

# FEELS FAIR?

## HOW TO CREATE AN AI ALGORITHM?



Meet **Techie**, a data scientist who wants to use AI for social good. Techie is approached by corporation C.O.R.P to automate their hiring processes. This AI lifecycle shows the steps of AI development.



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

Build us a fair AI hiring algorithm:

Create a binary classification model <sup>NN</sup> 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 has previously been 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

C.O.R.P. provides a dataset of 1000 CVs from their previous applicants, including a label, indicating who got hired (1) and who didn't (0).

PRE-PROCESS DATASET

### TRAINING

900 CVs are used to train the hiring algorithm



ADJUST MODEL

RE-PROCESS DATASET

### TESTING

100 CVs are used to test the algorithm

YES

IS IT ACCURATE?

NO

### DEPLOYMENT

C.O.R.P. now uses the algorithm to hire their new employees.

### REAL WORLD EVALUATION

C.O.R.P. evaluates how efficient and valuable the algorithm is for them. But is the algorithm also fair?

YOU MADE AN AI.  
BUT IS IT FAIR?

### TESTING THE ACCURACY

Techie tests the AI's performance in the hiring task by comparing its predictions to past decisions (Ground Truth<sup>NN</sup>):

	AI predicted hire	AI predicted reject
Ground Truth said hire	TRUE POSITIVE	FALSE NEGATIVE
Ground Truth said reject	FALSE POSITIVE	TRUE NEGATIVE

Techie calculates common accuracy metrics, which give insight into the rate of "correct" predictions. If these proportions are not satisfying<sup>NN</sup>, Techie will refine and retrain the model.

$$\text{Precision} = \frac{\text{True Positive Predictions}}{\text{All Positive Predictions}}$$

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

### NERD NOTES (NN)

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 friendliest introduction into how fairness is measured. In reality, a wide array of AI's are trained to perform more complex tasks such as image/video/text/audio generation based on a prompt. Even though those other algorithms aren't trained to perform binary classification, they do rely on the same binary fairness metrics.

We use two different coloured boxes to represent our AI model, to highlight a key phase difference: during training, we use labeled data, a situation in which we know the **Ground Truth** (in Techies case, was this person hired/not hired). This knowledge allows for testing the model accuracy. We deploy the model when the accuracy is **satisfying**, meaning above some predetermined threshold. After model deployment, we lack access to such historical decisions, instead relying on the model's presumed accuracy. This transition essentially turns the AI into a **black box** where its decision-making processes are not directly observable or comparable to the Ground Truth.

A project by the Alexander von Humboldt Institute for Internet and Society within the AI Society Lab.

IDEA - REALISATION Irma Mastenbroek, Irina Kühnlein, Birte Lübbert

DESIGN - Irma Mastenbroek, Larissa Wunderlich, Irina Kühnlein, Birte Lübbert

These posters and accompanying material are published under a Creative Commons Attribution Licence (CC-BY-4.0).