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

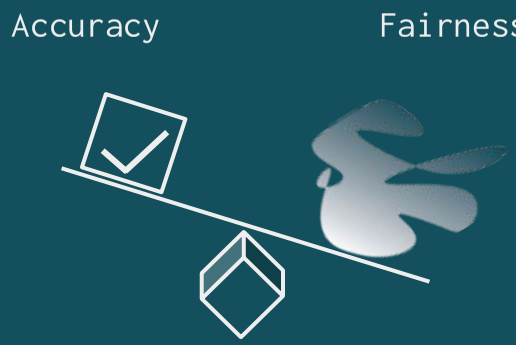
When the AI model is not explainable, the public does not get a say. But fairness is about power sharing!

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:**

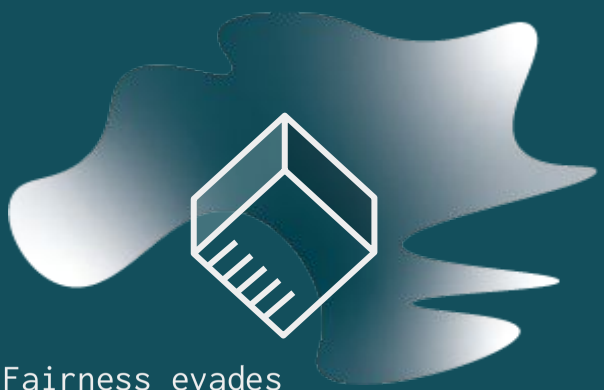
ACCURACY FOCUSED CRITICISM:



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?