WHAT ABOUT FAIRNESS & AI?

Many digital products, in all parts of society, use **AI and machine learning algorithms**. An AI algorithm consists of a set of instructions which, if executed correctly by a computer/software, can solve a problem or complete a task. Yet these systems are often a **black box** and it is not always clear to humans exactly how the AI system arrives at a decision. There are several examples where we as a society agree that AI systems reaffirm existing **(societal) biases**. Consider facial recognition software that recognises certain skin colours better than others. To remedy this, an algorithm can be used that takes these biases into account by applying certain '**fairness notions**'. However, many different notions exist and each is based on different definition of fairness.

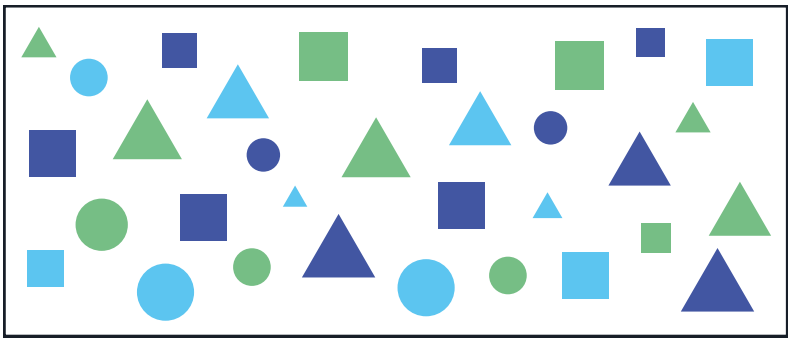
In this brainfood, we present some of these fairness principles using the example of an automated job application process. Which of these principles do you think is the most fair?

This brAInfood was prepared in consultation with Carmen Mazijn and An Jacobs of the IRP COMPASS project of the Vrije Universiteit Brussel.

Knowledge Centre Data & Society (2022). What about fairness & AI? Brussels.

This document is available under a CC BY 4.0 license.

ALL APPLICANTS FOR A JOB

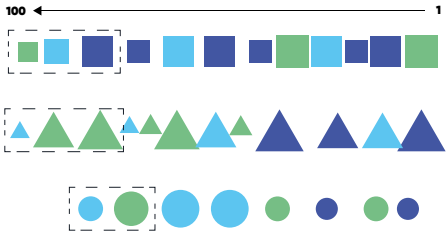


IMAGINE...

Above, you can see all the people who have sent in their CV to apply for a specific job. Each of these fictitious applicants is a little different, in shape, size or colour, but in theory each could be qualified for the job. Who gets invited for an interview and who does not? The company uses an AI system to make this decision.

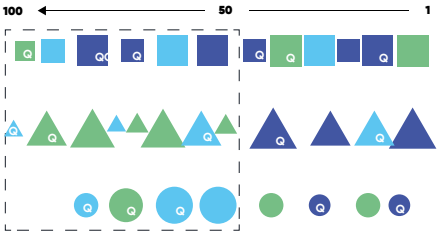
As a starting point, the algorithm calculates a score between 1 and 100 for each applicant **based on historical data**, and those with **the highest score** are invited to interview. But this is not necessarily a fair way of selection. For instance, the algorithm may disadvantage a certain group, like the dark blue circles in this example. To address this, we can apply several fairness principles.

NOTION 1: STATISTICAL PARITY



By applying '**statistical parity**' to your algorithm, candidates are first divided based on certain (protected) characteristics, in this case shape, and then sorted in groups. For each group, the same proportion or percentage of those with the highest scores are then selected for interview. In this way, **each group is considered equally** in the selection process.

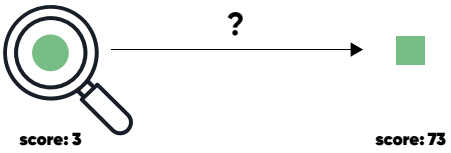
NOTION 2: EQUAL OPPORTUNITY



Instead of inviting a fixed number of people to an interview, a company can also choose to invite everyone who has achieved **at least a certain score**. We call this a **threshold**. In this example, the threshold is at score 50.

The Q indicates that this person is actually qualified for the job. This value cannot be known with certainty in advance and can only be tested after recruitment. Using the fairness principle '**equal opportunity**', you can check whether the **error rate** for people who were effectively qualified, but not above the threshold value, is **equal** in each group.

NOTION 3: CAUSAL DISCRIMINATION



The previous principles each looked at different groups. We can also look from the individual's point of view, which is called **individual fairness**. For example, '**causal discrimination**' assesses whether someone with the same characteristics, except one, would receive a significantly different score. If this is the case for a characteristic that is legally protectewd, such as gender or ethnicity, then the algorithm is not fair and correct.

CONCLUSION

There are many other fairness principles that developers can apply to their AI systems. In each context, a different principle may be most appropriate. This therefore poses **a challenge for developers** to choose which principle to use in which situation. Would you also like to join that discussion and **experience for yourself** how it feels when an AI system judges you? In the workshop "Does the computer give you the job you deserve?", you can experience the fairness notions for yourself.

THE OUTCOME OF THE ALGORITHM WITHOUT USING FAIRNESS NOTIONS