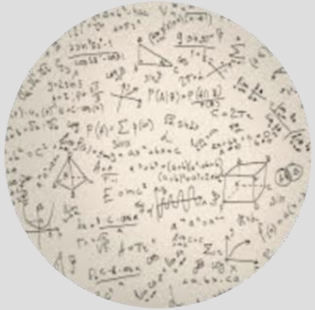
The Alan Turing Institute

Methods for detecting, measuring, and mitigating bias in Machine Learning

Dr. Ogerta Elezaj

What does it take to trust a decision made by a machine?

Is it fair?



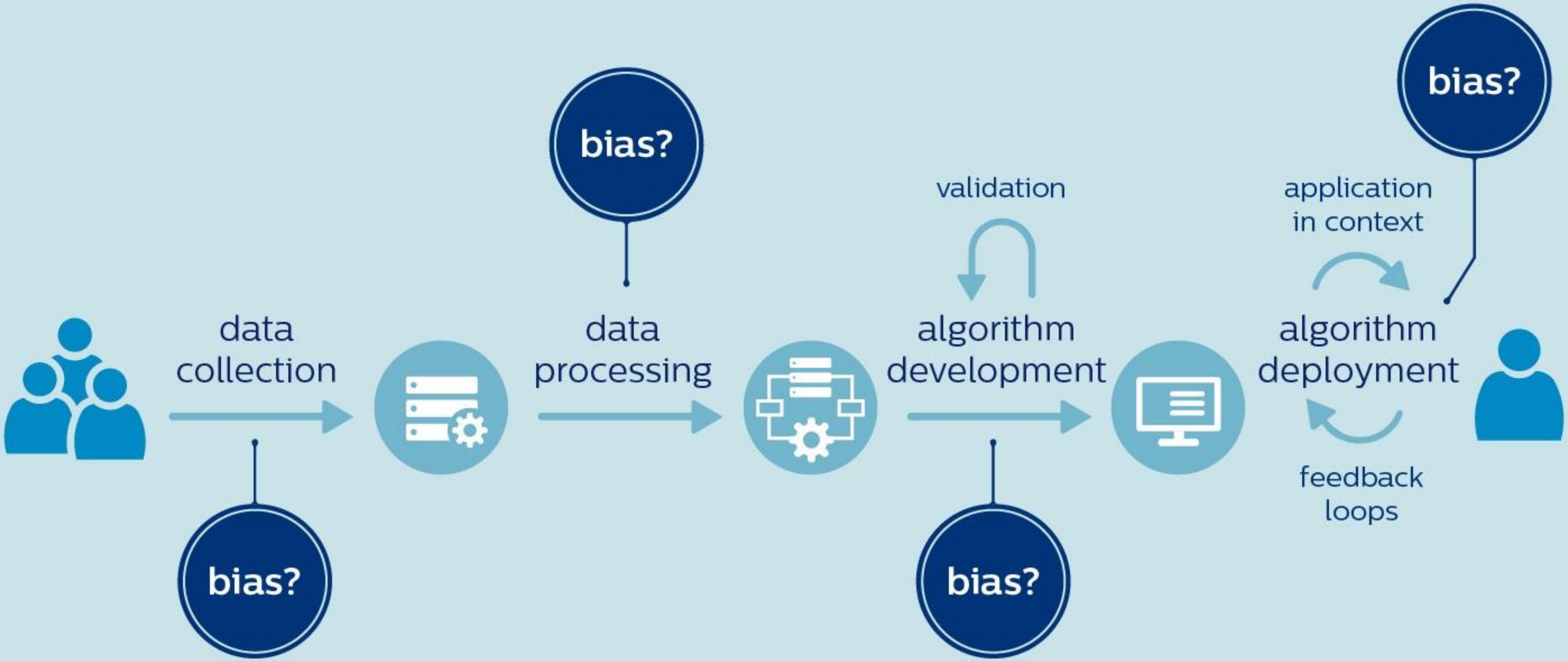
Is it easy to understand?

Did anyone tamper with it?

