI model is optimised for a certain quantifiable notion of fairness, does that really mean that the AI system is becoming fairer?

It also questions our motivation behind fair AI systems: Do we understand fairness as a measurable, manageable concept - or are we open to understanding and rethinking the complex processes of discrimination and really acting on them?

1

Let the green side of the paper face up

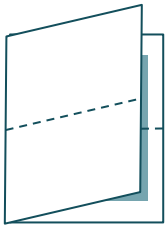
2

Fold it in half

3

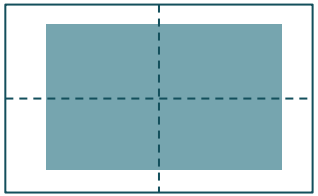
Open it again

4



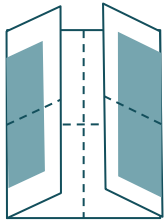
Fold it in half in the other direction

5



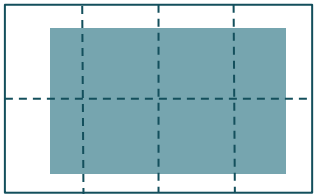
Open it again

6



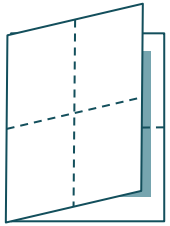
Flip the paper around, green side facedown. Now fold right and left sides of the paper to the crease in the middle.

7



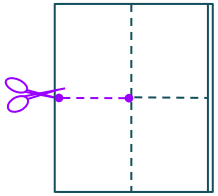
Open it again and flip, green side faces up

8



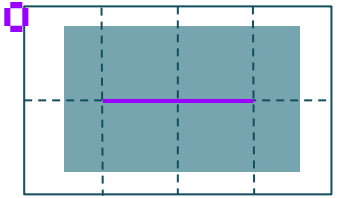
Fold it in half

9



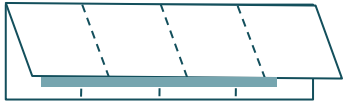
and cut from you fold inwards to the middle, where the two folds meet

10



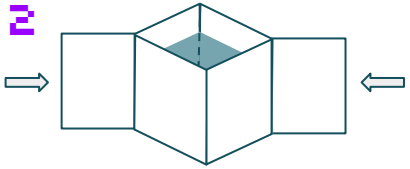
Open it again

11



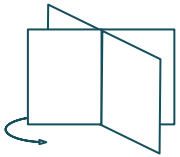
Fold it in half

12



and push both ends inwards so that your cut opens

13



now fold all sides together and flip until the front page turns forward

14

Tadaa! Enjoy reading :)