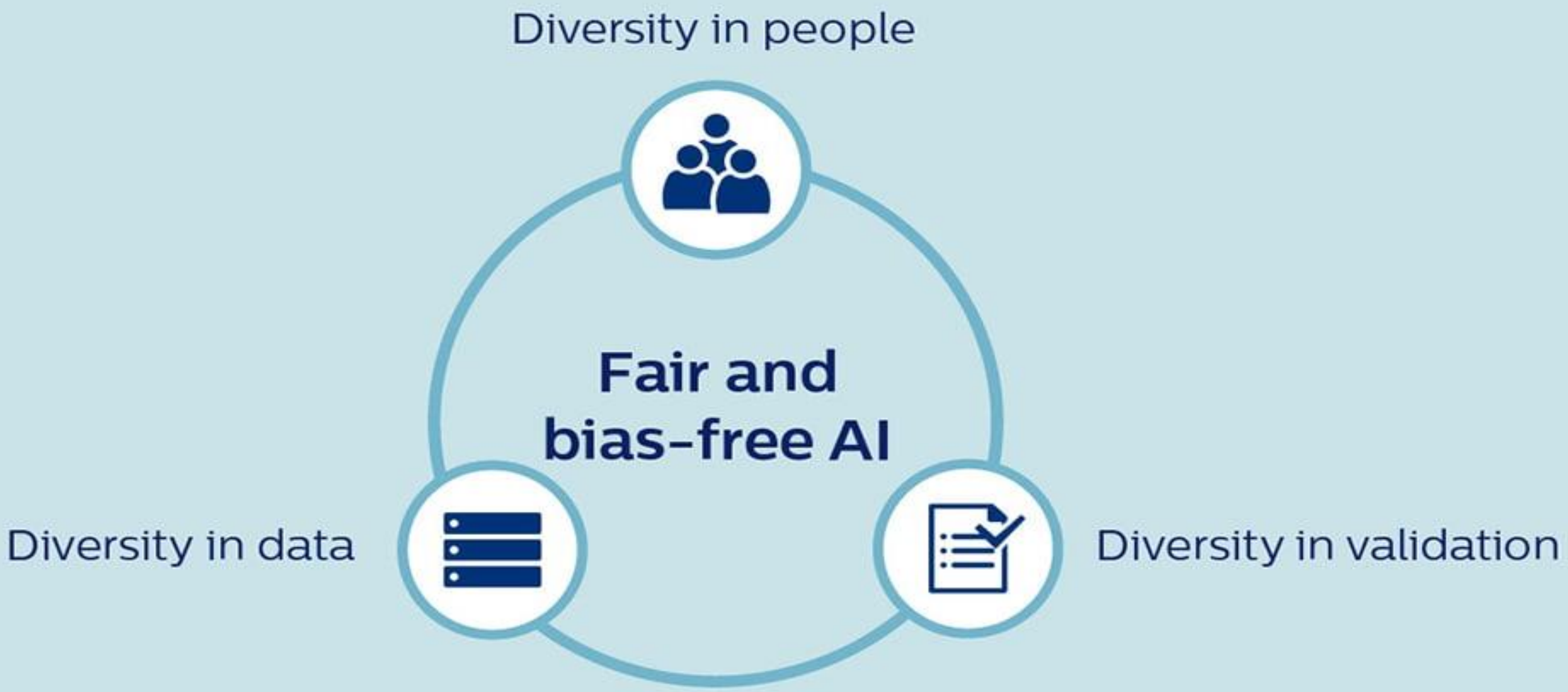
Is it accountable?

Where Does Bias Come From& How Do You Measure it ?

Bias Sources



Diversity that matters



What are some first steps in assessing your use case for fairness in machine learning?

- Consider your product's context and use.
  - Does your use case or product specifically use any of the following data: biometrics, race, skin color, religion, sexual orientation, socioeconomic status, income, country, location, health, language, or dialect?
  - Does your use case or product use data that's likely to be highly correlated with any of the personal characteristics that are listed above (for example, zip code and other geospatial data are often correlated with socioeconomic status and/or income; similarly, image/video data can reveal information about race, gender, and age)?
  - Could your use case or product negatively impact individuals' economic or other important life opportunities?

Open Source AI fairness Tools

Open Source AI fairness Tools

05/03/18	facebook	Facebook says it has a tool to detect bias in its artificial intelligence	Quartz
05/25/18	Microsoft	Microsoft is creating an oracle for catching biased AI algorithms	MIT Technology Review
05/31/18	pymetrics	Pymetrics open-sources Audit AI, an algorithm bias detection tool	VentureBeat
06/07/18	Google	Google Education Guide to Responsible AI Practices – Fairness	Google
06/09/18	accenture	Accenture wants to beat unfair AI with a professional toolkit	TechCrunch

AI Fairness 360

- Open Source Toolbox to Mitigate Bias
- Demos & Tutorials on Industry Use Cases
- Fairness Guidance
- Comprehensive Toolbox
  - 75+ Fairness metrics
  - 10+ Bias Mitigation Algorithms
  - Fairness Metric Explanations

AI Fairness 360

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

Python API Docs ↗ | Get Python Code ↗ | Get R Code ↗

Not sure what to do first? Start here!

Read More

Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

Try a Web Demo

Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.

Watch Videos

Watch videos to learn more about AI Fairness 360.

Read a paper

Read a paper describing how we designed AI Fairness 360.

Use Tutorials

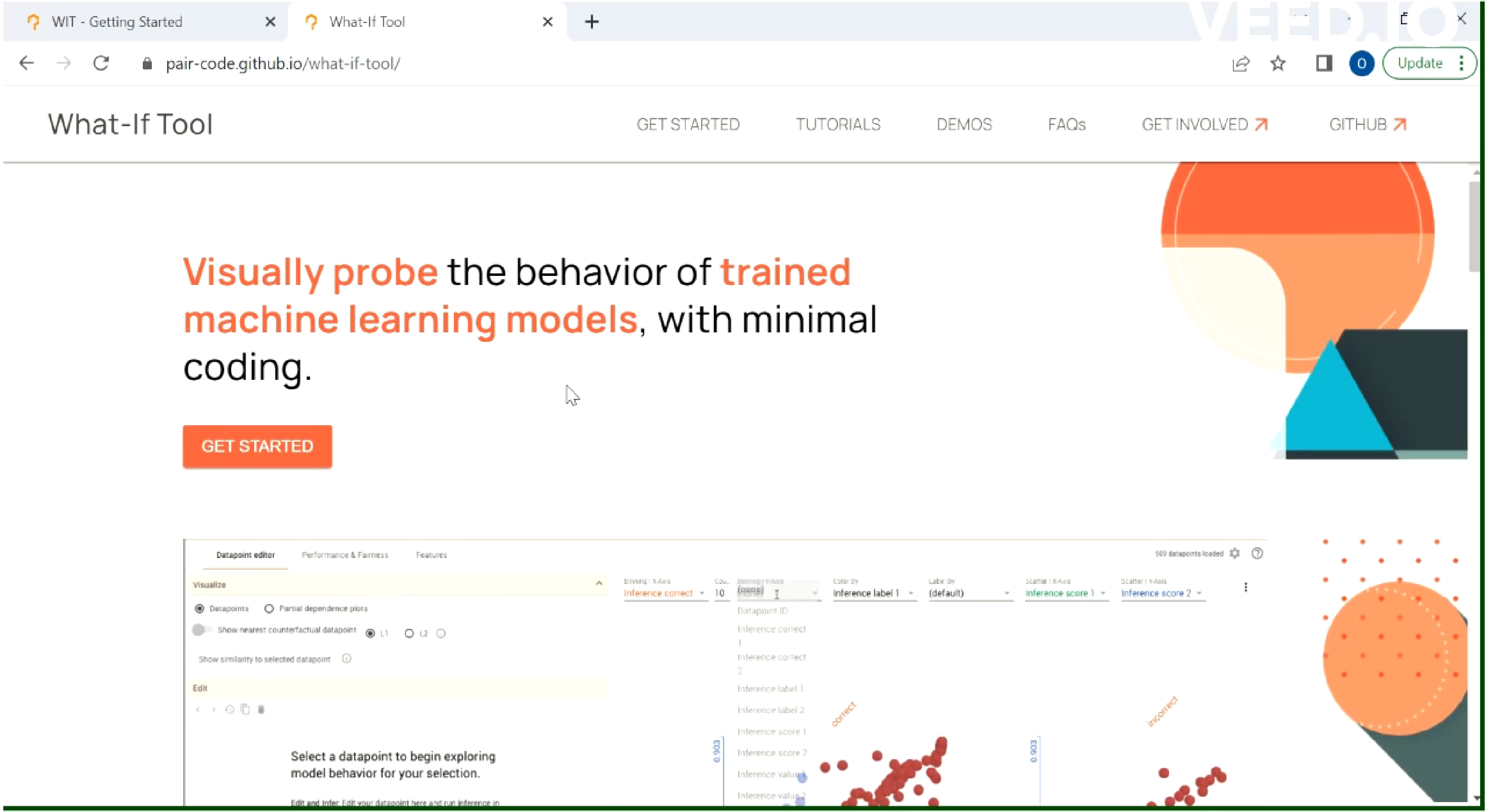
Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application domains.

AIF360 includes the top Algorithms In Industry/Academia

- Optimized Preprocessing (Calmon et al., NIPS 2017)
- Meta-Algorithm for Fair Classification (Celis et al., FAT* 2019)
- Disparate Impact Remover (Feldman et al., KDD 2015)
- Equalized Odds Postprocessing (Hardt et al., NIPS 2016)
- Reweighing (Kamiran and Calders, KIS 2012)
- Reject Option Classification (Kamiran et al., ICDM 2012)
- Prejudice Remover Regularizer (Kamishima et al., ECML PKDD 2012)
- Calibrated Equalized Odds Postprocessing (Pleiss et al., NIPS 2017)
- Learning Fair Representations (Zemel et al., ICML 2013)
- Adversarial Debiasing (Zhang et al., AIES 2018)

Google’s What-if tool

- Web application which allows users to analyze an ML model without the need of writing code
- Interactive visual interface to explore the model results
- Two major features
  - Counterfactuals
  - Performance and Algorithmic Fairness analysis


WIT - Getting Started
What-If Tool
pair-code.github.io/what-if-tool/
Update
What-If Tool
GET STARTED
TUTORIALS
DEMOS
FAQs
GET INVOLVED
GITHUB
Visually probe the behavior of trained machine learning models, with minimal coding.
GET STARTED
Datapoint editor
Performance & Fairness
Features
Visualize
Datapoints
Partial dependence plots
Show nearest counterfactual datapoint
Show similarity to selected datapoint
Edit
Select a datapoint to begin exploring model behavior for your selection.
Inference correct
Inference label 1
(default)
Inference score 1
Inference score 2
Datapoint ID
Inference label 2

Microsoft fairlean

- A tool to assess AI system’s fairness and mitigate any observed unfairness issues

- A Python package containing mitigation algorithms as well as a Jupyter widget for model assessment.

- Two major components
  - A dashboard -assessing which groups are negatively impacted by a model, and for comparing multiple models
  - Algorithms - mitigating unfairness in a variety of AI tasks and along a variety of fairness definitions

Detecting and mitigating gender bias on credit decisions

Biases and Machine Learning

A machine learning model makes predictions of an outcome for a particular instance. (Given an instance of a loan application, predict if the applicant will repay the loan.) The model makes these predictions based on a training dataset, where many other instances (other loan applications) and actual outcomes (whether they repaid) are provided. Thus, a machine learning algorithm will attempt to find patterns, or generalizations, in the training dataset to use when a prediction for a new instance is needed.

However, sometimes the patterns that are found may not be desirable or may even be illegal. For example, a loan repay model may determine that gender plays a role in the prediction of repayment because the training dataset happened to have better repayment for one age group than for another. This raises two problems: 1) the training dataset may not be representative of the true population of people of both gender groups, and 2) even if it is representative, it is illegal to base any decision on a applicant's gender, regardless of whether this is a good prediction based on historical data.

AI Fairness 360 is designed to help address this problem with *fairness metrics* and *bias mitigators*. Fairness metrics can be used to check for bias in machine learning workflows. Bias mitigators can be used to overcome bias in the workflow to produce a more fair outcome.

Fairlearn algorithms

Algorithm	**Description**	**Classification/Regression**	**Sensitive features**
fairlearn. reductions. ExponentiatedGradient	Black-box approach to fair classification described in **A Reductions Approach to Fair Classification**	binary classification	categorical
fairlearn. reductions. GridSearch	Black-box approach described in Section 3.4 of **A Reductions Approach to Fair Classification**	binary classification	binary
fairlearn. reductions. GridSearch	Black-box approach that implements a grid-search variant of the algorithm described in Section 5 of **Fair Regression: Quantitative Definitions and Reduction-based Algorithms**	regression	binary
fairlearn. postprocessing. ThresholdOptimizer	Postprocessing algorithm based on the paper **Equality of Opportunity in Supervised Learning**. This technique takes as input an existing classifier and the sensitive feature, and derives a monotone transformation of the classifier’s prediction to enforce the specified parity constraints.	binary classification	categorical

Conclusions

- Fairness is essentially contested and the appropriate measure of fairness is often context-dependent.
- While it is the responsibility of the user to decide what is ultimately fair or not, the toolkit should provide a wide range of fairness measures to aid its users in their justification.
- The existing solutions support metrics for **binary classification**.
- Support for **multi-classification** problems and other **non-supervised** learning problems seem to be lacking in these solutions.

Reading/References

- https://aif360.mybluemix.net
- https://pair-code.github.io/what-if-tool/
- https://github.com/fairlearn/fairlearn
- http://www.jennwv.com/papers/checklists.pdf
- ps://doi.org/10.1038/s41598-022-07939-1